Course:

Advanced Nonlinear Programming
a.k.a.
Lectures on Modern Convex Optimization
ISyE 7683 Spring 2019

- **Instructor:** Dr. Arkadi Nemirovski nemirovs@isye.gatech.edu, Groseclose 446, Office hours: Monday 10:00-12:00
- **Teaching Assistant:** none
- **Classes:** Monday-Wednesday 3:00-4:15, Groseclose 402
- **Lecture Notes, Transparencies:** course site and https://www2.isye.gatech.edu/~nemirovs/LMCO_LN.pdf
  https://www2.isye.gatech.edu/~nemirovs/Trans_ModConvOpt.pdf
- **Grading Policy:** Take Home Final Exam: 100%
A man searches for a lost wallet at the place where the wallet was lost.
A wise man searches at a place with enough light...

♣ Where should we search for a wallet? Where is “enough light” – what Optimization can do well?

The most straightforward answer is: we can solve well convex optimization problems.

The very existence of what is called Mathematical Programming stemmed from discovery of Linear Programming (George Dantzig, late 1940’s) – a modeling methodology accompanied by extremely powerful in practice (although “theoretically bad”) computational tool – Simplex Method. Linear Programming, which is a special case of Convex Programming, still underlies the majority of real life applications of Optimization, especially large-scale ones.
Around mid-1970’s, it was shown that

- Linear and, more generally, Convex Programming problems are efficiently solvable – under mild computability and boundedness assumptions, generic Convex Programming problems admit polynomial time solution algorithms.

As applied to an instance of a generic problem, like Linear Programming

\[
\mathcal{LP} = \left\{ \min_x \{c^T x : Ax \geq b\} : \begin{array}{l}
A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, \\
c \in \mathbb{R}^n, m, n \in \mathbb{Z}
\end{array} \right\},
\]

a polynomial time algorithm solves it to a whatever high required accuracy \( \epsilon \), in terms of global optimality, in a number of arithmetic operations which is polynomial in the size of the instance (the number of data entries specifying the instance, \( O(1)mn \) in the case of \( \mathcal{LP} \)) and the number \( \ln(1/\epsilon) \) of required accuracy digits.

\[\Rightarrow\text{Theoretical (and to some extent – also practical) possibility to solve convex programs of reasonable size to high accuracy in reasonable time}\]
\begin{itemize}
\item No polynomial time algorithms for general-type nonconvex problems are known, and there are strong reasons to believe that no such methods exist.
\end{itemize}

\textit{⇒} Solving general nonconvex problems of not too small sizes is usually a highly unpredictable process: with luck, we can improve somehow the solution we start with, but we never have a reasonable a priory bound on how long it will take to reach desired accuracy.
From purely academical viewpoint, polynomial time solvability of Convex Programming is a straightforward consequence of the following statement:

**Theorem** [circa 1976] Consider a convex problem

\[
\text{Opt} = \min_{x \in \mathbb{R}^n} \left\{ f(x) : g_i(x) \leq 0, 1 \leq i \leq m \right\}
\]

\[
\left. \begin{array}{c}
|f(x)| \leq 1, |g_j(x)| \leq 1 \\
1 \leq j \leq n
\end{array} \right\}
\]

normalized by the restriction

\[
|f(x)| \leq 1, |g_j(x)| \leq 1 \ \forall x \in B = \{|x_j| \leq 1 \forall j\}.
\]

For every \( \epsilon \in (0, 1) \), one can find an \( \epsilon \)-solution

\[
x_\epsilon \in B : f(x_\epsilon) - \text{Opt} \leq \epsilon, g_i(x_\epsilon) \leq \epsilon \ \forall i
\]

or to conclude correctly that the problem is infeasible at the cost of at most

\[
3n^2 \ln \left( \frac{2n}{\epsilon} \right)
\]

computations of the objective and the constraints, along with their (sub)gradients, at subsequently generated points of \( \text{int} B \), with \( O(1)n(n + m) \) additional arithmetic operations per every such computation.
The outlined Theorem is sufficient to establish theoretical efficient solvability of generic Convex Programming problems. In particular, it underlies the famous result (Leo Khachiyan, 1979) on polynomial time solvability of $\mathcal{L}P$ – the first ever mathematical result which made the C2 page of *New York Times* (Nov 27, 1979).

From practical perspective, however, polynomial type algorithms suggested by Theorem are too slow: the arithmetic cost of an accuracy digit is at least

$$O(n^2n(m + n)) \geq O(n^4),$$

which, even with modern computers, allows to solve in reasonable time problems with hardly more than 100 – 200 design variables.

The low (although polynomial time) performance of the algorithms in question stems from their *black box oriented* nature – these algorithms do not adjust themselves to the structure of the problem and use a priori knowledge of this structure solely to mimic *First Order oracle* reporting the values and (sub)gradients of the objective and the constraints at query points.
Note: A convex program always has a lot of structure — otherwise how could we know that the problem is convex? A good algorithm should utilize a priori knowledge of problem’s structure in order to accelerate the solution process.

Example: The LP Simplex Method is fully adjusted to the particular structure of an LP problem. Although not a polynomial time one, this algorithm in reality is capable to solve LP’s with tens and hundreds of thousands of variables and constraints — a task which is by far out of reach of the theoretically efficient “universal” black box oriented algorithms underlying the Theorem.
Since mid-1970’s, Convex Programming is the most rapidly developing area in Optimization, with intensive and successful research primarily focusing on

- discovery and investigation of novel well-structured generic Convex Programming problems (‘Conic Programming’, especially Conic Quadratic and Semidefinite)

- developing theoretically efficient and powerful in practice algorithms for solving well-structured convex programs, including large-scale nonlinear ones

- building Convex Programming models for a wide spectrum of problems arising in Engineering, Signal Processing, Machine Learning, Statistics, Management, Medicine, etc.

- extending modelling methodologies in order to capture factors like data uncertainty typical for real world situations

- software implementation of novel optimization techniques at academic and industry levels
“Structure-Revealing” Representation of Convex Problem: Conic Programming

♣ When passing from a Linear Programming program

$$\min_x \left\{ c^T x : Ax - b \geq 0 \right\} \quad (\ast)$$

to a nonlinear convex one, the traditional wisdom is to replace linear inequality constraints

$$a_i^T x - b_i \geq 0$$

with nonlinear ones:

$$g_i(x) \geq 0 \quad [g_i \text{ are concave}]$$

♠ There exists, however, another way to introduce nonlinearity, namely, to replace the coordinate-wise vector inequality

$$y \geq z \iff y - z \in \mathbb{R}^m_+ = \{u \in \mathbb{R}^m : u_i \geq 0 \ \forall i\}$$

[\(y, z \in \mathbb{R}^m\)]

with another vector inequality

$$y \geq_K z \iff y - z \in K \quad [y, z \in \mathbb{R}^m]$$

where \(K\) is a regular cone (i.e., closed, pointed and convex cone with a nonempty interior) in \(\mathbb{R}^m\).
♣ **LP problem:**

\[
\min_x \{c^T x : Ax - b \geq 0\} \iff \min_x \{c^T x : Ax - b \in \mathbb{R}_+^m\}
\]

♣ **General Conic problem:**

\[
\min_x \{c^T x : Ax - b \geq_K 0\} \iff \min_x \{c^T x : Ax - b \in K\}
\]

- \((A, b)\) – data of conic problem
- \(K\) - structure of conic problem

♠ **Note:** Every convex problem admits equivalent conic reformulation

♠ **Note:** With conic formulation, convexity is “built in”; with the standard MP formulation convexity should be kept in mind as an additional property.

♣ (??) A general convex cone has no more structure than a general convex function. Why conic reformulation is “structure-revealing”?

♣ (!!) As a matter of fact, just 3 types of cones allow to represent an extremely wide spectrum (“essentially all”) of convex problems!
\[
\min \{ c^T x : Ax - b \geq_0 K \} \iff \min \{ c^T x : Ax - b \in K \}
\]

♠ Three Magic Families of cones:

- \(LP\): Nonnegative orthants \( \mathbb{R}^m_+ \) – direct products of \( m \) nonnegative rays \( \mathbb{R}_+ = \{ s \in \mathbb{R} : s \geq 0 \} \) giving rise to Linear Programming programs
  \[
  \min_{s} \left\{ c^T x : a^T_{\ell} x - b_{\ell} \geq 0, 1 \leq \ell \leq q \right\}.
  \]

- \(CQP\): Direct products of Lorentz cones
  \[
  \mathbb{L}^p_+ = \{ u \in \mathbb{R}^p : u^p \geq \|[u_1; \ldots; u_{p-1}]||_2 \}
  \]
  giving rise to Conic Quadratic programs
  \[
  \min_{x} \left\{ c^T x : \|A_{\ell}x - b_{\ell}\|_2 \leq c^T_{\ell} x - d_{\ell}, 1 \leq \ell \leq q \right\}.
  \]

- \(SDP\): Direct products of Semidefinite cones \( \mathbb{S}^{p}_+ = \{ M \in \mathbb{S}^p : M \succeq 0 \} \)
  giving rise to Semidefinite programs
  \[
  \min_{x} \left\{ c^T x : {A}_{\ell}(x) \succeq 0, 1 \leq \ell \leq q \right\}.
  \]
  where \( \mathbb{S}^{p} \) is the space of \( p \times p \) real symmetric matrices, \( M \succeq 0 \) means that \( M \) is symmetric positive semidefinite, and \( A_{\ell}(x) \) are affine in \( x \) symmetric matrices.

Note: Constraint stating that a symmetric matrix affinely depending on decision variables is \( \succeq 0 \) is called \( \text{LMI} \) – Linear Matrix Inequality.
The nonnegative orthant $\mathbb{R}^3$  

The Lorentz cone $\mathcal{L}^3$

3 random 3D cross-sections of the semidefinite cone $\mathcal{S}^3_+$
Facts:
♠ Three “magic” families of conic problems – $LP, CQP, SDP$ – possess extremely strong “expressive abilities” and for all practical purposes cover the entire Convex Programming
♠ At the same time, the cones underlying the magic families are well understood and possess deep intrinsic mathematical similarity allowing for unified design of theoretically and practically efficient Interior Point polynomial time methods for $LP/CQP/SDP$.
♠ To enjoy the power of “computational toolbox” of $LP/CQP/SDP$, one should reformulate the problem of interest as a conic problem from a “magic” family, and this is where a priori knowledge of problem’s structure is used.
**Fact:** Modern Interior Point Polynomial Time methods for $LP/CQP/SDP$ are the best known so far techniques for finding high accuracy solutions to convex programs – after the program is reformulated as $LP/CQP/SDP$, such a solution usually is found in a moderate (few tens) number of iterations, an iteration reducing to assembling and solving a system of linear equations.

**However:** For extremely large-scale problems, the linear systems arising in Interior Point methods become too large to be solved in reasonable time

$\Rightarrow$ *In the large-scale case, utilizing ”computationally cheap” optimization techniques becomes a must.*

As far as constrained/nonsmooth large-scale convex problems are concerned, the scope of these “computationally cheap” techniques – *First Order algorithms* – is restricted to search for medium-accuracy solutions.
In our course, the emphasis will be on

1. Theory of Conic Programming, primarily, Conic Programming Duality

2. Expressive abilities and typical applications, primarily in Engineering, of Linear, Conic Quadratic, and Semidefinite Programming

3. Polynomial time solvability of Convex Programming and Interior Point algorithms for $LP/CQP/SDP$

4. First Order Algorithms for Large-scale problems with convex structure
I. FROM LINEAR TO CONIC PROGRAMMING
A Linear Programming

\[
\min_x \{ c^T x : Ax \geq b \} \quad [x \in \mathbb{R}^n, A \in \mathbb{R}^{m \times n}]
\]

Aside of modelling and algorithmic issues, the most important issue in LP is **LP Duality Theory**, which, essentially, answers the following basic question:

**(?)** *How to certify that a system of strict and nonstrict linear inequalities

\[
\begin{aligned}
Px &> p \\
Qx &\geq q
\end{aligned}
\]

**(S)** has no solutions?*

Note that it is easy to certify that *(S)* has a solution: **every solution is a certificate**!
**General Theorem on Alternative**

- **Question:** Given a finite system of strict and non-strict linear inequalities with \( n \) unknowns

\[
\begin{align*}
P_x &> p \ (a) \\
Q_x &\ge q \ (b)
\end{align*}
\]

how to certify that the system has *no* solutions?

**Example:** To certify that the system

\[
\begin{align*}
-4u & -9v + 5w > 2 \\
-2u & +6v \quad \ge -2 \\
7u & -5w \quad \ge 1
\end{align*}
\]

has no solutions, it suffices to point out that aggregating the inequalities of the system with weights 2, 3, 2, we get a contradictory inequality:

\[
\begin{align*}
2 \times & \quad -4u \quad -9v \quad +5w > 2 \\
+ & \quad -2u \quad +6v \quad \ge -2 \\
+ & \quad 7u \quad -5w \quad \ge 1 \\
\hline
0 \cdot u \quad +0 \cdot v \quad +0 \cdot w & > 0
\end{align*}
\]
General Theorem on Alternative

• **Question:** Given a finite system of strict and non-strict linear inequalities with $n$ unknowns

\[
\begin{align*}
Px & > p \ (a) \\
Qx & \geq q \ (b)
\end{align*}
\]

(S)

How to certify that the system has no solutions?

• **Simple sufficient condition for insolvability:**
Assume that we can get, as a “linear consequence” of (S) (i.e., by multiplying inequalities (a) by nonnegative weights $s_i$, inequalities (b) by nonnegative weights $y_j$ and adding the results) a *contradictory* (no solutions at all!) inequality:

There exist nonnegative weight vectors $s$ ($\text{dim } s = \text{dim } p$) and $y$ ($\text{dim } y = \text{dim } q$) such that the inequality

\[
[s^T P + y^T Q] x \Omega s^T p + y^T q \quad \Omega = \begin{cases} 
'' > '', & s \neq 0 \\
'' \geq '', & s = 0 
\end{cases}
\]

(*)

with unknowns $x$ has no solutions. Then (S) is infeasible.

1.3
\{P x > p, Q x \geq q\} \& \{s \geq 0, y \geq 0\} \Rightarrow \left[ s^T P + y^T Q \right] x \underbrace{\Omega s^T p + y^T q}_{(\ast)} \quad \Omega = \left\{\begin{array}{l} " > " , \quad s \neq 0 \\ " \geq " , \quad s = 0 \end{array}\right. \\

**Observation:** Inequality \((\ast)\) has no solutions iff \(P^T s + Q^T y = 0\) and

— either \(\Omega = " > " \) and \(s^T p + y^T q \geq 0\),

— or \(\Omega = " \geq " \) and \(s^T p + y^T q > 0\)

We have arrived at

**Proposition.** Given system of strict and nonstrict linear inequalities

\[
\begin{aligned}
P x & > p \\
Q x & \geq q,
\end{aligned}
\]

\((S)\)

let us associate with it the following two systems of linear equalities/inequalities with unknowns \(s,y\):

\[
\begin{array}{|l|}
\hline
T_1: & \begin{cases}
s, y \geq 0; \\
 P^T s + Q^T y = 0; \\
p^T s + q^T y \geq 0; \\
\sum_i s_i > 0.
\end{cases} \\
T_2: & \begin{cases}
y \geq 0; \\
Q^T y = 0; \\
q^T y > 0.
\end{cases}
\hline
\end{array}
\]

If one of the systems \(T_1, T_2\) has a solution, then \((S)\) has no solutions.

**General Theorem on Alternative.** The sufficient condition for infeasibility of \((S)\) stated by Proposition is in fact necessary and sufficient.
$S: \begin{cases} P_x > p \\ Q_x \geq q \end{cases} \quad T_I: \begin{cases} s, y \geq 0; \\ P^T s + Q^T y = 0; \\ p^T s + q^T y \geq 0; \\ \sum_i s_i > 0. \end{cases} \quad T_{II}: \begin{cases} y \geq 0; \\ Q^T y = 0; \\ q^T y > 0. \end{cases}$

**Remark:** By GTA applied to the system $Qx \geq q$, this system is unsolvable iff $T_{II}$ is solvable. Thus,

- **System** $(S_{NS})$ is unsolvable iff system $T_{II}$ is solvable;
- **Assume that system** $(S_{NS})$ is solvable. Then system $(S)$ is unsolvable iff system $T_I$ is solvable.
Corollaries: A. A system of linear inequalities

\[
a_i^T x \leq b_i, \; i = 1, \ldots, m
\]

is infeasible iff one can combine the inequalities of the system in a legitimate linear fashion (i.e., multiply the inequalities by weights and add the results, the sign of the weights making the summation legitimate) to get a contradictory inequality, namely, either the inequality \(0^T x \geq 1\), or the inequality \(0^T x > 0\).

B. [Inhomogeneous Farkas Lemma] A scalar linear inequality \(a_0^T x \leq b_0\) is a consequence of a solvable system of linear inequalities

\[
a_i^T x \leq b_i, \; i = 1, \ldots, m
\]

iff it can be obtained by taking weighted sum, with nonnegative weights, of inequalities from the system and the trivial identically true inequality \(0 \leq 1\):

\[
a_0 = \sum_{i=1}^{m} \lambda_i a_i, \; b_0 = \lambda_0 + \sum_i \lambda_i b_i \text{ for some } \lambda_i \geq 0, \; i = 0, 1, \ldots, m
\]
GTA is a really striking fact:

\[
\begin{align*}
-1 \leq u \leq 1 & \Rightarrow \begin{cases} u^2 \leq 1 & \Rightarrow u^2 + v^2 \leq 2 \\ v^2 \leq 1 \end{cases} \\
-1 \leq v \leq 1 & \Rightarrow u + v = 1 \times u + 1 \times v \leq \sqrt{1^2 + 1^2} \sqrt{u^2 + v^2} \leq \sqrt{2} \times \sqrt{2} = 2 \Rightarrow u + v \leq 2
\end{align*}
\]

In this “highly nonlinear” derivation, the premise is a solvable system of linear inequalities, and the conclusion is a linear inequality. How could we know in advance that every derivation of this type can be replaced just with linear aggregation of the inequalities in the premise and the trivial inequality \(0 \leq 1\)?

GTA heavily exploits the fact that we are speaking about linear inequalities:

\[
\begin{align*}
u \leq 1 & \Rightarrow u^2 \leq 1 & \text{— definitely true!} \\
-u \leq 1 & \Rightarrow -u^2 \leq 0 \\
& \Rightarrow -u^2 + 2u \leq 0, -u^2 + 2u \leq 1, ...
\end{align*}
\]

However, aggregating in a legitimate linear fashion inequalities from the premise and trivial (i.e., identically true) linear and quadratic inequalities, like

\[
0 \leq 1, -u^2 \leq 0, -u^2 + 2u \leq 1, ...
\]

you cannot get the concluding inequality.

1.7
Starting point: Homogeneous Farkas Lemma: A homogeneous linear inequality

\[ a^T x \geq 0 \]  \hspace{1cm} (I)

is a consequence of a system of homogeneous linear inequalities

\[ a_i^T x \geq 0, \ i = 1, \ldots, m, \]  \hspace{1cm} (H)

iff (I) can be obtained from (H) by linear aggregation:

\[ \exists y \geq 0 : a = \sum_{i} y_i a_i, \]

that is, iff \( a \) is a conic combination (linear combination with nonnegative coefficients) of \( a_1, \ldots, a_m \).
HFL ⇒ GTA: Given system

\[
\begin{align*}
Px & > p \\
Qx & \geq q
\end{align*}
\]

(S)

in variables \(x\), we associate with it system

\[
\begin{align*}
Px - tp - \epsilon & \geq 0 \\
Qx - tq & \geq 0 \\
t - \epsilon & \geq 0
\end{align*}
\]

(H)

in variables \(x, t, \epsilon\).

It is immediately seen that (S) has no solutions iff (H) has no solutions with \(\epsilon > 0\), i.e., iff the homogeneous linear inequality \(-\epsilon \geq 0\) is a consequence of the system of homogeneous linear inequalities (H). HFL says exactly when the latter happens, and this answer turns out to be exactly the statement of GTA.
HFL – Intelligent Proof

❖ A set $X \subset \mathbb{R}^n$ is called **polyhedral**, if it is a solution set of a finite system of nonstrict linear inequalities:

$$X \text{ is polyhedral} \iff \exists A, b : X = \{ x \in \mathbb{R}^n : Ax \leq b \}.$$  

❖ A **polyhedral representation** of a set $X \subset \mathbb{R}^n_x$ is a representation of $X$ as the projection of a polyhedral set

$$X^+ = \{ [x; u] : Ax + Bu \leq b \} \subset \mathbb{R}^n_x \times \mathbb{R}^k_u$$

under the projection $[x; u] \mapsto x : \mathbb{R}^n_x \times \mathbb{R}^k_u \rightarrow \mathbb{R}^n_x$:

$$X = \{ x \in \mathbb{R}^n : \exists u : [x; u] \in X^+ \}.$$
Fact: A set is polyhedral iff it admits polyhedral representation, or, equivalently, the projection $X$ of a polyhedral set

$$X^+ = \{[x; u] : Ax + Bu \leq c\}$$
on the space of $x$-variables can be represented as a solution set to a finite system of nonstrict linear inequalities in $x$-variables only.

Proof [Fourier-Motzkin Elimination]: It suffices to consider the case when $u$ is one-dimensional. Let us split all inequalities $a_i^T x + b_i u \leq c_i$, $1 \leq i \leq I$, into three groups:

- black: $b_i = 0$ ($i \in \text{Black}$). Black inequality says that $a_i^T x \leq c_i$;
- red: $b_i > 0$ ($i \in \text{Red}$). Red inequality says that $u \leq \alpha_i^T x + \beta_i$, i.e., it imposes an affine in $x$ lower bound on $u$.
- green: $b_i < 0$ ($i \in \text{Green}$). Green inequality says that $u \geq \alpha_i^T x + \beta_i$, $1 \leq i \leq I$, i.e., it imposes an affine in $x$ upper bound on $u$.

Observe that a vector $\bar{x}$ belongs to the projection of $X^+$ on the $x$-plane iff $\bar{x}$ satisfies all black inequalities $a_i^T \bar{x} \leq c_i \forall i \in \text{Black}$ and we can points out a real which meets all stemming from $\bar{x}$ upper and lower bounds on $u$, i.e.,

$$X := \{x : \exists u : Ax + ub \leq c\} = \left\{ x : \begin{cases} a_i^T x \leq c_i \forall i \in \text{Black} \\ \alpha_i^T x + b_i \geq \alpha_j^T x + \beta_j \forall (i \in \text{Red}, j \in \text{Green}) \end{cases} \right\}$$

and $X$ indeed is polyhedral.
Now we are ready to prove HFL. The only nontrivial part of the statement is *If $a$ is not a conic combination of $a_1, ..., a_n$, then $a^T d < 0$ for some $d$ with $a_i^T d \geq 0$, $i = 1, ..., n$.*

**Proof:** Let $a \not\in \text{Cone}(a_1, ..., a_n) = \left\{ \sum_{i=1}^n u_i a_i : u \geq 0 \right\}$. Observe that Cone$(a_1, ..., a_n)$ admits polyhedral representation:

$$\text{Cone}(a_1, ..., a_n) = \left\{ x : \exists u : \begin{array}{c} u \geq 0, \\ x - \sum_i u_i a_i = 0 \end{array} \right\}$$

By the above, Cone$(a_1, ..., a_n)$ is polyhedral: there exists a finite system of inequalities $p_j^T x \geq b_j$, $1 \leq j \leq J$, such that

$$\text{Cone}(a_1, ..., a_n) = \{ x : p_j^T x \geq q_j \}.$$

- Since $0 \in \text{Cone}(a_1, ..., a_n)$, we have $q_j \leq 0$ for all $j$;
- Since $a \not\in \text{Cone}(a_1, ..., a_n)$, we have $p_j^T a < q_{j^*}$ for some $j^*$, whence $p_{j^*}^T a < 0$;
- since $ta_i \in \text{Cone}(a_1, ..., a_n)$ for all $i$ and all $t > 0$, we should have $p_{j^*}^T (ta_i) \geq q_{j^*}$ for all $t > 0$, whence $p_{j^*}^T a_i \geq 0$ for all $i = 1, ..., n$.

$\Rightarrow$ with $d = p_{j^*}$ we have $a_i^T d \geq 0$ for all $i$ and $a^T d < 0$, as required.

1.12
Dual to a Linear Programming program

**Question:** When a real $a$ is a lower bound on the optimal value of an LP program

$$\min_x \left\{ c^T x : Ax - b \geq 0 \right\} \quad ? \quad (P)$$

**Answer:** We are asking when the linear inequality

$$c^T x \geq a$$

is a corollary of the finite system of linear inequalities

$$Ax \geq b.$$  

A *sufficient* condition for this is the possibility to get the target inequality by aggregation, with nonnegative weights, of the inequalities from the system and identically true inequality $1 \geq 0$:

$$\exists y \geq 0 : \quad A^T y = c, \quad y^T b \geq a$$

This sufficient condition is also *necessary*, provided that $(P)$ is feasible (Corollary B of GTA).
\[
\min_x \{ c^T x : Ax - b \geq 0 \} \quad (P)
\]

**Conclusion:** The optimal value in the optimization problem

\[
\max_y \{ b^T y : A^T y = c, \ y \geq 0 \} \quad (D)
\]

is a lower bound on the optimal value in \((P)\). If the optimal value in \((P)\) is finite, then \((D)\) is solvable, and

\[
\text{Opt}(P) = \text{Opt}(D).
\]
LP Duality Theorem. Consider an LP program
\[
\min_x \left\{ c^T x : Ax \geq b \right\} \quad (P)
\]
(the “primal” problem) along with its dual
\[
\max_y \left\{ b^T y : A^T y = c, y \geq 0 \right\} \quad (D)
\]
Then
- The duality is symmetric: the problem dual to dual is equivalent to the primal;
- The value of the dual objective at every dual feasible solution is \(\leq\) the value of the primal objective at every primal feasible solution
- The following 5 properties are equivalent to each other:
  (i) The primal is feasible and below bounded.
  (ii) The dual is feasible and above bounded.
  (iii) The primal is solvable.
  (iv) The dual is solvable.
  (v) Both primal and dual are feasible.
Whenever (i) \(\equiv\) (ii) \(\equiv\) (iii) \(\equiv\) (iv) \(\equiv\) (v) is the case, the optimal values in the primal and the dual problems are equal to each other:
\[
\text{Opt}(P) = \text{Opt}(D).
\]
\[
\min \{ c^T x : Ax \geq b \} \quad (P)
\]
\[
\max \{ b^T y : A^T y = c, y \geq 0 \} \quad (D)
\]

**Corollary.** [Necessary and sufficient optimality conditions in LP] Consider an LP program \((P)\) along with its dual \((D)\), and let \((x, y)\) be a pair of primal and dual feasible solutions. The pair is comprised of optimal solutions to the respective problems iff

\[
c^T x - b^T y = 0 \quad \text{[zero duality gap]}
\]
as well as iff

\[
y_i [Ax - b]_i = 0, \quad i = 1, ..., m, \quad \text{[complementary slackness]}
\]

Indeed, since \((P)\) and \((D)\) are feasible, they are solvable with equal optimal values, hence for primal-dual feasible \((x, y)\)

\[
\text{DualityGap}(x, y) \equiv c^T x - b^T y = \underbrace{c^T x - \text{Opt}(P)}_{\geq 0} + \underbrace{\text{Opt}(D) - b^T y}_{\geq 0}
\]
is always nonnegative and is 0 iff \(x, y\) are optimal for the respective problems. Next, for a primal-dual feasible \((x, y)\) we have

\[
\text{DualityGap}(x, y) = c^T x - b^T y = (A^T y)^T x - b^T y = [Ax - b]^T y
\]

\[
\Rightarrow c^T x - b^T y = 0 \iff \underbrace{(Ax - b)^T}_{\geq 0} \underbrace{y}_{\geq 0} = 0 \iff y_i [Ax - b]_i = 0 \forall i.
\]
The basic problem of Signal Processing is as follows: (??) “In the nature” there exists a signal represented by vector $x \in \mathbb{R}^n$. Given observation

$$y = Ax + \eta$$

- $A$: $m \times n$ sensing matrix
- $\eta$: observation noise

we want to recover $x$.

There are many different approaches to (??), depending primarily on the relation between $m$ and $n$ and on a priori information on $x$:

**Parametric case:** $m \gg n$: in principle, no a priori information on $x$ is needed. In the “no noise” case $\eta = 0$ and with a “general position” $A$, $x$ is readily given by $y$. When $\eta \neq 0$, the challenge is to reduce the influence of the noise on the estimate. A typical estimate is the Least Squares one:

$$\hat{x}(y) \in \text{Argmin}_{w \in \mathbb{R}^n} \| Aw - y \|_2^2.$$ 

Least Squares are commonly used when $\eta = \sigma \xi$, $\xi \sim \mathcal{N}(0, I_m)$.

**Nonparametric case:** $m \ll n$: In the “no noise” case $\eta = 0$ the equality $y = Ax$ does not define $x$ uniquely

$\Rightarrow$ A priori information on $x$ is needed!

— In Compressed Sensing, a priori information is that $x$ is sparse — has at most a given number $s \ll m$ of nonzero entries.
Fact: Many real-life signals $x$ when presented by their coefficients in properly selected basis ("dictionary") $B$:

$$x = Bu$$

- columns of $B$: vectors of basis $B$
- $u$: coefficients of $x$ in basis $B$

become sparse (or nearly so): $u$ has just $s \ll n$ nonzero entries (or can be well approximated by vector with $s \ll n$ nonzero entries). We do not assume the location of “meaningful coefficients” known in advance.
**Example I:** Typical audio signals become sparse (or nearly so) when representing them "in frequency domain" – as sums of harmonic oscillations of different frequencies:

Top: signal in time domain
Bottom: decomposition of signal into sum of harmonic oscillations
Illustration: 25 sec fragment of audio signal “Mail must go through” (dimension 1,058,400) and its "Fourier coefficients" – amplitudes of participating harmonic oscillations vs. the frequencies:

<table>
<thead>
<tr>
<th>% of leading Fourier coefficients kept</th>
<th>energy</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>25%</td>
<td>99.8%</td>
</tr>
<tr>
<td>15%</td>
<td>99.6%</td>
</tr>
<tr>
<td>5%</td>
<td>98.2%</td>
</tr>
<tr>
<td>1%</td>
<td>79.0%</td>
</tr>
</tbody>
</table>
**Example II:** The $256 \times 256$ image can be thought of as $256^2 = 65536$-dimensional vector (write down the intensities of pixels column by column). This image (same as other “non-pathological” images) is nearly sparse when represented in wavelet basis:

<table>
<thead>
<tr>
<th>Percentage of Coefficients Kept</th>
<th>Energy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>99.70%</td>
</tr>
<tr>
<td>5%</td>
<td>99.93%</td>
</tr>
<tr>
<td>10%</td>
<td>99.96%</td>
</tr>
<tr>
<td>25%</td>
<td>99.99%</td>
</tr>
</tbody>
</table>

1.21
When recovering a signal $x_*$ admitting a sparse (or nearly so) representation $Bu_*$ in a known basis $B$ from observations

$$y = Ax_* + \eta,$$

the situation reduces to the one when the signal to be recovered is just sparse. Indeed, we can first recover sparse $u_*$ from observations

$$y = Ax_* + \eta = [AB]u_* + \eta.$$  

After an estimate $\hat{u}$ of $u_*$ is built, we can estimate $x_*$ by $B\hat{u}$.

$\Rightarrow$ In fact, sparse recovery is about how to recover a sparse $n$-dimensional signal $x$ from $m \ll n$ observations

$$y = Ax_* + \eta.$$
\[ y = Ax + \eta, \|\eta\| \leq \delta, \|x\|_0 := \text{Card}\{i : x_i \neq 0\} \leq s \quad \therefore \quad \hat{x} \approx x \]

Let \( \delta = 0 \). When the number \( s \) of nonzero entries in \( x \in \mathbb{R}^n \) is essentially smaller than the number \( m = \dim y \) of observations, the recovery problem becomes well-posed and can be solved by, e.g., \( \ell_0 \) minimization:

\[ \hat{x} \in \text{Argmin}_{w \in \mathbb{R}^n} \{\|w\|_0 : Aw = y\} \]

**Simple fact:** Let every \( m \times 2s \) submatrix of the \( m \times n \) matrix \( A \) be of rank \( 2s \) (which is the case for a “general position” matrix \( A \), provided that \( 2s \leq \min[m,n] \)). Then in the noiseless case the \( \ell_0 \) minimization recovers exactly every \( s \)-sparse signal \( x \).

Indeed, \( x \) is feasible for the minimization problem \( \Rightarrow \|\hat{x}\|_0 \leq \|x\|_0 \leq s \Rightarrow \|x - \hat{x}\|_0 \leq 2s \), which combines with \( A(x - \hat{x}) = 0 \) and the assumption that every \( 2s \) columns of \( A \) are linearly independent to imply \( x - \hat{x} = 0 \).

**Bad news:** \( \ell_0 \) minimization requires to solve a disastrously complex combinatorial problem and as such is completely impractical.

**A remedy:** let us replace minimizing nonconvex (and even discontinuous) \( \|\cdot\|_0 \) with minimizing the “closest” to \( \|\cdot\| \) convex function \( \|\cdot\|_1 \), thus arriving at \( \ell_1 \) minimization, which in the noiseless case is

\[ \hat{x}(y) \in \text{Argmin}_{w \in \mathbb{R}^n} \{\|w\|_1 : Aw = y\}. \]

\[ \|z\|_1 = \sum_i |z_i| \]

Extensions of \( \ell_1 \) minimization to the case of noisy observation take different forms, depending on noise’s structure. For example, in the case of uncertain-but-bounded noise, where all we know is that \( \|\eta\| \leq \delta \), \( \|\cdot\| \) and \( \delta \) being given, a natural version of \( \ell_1 \) minimization is

\[ \hat{x}(y) \in \text{Argmin}_w \{\|w\|_1 : \|Aw - y\| \leq \delta\} . \]
\[ y = Ax + \eta, \|\eta\| \leq \delta \Rightarrow \hat{x}(y) \in \operatorname{Argmin}_{w \in \mathbb{R}^n} \{\|w\|_1 : \|Aw - y\| \leq \delta\} \]

**Note:** When \( \delta = 0 \), same as when \( \|w\| = \|w\|_\infty := \max_i |w_i| \), \( \ell_1 \) recovery reduces to solving an LP program!

**Basic questions:**

**A.** When \( A \) is \( s \)-good, that is, when \( \ell_1 \)-recovery in the noiseless case \( \delta = 0 \) recovers exactly every \( s \)-sparse signal \( x \)?

**B.** For \( s \)-good \( A \), what are the error bounds of \( \ell_1 \) recovery in the presence of noise?
A. When $A$ is $s$-good, that is, when $\ell_1$-recovery in the noiseless case $\delta = 0$ recovers exactly every $s$-sparse signal $x$?

**Answer to A** can be straightforwardly extracted from LP optimality conditions and is as follows:

(!) $A$ is $s$-good iff the nullspace property takes place: for every subset $I$ of cardinality $s$ of the index set $\{1,\ldots,n\}$ and for every $z \in \ker A\setminus\{0\}$ one has

$$\|z_I\|_1 < \frac{1}{2} \|z\|_1,$$

where $z_I$ is obtained from $z$ by keeping intact all entries with indexes from $I$ and zeroing out entries with indexes not in $I$.

**Only if:** Assume that for some $I$, $\text{Card}(I) \leq s$, and some nonzero $z \in \ker A$, one has $\|z_I\|_1 \geq \frac{1}{2} \|z\|_1$, or, equivalently, $\|z_I\|_1 \geq \|z_J\|_1$, $J = \{1,\ldots,n\} \setminus I$, and let us prove that $A$ is not $s$-good. Let the true signal be the $s$-sparse signal $x = z_I$. Then

$$Az = 0 \Rightarrow Ax = A[-z_J] \& \|z_J\|_1 \leq \|z_I\|_1 = \|x\|_1$$

$$\Rightarrow x \text{ is not the unique optimal solution to } \min_w \{\|w\|_1 : Aw = Ax\}$$

$$\Rightarrow A \text{ is not } s\text{-good.}$$

**If:** Let the nullspace property take place, let $x$ be $s$-sparse, so that $x = x_I$ for some $I$, $\text{Card}(I) \leq s$, and let $\hat{x} \in \text{Argmin}_w \{\|w\|_1 : Aw = Ax\}$. Let $J = \{1,\ldots,n\} \setminus I$ and $z = \hat{x} - x$. Assuming $z \neq 0$, let us lead this assumption to a contradiction. Since $0 \neq z \in \ker A$, we have by nullspace property $\|z_I\|_1 < \|z_J\|_1$, so that

$$\|x_I\|_1 - \|\hat{x}_I\|_1 \leq \|z_I\|_1 < \|z_J\|_1 = \|\hat{x}_J\| \Rightarrow [\|x\|_1 =] \|x_I\|_1 < \|\hat{x}\|_1$$

and the concluding inequality contradicts the origin of $\hat{x}$.  

1.25
B. For s-good $A$, what are the error bounds of $\ell_1$ recovery in the presence of noise? Let us set
\[
\|x\|_{s,1} := \max_{I: \text{Card}(I) \leq s} \|x_I\|_1 = \max_u \{u^T x : \|u\|_{\infty} \leq 1, \|u\|_1 \leq s\}
\]

Note: (!) is due to the evident fact that for a positive integer $s \leq n$, the extreme points of the convex polytope
\[
U_s = \{u \in \mathbb{R}^n : \|u\|_{\infty} \leq 1, \sum_i |u_i| \leq s\}
\]
are exactly the vectors with $s$ nonzero entries equal to $\pm 1$.

Observation: $A$ is s-good iff the quantity
\[
\kappa_s(A) = \max_x \{\|x\|_{s,1} : Ax = 0, \|x\|_1 \leq 1\} = \max_{x,u} \{u^T x : u \in U_s, Ax = 0, \|x\|_1 \leq 1\}
\]
is $< 1/2$.
Indeed, the nullspace property says that $\|x_I\|_1 < \frac{1}{2}\|x\|_1$ for all $0 \neq x \in \text{Ker}A$ and every $I$ with $\text{Card}(I) \leq s$, or, which is the same as $\|x\|_{s,1} < 1/2$ whenever $x \in \text{Ker}A$ and $\|x\|_1 \leq 1$.

Observation: For every integer $s \leq n$, every $m \times n$ matrix $A$ and every norm $\| \cdot \|$ on the image space $\mathbb{R}^m$ of $A$ there exists $\beta < \infty$ such that
\[
\forall x \in \mathbb{R}^n : \|x\|_{s,1} \leq \beta \|Ax\| + \kappa_s(A)\|x\|_1.
\]

(*)
The infimum of $\beta$'s satisfying this property will be denoted $\beta_s(A, \| \cdot \|)$.

Indeed, let $P$ be orthogonal projector on $\text{Ker}A$. For some $\alpha < \infty$ and all $z$ we have
\[
\|(I - P)z\|_1 \leq \alpha\|(I - P)z\|, \text{ whence }
\]
\[
\|z\|_{s,1} \leq \|(I - P)z\|_{s,1} + \|Pz\|_{s,1} \leq \|(I - P)z\|_1 + \kappa_s(A)\|Pz\|_1 \leq \|(I - P)z\|_1 + \kappa_s(A)(\|z\|_1 + \|(I - P)z\|_1)
\]
\[
\leq (1 + \kappa_s(A))\|(I - P)z\|_1 + \kappa_s(A)\|z\|_1 \leq \alpha(1 + \kappa_s(A))\|A(I - P)z\|_1 + \kappa_s(A)\|z\|_1
\]
\[
= \frac{\alpha(1 + \kappa_s(A))\|Az\| + \kappa_s(A)\|z\|_1}{\beta}
\]

Note: (*) together $\kappa_s(A) < 1/2$ implies nullspace property.
\[
\forall z \in \mathbb{R}^n : \|z\|_{s, 1} \leq \beta \|Az\| + \kappa_s(A) \|z\|_1. 
\]

\[\text{(*)}\]

\[\oplus\] The quantities \(\kappa_s(A)\) and \(\beta_s(a, \|\cdot\|)\) are responsible for the error bound in imperfect \(\ell_1\) recovery:

**Theorem.** Let \(A\) be \(m \times n\) sensing matrix and \(s\) be a positive integer. Assume that
\* signal \(x \in \mathbb{R}^n\) is nearly \(s\)-sparse: \(\|x - x^s\|_1 \leq \nu\) for some \(s\)-sparse vector \(x^s\);
\* noise \(\eta\) in the observation \(y = Ax + \eta\) satisfies \(\|\eta\| \leq \delta\) for given \(\delta \geq 0\) and norm \(\|\cdot\|\);
\* \(\hat{x}\) is obtained from \(A, y, \delta\) by imperfect \(\ell_1\)-recovery:

\[
\|\hat{x}\|_1 \leq \nu + \min_w \{\|w\|_1 : \|Aw - y\| \leq \delta\} \quad \& \quad \|A\hat{x} - y\| \leq \delta + \epsilon.
\]

**Assuming (\*) and \(\kappa_s(A) < 1/2\), the following error bound holds true:**

\[
\|x - \hat{x}\|_1 \leq \frac{2\beta_s(A, \|\cdot\|)[2\delta + \epsilon] + 2\nu + \nu}{1 - 2\kappa_s(A)}.
\]

**Proof.** W.l.o.g. we can take \(x^s = x_I\), where \(I\) is the collection of indexes of the \(s\) largest in magnitude entries in \(x\), and \(x_I\) is obtained from \(x\) by zeroing out the entries with indexes outside of \(I\). Let \(J = \{1, \ldots, n\} \setminus I\) and \(z = \hat{x} - x\), so that \(\|x_J\|_1 = \nu\). Setting \(\kappa = \kappa_s(A), \beta = \beta_s(A, \|\cdot\|),\) have

\[
\begin{align*}
(a) \quad & \|x\|_1 \leq \text{Opt} + \nu \leq \|x_I\|_1 + \nu = \|x_I\|_1 + \|x_J\|_1 + \nu, \\
(b) \quad & \|Az\| \leq \|[A\hat{x} - y] + [y - Ax]\| \leq \|A\hat{x} - y\| + \|Ax - y\| \leq 2\delta + \epsilon, \\
(c) \quad & \|x_J\|_1 - \|x_I\|_1 \leq \|\hat{x}\|_1 - \|x_I\|_1 - \|x_J\|_1 \leq \nu + \|x_I\|_1 - \|x_I\|_1 \leq \nu + \|z_I\|_1; \quad \text{[by (a)]} \\
(d) \quad & \|z_I\|_1 \leq \beta\|Az\| + \kappa\|z_I\|_1 = \beta\|Az\| + \kappa[\|z_I\|_1 + \|z_J\|_1] \\
\Rightarrow \quad & \|z_I\|_1 \leq \beta\frac{\|Az\|_1}{1-\kappa} + \frac{\kappa}{1-\kappa}\|z_I\|_1 \leq \beta\frac{2\delta + \epsilon}{1-\kappa} + \kappa\|z_J\|_1; \quad \text{[see (b)]} \\
(e) \quad & \|z\|_1 \leq \beta\frac{\|Az\|_1}{1-\kappa} + \frac{\kappa}{1-\kappa}\|z_J\|_1 \leq \beta\frac{2\delta + \epsilon}{1-\kappa} + \frac{\kappa}{1-\kappa}\|z_J\|_1. \quad \text{[by (d)]}
\end{align*}
\]
We have

\[ \|\hat{x}_J\|_1 - \|x_J\|_1 \leq \nu + \|z_I\|_1 \leq \nu + \beta(2\delta + \epsilon) + \frac{\kappa}{1 - \kappa} \|z_J\|_1 \leq \nu + \beta(2\delta + \epsilon) + \frac{\kappa}{1 - \kappa} \left[\|x_J\|_1 + \|\hat{x}_J\|_1\right] \]

\[ \Rightarrow \frac{1 - 2\kappa}{1 - \kappa} \|x_J\|_1 \leq \nu + \frac{\beta(2\delta + \epsilon)}{1 - \kappa} + \frac{1}{1 - \kappa} \|x_J\|_1 \Rightarrow \frac{1 - 2\kappa}{1 - \kappa} \|z_J\|_1 \leq \nu + \frac{\beta(2\delta + \epsilon)}{1 - \kappa} + 2\|x_J\|_1 \]

\[ \Rightarrow \|z_J\|_1 \leq \nu(1 - \kappa) + \beta(2\delta + \epsilon) + 2(1 - \kappa)\|x_J\|_1 \]

\[ \Rightarrow \|z_J\|_1 \leq \nu(1 - \kappa) + \beta(2\delta + \epsilon) + 2(1 - \kappa)\|x_J\|_1 \]

Invoking (e), we arrive at the desired bound

\[ \|x - \hat{x}\|_z \leq \frac{2\beta_s(A, \|\cdot\|)[2\delta + \epsilon] + 2\nu + \nu}{1 - 2\kappa_s(A)}. \]
We have defined the quantities $\kappa_s(A)$, $\beta_s(A,\|\cdot\|)$ responsible for $s$-goodness of $A$ and for the error bound for imperfect $\ell_1$ recovery. But: It is unclear how to compute efficiently $\kappa_s(A)$. Moreover, no ways to verify the nullspace property in reasonable time are known, unless $s$ is "very small," like 1 or 2. ⇒ We need verifiable sufficient conditions for $s$-goodness, or, which is basically the same, an efficiently computable upper bound $\kappa_s^+(A)$ on the quantity

$$\kappa_s(A) = \max_{u,x} \{ u^T x : \|u\|_\infty \leq 1, \|u\|_1 \leq s, \|x\|_1 \leq 1, Ax = 0 \};$$

Equipped with such a bound, we could use the verifiable condition $\kappa_s^+(A) < 1/2$ as a sufficient condition for $s$-goodness of $A$.

Computationally Efficient Upper-Bounding of $\kappa_s(A)$: For $H \in \mathbb{R}^{m \times n}$ we have

$$\kappa_s(A) := \max_{u,x} \{ u^T x : \|u\|_\infty \leq 1, \|u\|_1 \leq s, \|x\|_1 \leq 1, Ax = 0 \}$$

$$= \max_{u,x} \{ u^T [x - HTAx] : \|u\|_\infty \leq 1, \|u\|_1 \leq s, \|x\|_1 \leq 1, Ax = 0 \}$$

$$\leq \max_{u,x} \{ u^T [I - HTA]x : \|u\|_\infty \leq 1, \|u\|_1 \leq s, \|x\|_1 \leq 1 \}$$

$$= \max_{u,j} \{ u^T \text{Col}_j[I - HTA] : u \in U_s \}$$

$$= \max_j \|\text{Col}_j[I - HTA]\|_{s,1}$$

⇒ The efficiently computable quantity

$$\kappa_s^+(A) = \min_{H \in \mathbb{R}^{m \times n}} \max_j \|\text{Col}_j[I - HTA]\|_{s,1}$$

is an upper bound on $\kappa_s(A)$, and thus the efficiently verifiable condition $\kappa_s^+(A) < 1/2$ is sufficient for $s$-goodness of $A$. 1.29
What is inside

Observation: \( \kappa_s(A) \) is the maximum of convex function \( \|u\|_{s,1} \) on the polytope

\[
X = \text{Conv}\{\pm e_1, \ldots, \pm e_n\} \cap \{x : Ax = 0\} = \{x : Ax = 0, \|x\|_1 \leq 1\}.
\]

A recipe for upper-bounding a convex function \( \phi(x) \) over polytope

\[
X = \text{Conv}\{f_1, \ldots, f_N\} \cap \{x : Ax = 0\}
\]

which we used is as follows: 

For every \( H \in \mathbb{R}^{m \times n} \) we have

\[
\phi_* := \max_{x \in X} \phi(x) = \max \{ \phi(x) : x \in \text{Conv}\{f_1, \ldots, f_N\}, Ax = 0 \}
\]

\[
= \max \{ \phi([I - A^T]A)x) : x \in \text{Conv}\{f_1, \ldots, f_N\}, Ax = 0 \} \leq \max_x \{ \phi([I - H^T]A)x) : x \in \text{Conv}\{f_1, \ldots, f_N\} \}
\]

\[
= \max_{j \leq N} \phi([I - H^T]f_j)
\]

\[
\Rightarrow \phi_* \leq \phi_*^+ := \min_H \left[ \max_j \phi([I - H^T]f_j) \right],
\]

and \( \phi_*^+ \) is efficiently computable – this is the optimal value in explicit convex optimization problem.
Two birds with one stone

Assume that we can certify $s$-goodness of $A$ by the above verifiable sufficient condition, that is, we have at our disposal a matrix $H$ such that

$$\kappa^+ := \max_j \|\text{Col}_j[\Delta]\|_{s,1} < 1/2, \quad \Delta = I - H^T A$$

Then for every $x \in \mathbb{R}^n$ we have $x = [\Delta + H^T A]x$, whence

$$\|x\|_{s,1} \leq \|H^T A x\|_{s,1} + \|\Delta x\|_{s,1} \leq s \|H^T A x\|_\infty + \sum_{j=1}^n |x_j| \|\text{Col}_j(\Delta)\|_{s,1}$$

$$\leq \beta \|Ax\| + \kappa^+ \|x\|_1, \quad \beta = s \max_j \|\text{Col}_j[H]\|_*$$

$$\|f\|_* = \max_{\|u\| \leq 1} f^T u$$

We arrive at

$$\kappa_s(A) \leq \kappa^+ < \frac{1}{2} \quad \text{and} \quad \beta_s(A, \|\cdot\|) \leq s \max_j \|\text{Col}_j[H]\|_*.$$
Remarks:

A. Computing $\kappa_s^+(A)$ and the associated $H$ reduces to LP.
Indeed, for $z \in \mathbb{R}^n$ we have

$$\|z\|_{s,1} = \max_u \{ z^T u : \|u\|_{\infty} \leq 1, \|u\|_1 \leq s \}$$

$$= \min_{y,t} \left\{ st + \sum_{j=1}^n y_j : y \geq 0, |z_j| \leq y_j + t \forall j \right\} \quad \text{[LP duality]}$$

$$\Rightarrow \kappa_s^+(A) := \min_{H,\tau} \{ \tau : \|\text{Col}_j[I - H^T A]\|_{s,1} \leq \tau \}$$

$$= \min_{y^1,\ldots,y^n,t_1,\ldots,t_n,H,\tau} \left\{ \tau : \begin{cases} -y^j - t_j 1 \leq \text{Col}_j[I - H^T A] \leq y^j + t_j 1, 1 \leq j \leq n \\ y^j \geq 0, \sum_{i=1}^n y^j_i + s \tau \leq \tau, 1 \leq j \leq n \end{cases} \right\}$$

B. One has

$$\kappa_1^+(A) = \kappa_1(A) = \max_{j \leq n} \max_x \{ x_j : Ax = 0, \|x\| \leq 1 \} = \min_H \max_{i,j} |(I - H^T A)_{ij}|$$

where the concluding equality is due to LP Duality Theorem.

C. Let $H$ certify $\kappa_p^+(A)$: $\kappa_p^+(A) = \max_j \|\text{Col}_j[I - H^T A]\|_{p,1}$. Since $\|u\|_{s,1} \leq \frac{s}{p}\|u\|_{p,1}$ whenever $p \leq s$, $H$ certifies that

$$\kappa_s^+(A) \leq \frac{s}{p}\kappa_p^+(A), \ p \leq s$$

In particular,

$$\kappa_1^+(A) < \frac{1}{2s} \Rightarrow \kappa_s^+(A) \leq s\kappa_1^+(A) < \frac{1}{2} \Rightarrow A \text{ is } s\text{-good}$$
Mutual Incoherence of $A = [A_1, ..., A_n]$ is defined as

$$\mu(A) = \max_{i \neq j} |A_i^T A_j|/A_j^T A_j.$$  

Setting $H = \frac{1}{1+\mu(A)} [A_1/(A_1^T A_1), A_2/(A_2^T A_2), ..., A_n/(A_n^T A_n)]$:

— diagonal entries in $H^T A$ are $\frac{1}{1+\mu(A)}$,

— magnitudes of off-diagonal entries in $H^T A$ are $\leq \frac{\mu(A)}{1+\mu(A)}$.

$\Rightarrow H$ certifies that $\kappa_1^+(A) \leq \frac{\mu(A)}{\mu(A)+1} \Rightarrow A$ is $s$-good whenever $\frac{2s\mu(A)}{\mu(A)+1} < 1$.

**Note:** When entries of $A$ are drawn at random from $\mathcal{N}(0,1)$ or from Uniform$\{-1, 1\}$, the typical value of $\mu(A)$ is as small as $O(1)\sqrt{\ln(n)/m}$

$\Rightarrow$ our simplified verifiable sufficient condition for $s$-goodness “$\kappa_1^+(A) < \frac{1}{2s}$” certifies that typical $A$ from the above random ensembles is $O(\sqrt{m/\ln(n)})$-good.

1.33
**Bad news:** When $A$ is “essentially non-square,” namely, $n \geq 2m$, our verifiable sufficient condition can certify $s$-goodness only when $s \leq O(1)\sqrt{m}$.

Indeed, assume that $n \geq 2m$ and $H$ certifies that $\kappa_s^+(A) < 1/2$. Setting $\bar{n} = 2m$ and denoting by $D$ the angular $\bar{n} \times \bar{n}$ submatrix of $H^TA$, we have $\text{Rank}D \leq m$, whence $I_{\bar{n}} - D$ has at least $\bar{n} - m \geq m$ singular values $\geq 1$ and thus

$$\sum_{i,j=1}^{\bar{n}} [I_{\bar{n}} - D]_{ij}^2 \geq m.$$ 

On the other hand, it is easily seen that

$$u \in \mathbb{R}^{\bar{n}} \Rightarrow \|u\|_2^2 \leq \max \left[ \frac{\bar{n}}{s^2}, 1 \right] \|u\|_{s,1}^2,$$

and since

$$\|\text{Col}_j[I_{\bar{n}} - D]\|_{s,1} \leq \|\text{Col}_j[I_n - H^TA]\|_{s,1} \leq \kappa_s^+(A) < 1/2,$$

we get $\|\text{Col}_j[I_{\bar{n}} - D]\|_2^2 \leq \max \left[ \frac{\bar{n}}{s^2}, 1 \right] \cdot \frac{1}{4}$, whence

$$\sum_{i,j=1}^{\bar{n}} [I_{\bar{n}} - D]_{ij}^2 \leq \bar{n} \max \left[ \frac{\bar{n}}{s^2}, 1 \right] \cdot \frac{1}{4} = \max \left[ \frac{4m^2}{s^2}, 2m \right] \cdot \frac{1}{4}$$

Thus,

$$m \leq \max \left[ \frac{m^2}{s^2}, \frac{m}{2} \right] \Rightarrow s \leq \sqrt{m}.$$
“True” upper bounds on $s$-goodness

It is known that $m \times n$ matrices from typical random ensembles, e.g., Gaussian (i.i.d. entries $\sim \mathcal{N}(0, 1/m)$) or Rademacher (i.i.d. entries taking values $\pm 1/\sqrt{m}$ with probabilities $1/2$) with probability approaching 1 as $m, n$ grow are $s$-good with $s$ as large as $O(1)m/\log(2n/m)$, which is by far better than the maximal level of goodness $O(\sqrt{m})$ which can be certified by our verifiable sufficient conditions.

Specifically, let us say that an $m \times n$ matrix $A$ possesses Restricted Isometry Property with parameters $\delta, k$ ($A$ is RIP($\delta, k$) for short), if

$$(1 - \delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta)\|x\|_2^2$$

for all $k$-sparse vectors $x$.

It is known that

A. A random Gaussian/Rademacher $m \times n$ matrix is, with probability approaching 1 as $m, n$ grow, RIP(0.1, $k$) with $k$ as large as $O(m/\ln(2n/m))$;

B. Whenever $A$ is RIP($\delta, 2s$) with $\delta < 1/3$, $A$ is $s$-good.
B. Whenever $A$ is RIP($\delta, 2s$) with $\delta < 1/3$, $A$ is $s$-good.

Verification of B: Let $A$ be RIP($\delta, 2s$), $\delta < 1/3$, and let $x \in \mathbb{R}^n$. Let $x^1$ be obtained from $x$ by zeroing out all but the $s$ largest in magnitude entries, $x^2$ be obtained in the same fashion from $x - x^1$, $x^3$ obtained in the same fashion from $x - x^1 - x^2$, etc. In other words, if $i_1, i_2, \ldots, i_n$ is the reordering of indexes such that $|x_{i_1}| \geq |x_{i_2}| \geq |x_{i_3}| \geq \ldots$ and $I_p = \{i_{(p-1)s+1}, \ldots, i_{ps}\}$, $1 \leq p \leq d = \lfloor n/s \rfloor$, then $x^p = x_{I_p}$.

We have $\|x^{p+1}\|_\infty \leq \|x^p\|_1/s$, $\|x^{p+1}\|_1 \leq \|x^p\|_1 \Rightarrow \|x^{p+1}\|_2 \leq \sqrt{\|x^{p+1}\|_\infty \|x^{p+1}\|_1} \leq \sqrt{s^{-1/2}} \|x^p\|_1$. We further have

$$\|Ax^i\|_2 \|Ax\|_2 \geq \|Ax^1\|^T[Ax] = \sum_{p=1}^d \|Ax^1\|^T[Axp] \geq \|Ax^1\|_2^2 - \sum_{p=2}^d \|Ax^1\|^T[Axp] \quad (*)$$

**Lemma:** If $A$ is RIP($\delta, 2s$) and $u, v$ are $s$-sparse with non-intersecting supports, then $\|u^T A^T Av\| \leq \delta \|u\|_2 \|v\|_2$.

Indeed, Lemma states that if $Q$ is symmetric matrix such that $(1 - \delta)y^T y \leq y^T Q y \leq (1 + \delta)y^T y$ for all $y$, then $\|u^T Q v\| \leq \delta \|u\|_2 \|v\|_2$ whenever $u^T v = 0$. This is evident, since from the premise it follows that the eigenvalues of $Q$ are in-between $1 - \delta$ and $1 + \delta$, whence the spectral norm of $Q - I$ is $\leq \delta$, whence for $u, v$ in question $\|u^T Q v\| = |u^T v + u^T (Q - I) v| = |u^T (Q - I) v| \leq \delta \|u\|_2 \|v\|_2$.

Applying Lemma, $(*)$ leads to

$$\|Ax^1\|_2 \|Ax\|_2 \geq \|Ax^1\|^2 - \delta \sum_{p=2}^d \|x^1\|_2 \|x^p\|_2 \geq \|Ax^1\|^2 - \delta s^{-1/2} \|x^1\|_2 \sum_{p=1}^{d-1} \|x^p\|_1$$

$$\Rightarrow \|Ax^1\|_2 \leq \|Ax\|_2 + \delta s^{-1/2} \|x^1\|_2 \|Ax^1\|_2 \Rightarrow \|x\|_2 \leq \frac{1}{\sqrt{1 - \delta}} \|Ax\|_2 + s^{-1/2} \frac{\delta}{1 - \delta} \|x\|_1$$

whence

$$\|x\|_{s,1} \leq s^{1/2} \|x\|_2 \leq \frac{s^{1/2}}{\sqrt{1 - \delta}} \|Ax\|_2 + \frac{\delta}{1 - \delta} \|x\|_1 \Rightarrow \kappa_s(A) \leq \frac{\delta}{1 - \delta} < 1/2, \quad \beta_s(A, \|\cdot\|_2) \leq \frac{s^{1/2}}{\sqrt{1 - \delta}}.$$
\[ \|x^1\|_2 \leq \frac{1}{\sqrt{1 - \delta}} \|Ax\|_2 + s^{-1/2} \frac{\delta}{1 - \delta} \|x\|_1 \]

Observing that \( \|x^1\|_{\infty} \leq \|x^1\|_2 \), we derive from (!) that

\[ \|x\|_{1,1} \leq \frac{1}{\sqrt{1 - \delta}} \|Ax\|_2 + \frac{s^{-1/2}\delta}{1 - \delta} \|x\|_1, \]

meaning that whenever \( A \) satisfies RIP(\( \delta, k \)) with \( \delta < 1/3 \), we have \( \kappa_1^+(A) \leq \frac{s^{-1/2}\delta}{1 - \delta} \), and the corresponding certificate \( H \) of \( s \)-goodness can be chosen to have \( \|\text{Col}_j(H)\|_2 \leq \frac{1}{\sqrt{1 - \delta}} \), \( 1 \leq j \leq n \).

**Bottom line:** Our verifiable sufficient condition for \( s \)-goodness, even in its simplest form, allows to certify at least the square root of the goodness level as guaranteed by (heavily computationally intractable) RIP. On the other hand, whenever \( n \geq 2m \), our condition for \( s \)-goodness fails to certify goodness level better than \( \sqrt{m} \).
Numerical illustration:
Efficiently Computable Lower and Upper bounds on $s_*(A) = \max \{ s : A \text{ is } s\text{-good} \}$

<table>
<thead>
<tr>
<th>$m \times 256$ random submatrix of $256 \times 256$ matrix</th>
<th>$m$</th>
<th>LB I</th>
<th>LB II</th>
<th>UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fourier matrix</td>
<td>128</td>
<td>3</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>178</td>
<td>3</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>242</td>
<td>5</td>
<td>11</td>
<td>26</td>
</tr>
<tr>
<td>Hadamard matrix</td>
<td>128</td>
<td>2</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>178</td>
<td>4</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>242</td>
<td>12</td>
<td>26</td>
<td>31</td>
</tr>
<tr>
<td>Rademacher matrix</td>
<td>128</td>
<td>1</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>178</td>
<td>2</td>
<td>8</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>242</td>
<td>2</td>
<td>23</td>
<td>47</td>
</tr>
<tr>
<td>Gaussian matrix</td>
<td>128</td>
<td>1</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>178</td>
<td>2</td>
<td>8</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>242</td>
<td>2</td>
<td>23</td>
<td>47</td>
</tr>
</tbody>
</table>

- LB I: Lower Bound on $s_*(A)$ based on Mutual Incoherence
- LB II: Lower Bound on $s_*(A)$ based on $\kappa^+_s(A)$
- UB: Upper Bound on $S_*(A)$

- $\kappa^+_s$-based goodness bounds significantly outperform bounds based on mutual incoherence
- Computation has its price: for random matrices, there is a significant gap between upper and lower goodness bounds

1.38
### Efficiently computable goodness bounds

<table>
<thead>
<tr>
<th>$m$</th>
<th>LB I</th>
<th>LB II</th>
<th>UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>102</td>
<td>2</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>204</td>
<td>2</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>307</td>
<td>2</td>
<td>6</td>
<td>30</td>
</tr>
<tr>
<td>409</td>
<td>3</td>
<td>7</td>
<td>44</td>
</tr>
<tr>
<td>512</td>
<td>3</td>
<td>10</td>
<td>61</td>
</tr>
<tr>
<td>614</td>
<td>3</td>
<td>12</td>
<td>78</td>
</tr>
<tr>
<td>716</td>
<td>3</td>
<td>15</td>
<td>105</td>
</tr>
<tr>
<td>819</td>
<td>4</td>
<td>21</td>
<td>135</td>
</tr>
<tr>
<td>921</td>
<td>4</td>
<td>32</td>
<td>161</td>
</tr>
<tr>
<td>960</td>
<td>0</td>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>

- Matrices with “personal story” seem to have smaller and easier to estimate goodness than random matrices of the same sizes.
Note: At least in the case of random matrices $A$, there exists a significant gap between $s$-goodness (the ability of $\ell_1$ recovery to recover well all $s$-sparse signals in the noiseless case) and “near $s$-goodness” – the ability of $\ell_1$ recovery to reproduce well with high reliability random $s$-sparse signals in the noiseless case.

For a randomly selected $256 \times 512$ submatrix $A$ of the $512 \times 512$ Hadamard matrix,
— lower bound on $s$-goodness, as given by the condition $\kappa_s^+(A) < 0.5$, is $s = 8$
— upper bound on $s$-goodness is $s = 15$. Here is a badly recovered in the noiseless case $16$-sparse signal:

True $16$-sparse signal (magenta) and its recovery (blue)

However, in a series of 100 experiments with noiseless $\ell_1$ recovery of randomly generated $81$-sparse signals, not a single erroneous recovery was observed!
Consider time-varying discrete time linear dynamical system

\[
\begin{align*}
  x_0 &= z \
  x_{t+1} &= A_t x_t + B_t u_t + R_t d_t \
  y_t &= C_t x_t + D_t d_t
\end{align*}
\]

(initial state)

state equations
- \( x_t \): state
- \( u_t \): control
- \( d_t \): external disturbance

(observed output)

“closed” by affine output-based control law

\[
  u_t = g_t + \sum_{\tau=0}^{t} G_t^\tau y_\tau.
\]

Given finite time horizon \( 0 \leq t \leq N \), we want to specify a control law (*) which ensures that the state-control trajectory \( w = [x_0; ...; x_{N+1}; u_0; ...; u_N] \) satisfies given design specifications

\[
  A w \leq b \iff a_i^T w \leq b_i, \quad 1 \leq i \leq I
\]

robustly w.r.t. the “perturbation” \( \zeta = [z; d_0; ...; d_N] \) running through a given set \( \mathcal{Z} \).

Good news: by linearity of the system and the control law, the trajectory is affine in \( \zeta \):

\[
  w = w^0 + W \zeta
\]

The Analysis problem: check whether a given control law (*) robustly meets the design specifications reduces to verifying whether a system of affine constraints on \( \zeta \) is satisfied by all \( \zeta \in \mathcal{Z} \). This is easy, provided \( \mathcal{Z} \) is “tractable.”
System:
\[
\begin{align*}
x_0 &= z \\
x_{t+1} &= A_t x_t + B_t u_t + R_t d_t \\
y_t &= C_t x_t + D_t d_t
\end{align*}
\]
[initial state]
\[
\begin{align*}
\text{state equations} \\
&\begin{aligned}
&\bullet x_t: \text{state} \\
&\bullet u_t: \text{control} \\
&\bullet d_t: \text{external disturbance}
\end{aligned}
\end{align*}
\]
[observed output]

Controller:
\[u_t = g_t + \sum_{\tau=0}^{t} G_t^\tau y_\tau.\] \hfill (*)

Trajectory: \[w = [x_0; \ldots; x_{N+1}; u_0; \ldots; u_N] \]

Design specifications:
\[A w \leq b \iff a_i^T w \leq b_i, \ 1 \leq i \leq I \]
\hfill (!)

\[\star\] From now on, assume that \(\mathcal{Z}\) is given by polyhedral representation:
\[
\mathcal{Z} = \{\zeta : \exists v : P\zeta + Qv \leq r\}
\]

Then to check whether (*) ensures (!) for all \(\zeta \in \mathcal{Z}\) is the same as to check whether
\[
\max_{\zeta, v} \{a_i^T [w^0 + W \zeta] : P\zeta + Qv \leq r\} \leq b_i, \ 1 \leq i \leq I.
\]

\[\Rightarrow \text{Verification requires solving I LO programs.}\]
\[
\begin{align*}
x_0 &= z \\
x_{t+1} &= A_t x_t + B_t u_t + R_t d_t \\
y_t &= C_t x_t + D_t d_t \\
u_t &= g_t + \sum_{\tau=0}^{t} G^\tau_t y_{\tau} \\
\end{align*}
\]

(S)

**Bad news:** the trajectory is highly nonlinear in the parameters \(\gamma = \{g_t, G^\tau_t\}\) of the control law (\(*\))

\(\Rightarrow\) **The Synthesis problem:** find control law (\(*\)), if it exists, which robustly meets the design specifications seems to be intractable.

**Remedy:** pass to affine purified-output-based control laws.

\(\blacklozenge\) Consider, along with system (\(S\)) “closed” by some control law, its *model*

\[
\begin{align*}
\hat{x}_0 &= 0 \\
\hat{x}_{t+1} &= A_t \hat{x}_t + B_t u_t \\
\hat{y}_t &= C_t \hat{x}_t \\
\end{align*}
\]

(M)

which we “feed” by the same controls \(u_t\) as (\(S\)). We can run the model in an on-line fashion, and thus at time \(t\), before the decision on \(u_t\) should be made, we have at our disposal *purified output* \(v_t = y_t - \hat{y}_t\)

**Observation:** purified outputs are known in advance affine functions of \(\zeta\) completely independent on the control law in use

Indeed, setting \(\Delta_t = x_t - \hat{x}_t\), we clearly have

\[
v_t = C_t \Delta_t + D_t d_t, \quad \Delta_0 = z, \quad \Delta_{t+1} = A_t \Delta_t + R_t d_t.
\]
\[
\begin{align*}
\text{System:} & & \text{Model:} \\
\quad x_0 &= z & \quad \hat{x}_0 &= 0 \\
\quad x_{t+1} &= A_t x_t + B_t u_t + R_t d_t & \quad \hat{x}_{t+1} &= A_t \hat{x}_t + B_t u_t \quad (S) \\
\quad y_t &= C_t x_t + D_t d_t & \quad \hat{y}_t &= C_t \hat{x}_t \quad (M)
\end{align*}
\]

**Purified outputs:** \( v_t = y_t - \hat{y}_t \)

\[
\begin{align*}
u_t = \begin{cases} 
g_t + \sum_{\tau=0}^{t} G_{t \tau} y_{\tau} & \text{[output-based affine law]} \\
h_t + \sum_{\tau=0}^{t} H_{t \tau} v_{\tau} & \text{[purified-output-based affine law]} \end{cases} \quad (\ast)
\end{align*}
\]

**Facts:**

\(\heartsuit\) **Affine purified-output-based and output-based controls laws are equivalent:** every mapping \(\zeta \to w\) which can be obtained when “closing” \((S)\) by a law \((\ast)\), can be obtained by closing \((S)\) by a law \((+)\), and vice versa.

\(\heartsuit\) **When \((S)\) is closed by a purified-output-based affine control law \((+)\), the trajectory \(w = W[\zeta, \eta]\) becomes bi-affine in \(\zeta\) and in the parameters \(\eta = \{h_t, H_t^\tau\}\) of the control law:**

\[
w = w^0[\eta] + W[\eta] \zeta \quad \text{with} \quad w^0[\eta],\ W[\eta] \text{ affine in } \eta.
\]
The state-control trajectory of system “closed” with purified-output-based control law with parameters $\eta$:

$$w = w^0[\eta] + W[\eta] \zeta$$

with known affine $w^0[\cdot], W[\cdot]$

What we want:

$$Aw \leq b \forall \zeta : \exists v : P\zeta + Qw \leq r$$

**Facts (continued):**

**Sticking to purified-output-based control laws, the Synthesis problem**

Given design specifications $a_i^T w \leq b_i$, $1 \leq i \leq I$, on the state-control trajectory, find a control law, if one exists, which meets these specifications robustly w.r.t. $\zeta = [z; d_0; ...; d_N] \in \mathbb{Z}$

becomes an infinite system of linear constraints on $\eta$:

$$a_i^T [w^0[\eta] + W[\eta] \zeta] \leq b_i \forall \zeta \in \mathbb{Z}, 1 \leq i \leq I.$$  

which is fact is equivalent to an explicit finite “moderate size” system of linear constraints on $\zeta$ and additional variables.
**Question:** What the infinite system of linear constraints on \( \eta \):

\[
\forall (\zeta : \exists v : P\zeta + Qv \leq r) : a^T_i \left[ w^0[\eta] + W[\eta] \zeta \right] \leq b_i, \ i \leq I
\]

"wants" from \( \eta \)?

**Answer:** It wants the optimal values in \( I \) feasible parametric LP's:

\[
\text{Opt}_{i}[\eta] = \max_{\zeta, v} \left\{ a^T_i W[\eta] \zeta : P\zeta + Qv \leq r \right\}
= \min_{y^i} \left\{ r^T y^i : [y^i]^T P + a^T_i W[\eta] = 0, Q^T y^i = 0, y^i \geq 0 \right\} \quad [\text{LP duality}]
\]

to satisfy the constraints \( a^T_i w^0[\eta] + \text{Opt}_i[\eta] \leq b_i, \ i \leq I, \Rightarrow \) the set of desirable \( \eta \) admits polyhedral representation

\[
\left\{ \eta : \exists y^1, \ldots, y^I : \begin{align*}
[y^i]^T P + a^T_i W[\eta] & = 0, Q^T y^i = 0, y^i \geq 0 \\
\left[ a^T_i w^0[\eta] + r^T y^i \right] & \leq b_i
\end{align*} \right\}
\]

**(S)**

**Bottom line:** A purified-output-based affine control law with parameters \( \eta \) meets the design specifications \( a^T_i w \leq b_i, \ 1 \leq i \leq I, \Rightarrow \) robustly in \( \zeta \in \mathbb{Z} \) iff \( \eta \) can be extended by properly chosen \( y^i, \ i \leq I, \) to a feasible solution of \((S)\).
How it Works: Controlling 3-Level Serial Inventory

- Level 1 supplies external demand
- Level 2 supplies Level 1
- Level 3 supplies Level 2 and is supplied from Factory
- There is 2-period delay in executing replenishment orders

The Inventory can be modeled as the 9-state LDS

\[
\begin{align*}
    x_1(t+1) &= x_1(t) + x_{1,1}(t) - d_t \\
    x_{1,1}(t+1) &= x_{1,2}(t) \\
    x_{1,2}(t+1) &= x_{2}(t) + x_{2,1}(t) - u_1(t) \\
    x_2(t+1) &= x_2(t) + x_{3,1}(t) - u_2(t) \\
    x_{2,1}(t+1) &= x_{2,2}(t) \\
    x_{2,2}(t+1) &= x_{3}(t) + x_{3,1}(t) - u_3(t) \\
    x_{3}(t+1) &= x_{3}(t) + x_{3,1}(t) - u_3(t) \\
    y(t) &= x(t)
\end{align*}
\]

- \( x_1(\cdot), x_2(\cdot), x_3(\cdot) \) — inventory levels
- \( u_1(\cdot), u_2(\cdot), u_3(\cdot) \) — replenishment orders
- \( d_t \) — demands

1.47
It is well known that serial inventories with delays (and supply chains in general) suffer from bullwhip effect: variations of states (e.g., inventory levels) are severely amplified when moving upward from external demand to production units along the supply chain. This phenomenon badly affects the production.

• This is what happens with “naive” affine controller:

Note: variations of the demand in the range \([-1,1]\) result in huge (hundreds!) oscillations in the level #3 and in the replenishment orders.
To reduce the bullwhip effect, we can look for the best — with the largest decay rate as certified by Lyapunov Stability Certificate, whatever it means — linear feedback control law

\[ u(t) = Ky(t) \ [= Kx(t)]. \]

With this control, the picture looks much better:

---

**Good linear feedback**

Top: time-dependent demand \( d_t \in [-1, 1] \)
Middle: replenishment orders \( u_1(t), u_2(t), u_3(t) \in [-15, 5] \)
Bottom: inventory levels \( x_1(t), x_2(t), x_3(t) \in [-5, 15] \)

**But:** At the very beginning, we still have unpleasant jumps in the inventory levels and replenishment orders.
To improve the behaviour of the process in the beginning, we can use purified-output-based affine control aimed at minimizing the initial jumps and eventually switching to the above feedback control. This is what we get:

**Combined p.o.b./feedback control**
Top: time-dependent demand \( d_t \in [-1, 1] \)
Middle: replenishment orders \( u_1(t), u_2(t), u_3(t) \)
Bottom: inventory levels \( x_1(t), x_2(t), x_3(t) \)
This is what we gain in the beginning, while loosing nothing in the long run:

**Pure feedback control (left)**

**vs.**

**combined p.o.b/feedback control (right)**

Top: time-dependent demand varying in $[-1, 1]$
Middle: replenishment orders $u_1(t), u_2(t), u_3(t)$
Bottom: inventory levels $x_1(t), x_2(t), x_3(t)$
From Linear to Conic Programming

When passing from a generic LP problem
\[
\min_x \{ c^T x : Ax - b \geq 0 \} \quad [A : m \times n] \tag{LP}
\]
to nonlinear extensions, some components of the problem become nonlinear. The traditional way is to allow nonlinearity of the objective and the constraints:

\[
c^T x \mapsto c(x); \quad a_i^T x - b_i \mapsto a_i(x)
\]

and to preserve the “coordinate-wise” interpretation of the vector inequality \( A(x) \geq 0 \):

\[
A(x) \equiv \begin{bmatrix}
a_1(x) \\
\vdots \\
a_m(x)
\end{bmatrix} \geq 0 \iff a_i(x) \geq 0, \quad i = 1, \ldots, m.
\]

An alternative is to preserve the linearity of the objective and the constraint functions and to modify the interpretation of the vector inequality \( \geq \). In Convex Programming, both approaches are equivalent.

The second option turns out to be more preferable, since it “reveals the structure” of a convex program: an extremely wide variety of convex programs can be captured by vector inequalities of just 3 “standard” and well understood types.
**Example:** The problem with nonlinear objective and constraints

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>minimize</strong></td>
<td>( \sum_{\ell=1}^{n} x_{\ell}^2 )</td>
</tr>
<tr>
<td><strong>(a)</strong></td>
<td>( x \geq 0; )</td>
</tr>
<tr>
<td><strong>(b)</strong></td>
<td>( a_\ell^T x \leq b_\ell, \ \ell = 1, \ldots, n; )</td>
</tr>
<tr>
<td><strong>(c)</strong></td>
<td>( |Px - p|_2 \leq c^T x + d; )</td>
</tr>
<tr>
<td><strong>(d)</strong></td>
<td>( x_{\ell}^{i+1} \leq e_\ell^T x + f_\ell, \ \ell = 1, \ldots, n; )</td>
</tr>
<tr>
<td><strong>(e)</strong></td>
<td>( x_{\ell}^{i+3} x_{\ell+1}^{i+3} \geq g_\ell^T x + h_\ell, \ \ell = 1, \ldots, n - 1; )</td>
</tr>
<tr>
<td><strong>(f)</strong></td>
<td>Det ( \begin{bmatrix} x_1 &amp; x_2 &amp; x_3 &amp; \cdots &amp; x_n \ x_2 &amp; x_1 &amp; x_3 &amp; \cdots &amp; x_{n-1} \ x_3 &amp; x_2 &amp; x_1 &amp; \cdots &amp; x_{n-2} \ \vdots &amp; \vdots &amp; \vdots &amp; \ddots &amp; \vdots \ x_n &amp; x_{n-1} &amp; x_{n-2} &amp; \cdots &amp; x_1 \end{bmatrix} \geq 1; )</td>
</tr>
<tr>
<td><strong>(g)</strong></td>
<td>( 1 \leq \sum_{\ell=1}^{n} x_{\ell} \cos(\ell \omega) \leq 1 + \sin^2(5\omega) \ \forall \omega \in [-\frac{\pi}{7}, 1.3] )</td>
</tr>
</tbody>
</table>

can be converted, in a systematic way, into an equivalent problem

\[
\min_x \{ c^T x : Ax - b \succeq 0 \},
\]

"\( \succeq \)" being one of the 3 standard vector inequalities, so that seemingly highly diverse constraints of the original problem allow for unified treatment.

1.53
A significant part of nice mathematical properties of an LP program
\[
\min_x \{ c^T x : Ax - b \geq 0 \}
\]
stems from the fact that the underlying coordinate-wise vector inequality
\[
a \geq b \iff a_i \geq b_i, \ i = 1, \ldots, m \quad [a, b \in \mathbb{R}^m]
\]
satisfies a number of quite general axioms, namely:

**I.** It defines a *partial ordering* of \( \mathbb{R}^m \), i.e., is

I.a) *reflexive:* \( a \geq a \) for all \( a \in \mathbb{R}^m \)

I.b) *anti-symmetric:* if \( a \geq b \) and \( b \geq a \), then \( a = b \)

I.c) *transitive:* if \( a \geq b \) and \( b \geq c \), then \( a \geq c \)

**II.** It is *compatible with linear structure of* \( \mathbb{R}^m \), i.e., is

II.a) *additive:* if \( a \geq b \) and \( c \geq d \), then \( a + c \geq b + d \)

II.b) *homogeneous:* if \( a \geq b \) and \( \lambda \) is nonnegative real, then \( \lambda a \geq \lambda b \).
“Good” vector inequalities

- A vector inequality $\succeq$ on $\mathbb{R}^m$ is a binary relation – a set of ordered pairs $(a, b)$ with $a, b \in \mathbb{R}^m$. The fact that a pair $(a, b)$ belongs to this set is written down as $a \succeq b$ (“$a$ $\succeq$-dominates $b$”).
- Let us call a vector inequality $\succeq$ good, if it satisfies the outlined axioms, namely, is reflexive, antisymmetric, transitive, additive, and homogeneous.

**Observation:** A good vector inequality $\succeq$ on $\mathbb{R}^m$ is uniquely defined by the set

$$K = \{a \in \mathbb{R}^m : a \succeq 0\}$$

of all $\succeq 0$-nonnegative vectors, specifically,

$$a \succeq b \iff a - b \succeq 0 \iff a - b \in K$$

A set $K \subset \mathbb{R}^m$ specifies, in the above fashion, a good vector inequality iff $K$ is a pointed convex cone, that is,

- is nonempty,
- is conic: $a \in K$, $\lambda \geq 0 \Rightarrow \lambda a \in K$
- is convex,
- is pointed: $a \in K$ and $-a \in K$ iff $a = 0$,

or, equivalently, is a nonempty subset of $\mathbb{R}^m$ closed w.r.t. taking conic combinations (linear combinations with nonnegative coefficients) of its elements and not containing lines passing through the origin.
Example: The entrywise vector inequality $\geq$ stems from the nonnegative orthant $\mathbb{R}^+_m$:

$$a \geq b \Leftrightarrow a - b \geq 0 \Leftrightarrow a - b \in \mathbb{R}^+_m = \{x \in \mathbb{R}^m : x_i \geq 0, 1 \leq i \leq m\}.$$ 

The nonnegative orthant $\mathbb{R}^+_m$, along with being convex cone, possesses two additional properties:

- is closed, and
- possesses nonempty interior.

The first of this properties allows to pass to termwise limits in $\geq$ inequalities:

$$a_i \geq b_i \& a = \lim_{i} a_i \& b = \lim_{i} b_i \Rightarrow a \geq b.$$

The second property allows to define strict version $>$ of $\geq$:

$$a > b \Leftrightarrow a - b \in \text{int} \mathbb{R}^+_m \[= \{x \in \mathbb{R}^m : x_i > 0, i \leq m\}\]$$

which is stable w.r.t. small enough perturbations of $a, b$.

It makes sense to incorporate these useful properties into the definition of a "good" vector inequality.
**Bottom line:** From now on, a good vector inequality on $\mathbb{R}^m$ is the relation $\geq_K$ specified by a regular cone (closed convex pointed cone with a nonempty interior) $K \subset \mathbb{R}^m$ according to

$$a \geq_K b \iff a - b \geq_K 0 \iff a - b \in K.$$  

Along with $\geq_K$, the cone $K$ specifies the strict inequality $>_K$:

$$a >_K b \iff a - b >_K 0 \iff a - b \in \text{int} K.$$  

**Note:** Arithmetics and elementary topology of good vector inequalities $\geq_K$, $>_K$ is exactly the same as for entrywise vector inequality $\geq$ (and the scalar $\geq$), e.g.

- sum of two valid nonstrict/strict $K$-inequalities is a valid nonstrict $K$-inequality, and is strict, if at least one of the two inequalities we are summing up is strict;
- multiplying both sides of a valid nonstrict/strict $K$-inequality by a nonnegative real, we get valid nonstrict $K$-inequality which is strict, provided that the real is positive and the original inequality was strict;
- small enough perturbations in both sides of a valid strict $K$-inequality preserve inequality's validity;
- if left- and right hand sides in a sequence of valid $K$-inequalities have limits, these limits are linked by valid nonstrict $K$-inequality.
Facts:
A. The entrywise vector inequality
\[ a \geq b \iff a_i \geq b_i, \ i = 1, \ldots, m \]
is neither the only possible, nor the only interesting good vector inequality on \( \mathbb{R}^m \).

B. A good vector inequality \( \geq_K \) gives rise to generic conic program
\[
\min_x \{ c^T x : Ax - b \geq_K 0 \},
\]
and these programs inherit significant part of nice theoretical properties of LP's.

At the same time, "playing with \( K \)" – working with regular cones different from non-negative orthants – extends dramatically the scope of convex optimization problems we can handle. Moreover, for all practical purposes just three "magic" families of regular cones cover the entire Convex Programming.
Direct products of nonnegative rays — nonnegative orthants — give rise to the entrywise vector inequalities and thus — to generic Linear Programming problem

\[
\min_{x \in \mathbb{R}^n} \left\{ c^T x : Ax - b \geq 0 \right\} \quad \quad [A \in \mathbb{R}^{m \times n}]
\]
$m$-dimensional Lorentz cone (a.k.a. Second Order, or Ice-Cream, cone) is defined as

\[ L^m = \left\{ x = [x_1; \ldots; x_m] \in \mathbb{R}^m : x_m \geq \sqrt{\sum_{i=1}^{m-1} x_i^2} \right\} \]
Direct products of Lorentz cones give rise to Conic Quadratic (a.k.a. Second Order Conic) programs. A generic Conic Quadratic problem is of the form

\[
\min_{x} \{c^T x : \|D_i x + d_i\|_2 \leq e_i^T x + f_i, 1 \leq i \leq m\}
\]

\[
\min_{x} \left\{ c^T x : Ax - b \equiv \begin{bmatrix}
D_1 x + d_1 \\
\vdots \\
D_m x + d_m \\
\end{bmatrix} \begin{bmatrix}
e_1^T x + f_1 \\
\vdots \\
e_m^T x + f_m \\
\end{bmatrix} \geq K 0 \right\},
\]

\[K = L^{m_1} \times \ldots \times L^{m_k}\]

is a direct product of Lorentz cones.
Magic families of cones, III
Directs products of semidefinite cones

♣ **Semidefinite cone** $S^m_+$ lives in the space $S^m$ of real symmetric $m \times m$ matrices and is comprised of all $m \times m$ symmetric matrices $A$ which are *positive semidefinite*, that is, produce everywhere nonnegative quadratic forms $x^T A x$ or, equivalently, have nonnegative eigenvalues.

3 random 3D cross-sections of $S^3_+$
Direct products of semidefinite cones give rise to semidefinite programs

\[
\min_x \left\{ \begin{array}{l}
    c^T x : A_i(x) := \sum_j x_j A_{ij} - B_i \succeq 0, \ i \leq I \\
\end{array} \right. 
\]

where \( A_{ij}, B_i \) are symmetric matrices of size \( m_i \), and \( P \succeq Q (\equiv Q \preceq P) \) means that \( P, Q \) are symmetric matrices of the same size such that \( P - Q \) is positive semidefinite.

**Note:** Semidefinite program is the program of minimizing a linear objective under the bunch of LMI (Linear Matrix Inequality) constraints stating each that a variable symmetric matrix with entries affine in the decision vector \( x \) should be positive semidefinite.

**Note:** We can always write down a semidefinite program as a program with single LMI constraint:

\[
\min_x \{ c^T x : A_i(x) \succeq 0, \ i \leq m \} \Leftrightarrow \min_x \{ c^T x : A(x) := \text{Diag}\{A_1(x), \ldots, A_m(x)\} \succeq 0 \}. 
\]
Conic Duality

- Let us look at the origin of the problem dual to an LP program

\[
\min_x \{ c^T x : Ax - b \geq 0 \}. \tag{LPr}
\]

Observing that any nonnegative “weight vector” \( y \in \mathbb{R}_+^m \) is “admissible” for the constraint-wise vector inequality on \( \mathbb{R}^m \):

\[
\forall a, b, y \in \mathbb{R}^m : a \geq b \land y \geq 0 \Rightarrow y^T a \geq y^T b
\]

we conclude that all scalar linear inequalities of the type

\[
[ A^T y]^T x \geq b^T y \quad \text{with } y \geq 0
\]

with variables \( x \) are consequences of the constraints of (LPr). Thus, (*): If \( y \geq 0 \) is such that \( A^T y = c \), then \( b^T y \) is a lower bound on the optimal value in (LPr).

- The LP dual to (LPr) is exactly the problem

\[
\min_y \{ b^T y : A^T y = c, y \geq 0 \} \tag{LDI}
\]

of finding the best – the largest – lower bound on the optimal value of (LPr) among those given by (*).
Conic Duality, same as the LP one, is inspired by the desire to bound from below the optimal value in a conic program

$$\min_{x} \{ c^T x : Ax - b \geq_K 0 \}$$  \hspace{1cm} (CP)

and follows the just outlined scheme based on “conversion” of vector inequalities into the scalar ones:

$$a \geq_K b \Rightarrow y^T a \geq y^T b,$$

\hspace{1cm} (*)

**Crucial question** is:

*What are the "aggregation weights" $y$ which make (*) valid?*

**Answer:** A necessary and sufficient condition for the implication (*) to be true is

$$y \in K_* := \{ y : y^T x \geq 0 \forall x \in K \}$$

**Note:** $K_*$ is called the cone dual to $K$. Whenever $K$ is a regular cone, so is $K_*$, and

$$K = (K_*)_*.$$
We are ready to build the dual of a conic program. It is convenient to start with the primal problem in the form

$$\text{Opt}(P) = \min_x \{c^T x : Ax - b \in K, Rx = r\} \quad (P)$$

To build the dual, we equip the constraints of $(P)$ with *Lagrange multipliers* $y \in K^*, s \in \mathbb{R}^{\dim r}$

**Note:** the "aggregated constraint"

$$[A^T y]^T x + [R^T s]^T x \geq b^T y + r^T s,$$

by its origin is a consequence of the constraints of $(P)$. Consequently, *Whenever $A^T y + R^T s = c$, the quantity $b^T y + r^T s$ is a lower bound on Opt$(P)$. The problem*

$$\max_{y,s} \{b^T y + r^T s : y \in K^*, A^T y + R^T s = c\}$$

*dual to $(P)$ is to find the best – the largest – bound of this type.*
"In real life" a conic problem arises as

\[ \text{Opt}(P) = \min_x \{ c^T x : A_i x - b_i \in K^i, i \leq m, Rx = r \} \]  \hspace{1cm} (P)

that is, the associated regular cone is the direct product \( K = K^1 \times \ldots \times K^m \). We clearly have

\[ K_* = K^1_* \times \ldots \times K^m_* , \]

implying that the recipe for building the dual problem is as follows:

- we equip conic constraints \( A_i x - b_i \in K^i \) with Lagrange multipliers \( y^i \in K^i_* \), and the linear equality constraints – with Lagrange multiplier \( s \in \mathbb{R}^{\dim r} \)
- we induce from the constraints of \( (P) \) that \( y^T_i [A_i x - b_i] \geq 0 \) and \( s^T [Rx - r] \geq 0 \), so that the aggregated constraint

\[ \left[ \sum_i A^T_i y^i + R^T s \right]^T x \geq \sum_i b^T_i y^i + r^T s \]

is the consequence of the constraints of \( (P) \). In particular, whenever \( y^i \in K^i_* \) and \( s \) satisfy \( \sum_i A^T_i y^i + R^T s = c \), the quantity \( \sum_i b^T_i y^i + r^T s \) is a lower bound on \( \text{Opt}(P) \). The dual problem

\[ \text{Opt}(D) = \max_{y^i,s} \left\{ \sum_i b^T_i y^i + r^T s : y^i \in K^i_*, i \leq m, \sum_i A^T_i y^i + R^T s = c \right\} \]

is to find the best – the largest – of these lower bounds on \( \text{Opt}(P) \).

**Note:** The dual problem is conic along with the primal problem.

**Note:** The magic cones are self-dual, so that in this case \( (D) \) involves the same cones as \( (P) \).
\[
\text{Opt}(P) = \min_x \{ c^T x : Ax - b \in K, Rx = r \} \quad (P)
\]
\[
\text{Opt}(D) = \max_{y,s} \{ b^T y + r^T s : y \in K_*, A^T y + R^T s = c \} \quad (D)
\]

♠ The origin of the dual problem yields the

**Weak Duality Theorem:** \( \text{Opt}(P) \geq \text{Opt}(D) \).

**Equivalently:** The value of the primal objective \( c^T x \) at every primal feasible solution (one feasible for \((P)\)) is \( \geq \) the value of the dual objective \( b^T y + r^T s \) at every dual feasible solution \([y; s]\) (one feasible for \((D)\)).

**Equivalently:** The duality gap

\[
\text{DualityGap}(x; y, s) = c^T x - [b^T y + r^T s]
\]

evaluated at a primal-dual feasible pair \( x, [y; s] \), always is nonnegative.
Geometry of primal-dual pair of conic problems

\[ \text{Opt}(P) = \min_x \{ c^T x : Ax - b \in K, Rx = r \} \quad (P) \]
\[ \text{Opt}(D) = \max_{y,s} \{ b^T y + r^T s : y \in K^*, A^T y + R^T s = c \} \quad (D) \]

**Assumption:** The systems of linear equality constraints in \((P)\) and \((D)\) are solvable:
\[
\exists \bar{x}, [\bar{y}, \bar{s}] : R\bar{x} = r, A^T \bar{y} + R^T \bar{s} = c.
\]

**A:** Let us pass in \((P)\) from variable \(x\) to primal slack \(\eta = Ax - b\). Whenever \(x\) satisfies \(Rx = r\), we have
\[
c^T x = [A^T \bar{y} + R^T \bar{s}]^T x = \bar{y}^T Ax + \bar{s}^T Rx = \bar{y}^T [Ax - b] + [b^T \bar{y} + r^T \bar{s}]
\]
\[\Rightarrow (P) \text{ is equivalent to the conic problem} \]
\[ \text{Opt}(P) = \min_{\eta} \{ \bar{y}^T \eta : \eta \in [L - \bar{\eta}] \cap K \}, \quad L = \{Ax : Rx = 0\}, \quad \bar{\eta} = b - A\bar{x} \]
\[ (P) \]

**Explanation:** \((P)\) wants of \(\eta := Ax - b\) (a) to belong to \(K\), and (b) to be representable as \(Ax - b\) for some \(x\) satisfying \(Rx = r\). (b) says that \(\eta\) should belong to the **primal affine plane** \(\{Ax - b : Rx = r\}\), which is the shift of the parallel linear subspace \(L = \{Ax : Rx = 0\}\) by a (whatever) vector from the primal affine plane, e.g., the vector \(-\bar{\eta} = A\bar{x} - b\).
\[ \text{Opt}(P) = \min_x \{ c^T x : Ax - b \in K, Rx = r \} \quad (P) \]
\[ \text{Opt}(D) = \max_{y,s} \{ b^T y + r^T s : y \in K_*, A^T y + R^T s = c \} \quad (D) \]

**B.** Let us pass in \((D)\) from variables \([y; s]\) to variable \(y\). Whenever \([y; s]\) satisfies \(A^T y + R^T s = c\), we have

\[ b^T y + r^T s = b^T y + \bar{x}^T R^T s = b^T y + \bar{x}^T [c - A^T y] = [b - A\bar{x}]^T y + c^T \bar{x} = \eta^T y + c^T \bar{x}, \]

\( \Rightarrow (D) \) *is equivalent to the conic problem*

\[ \text{Opt}(D) = \max_y \{ \eta^T y : y \in [\mathcal{L}^\perp + \bar{y}] \cap K_* \} \]

\[ \text{[Opt}(D) = \text{Opt}(D) - c^T \bar{x}] \quad (D) \]

Explanation: \((D)\) wants of \(y\) (a) to belong to \(K_*\), and (b) to satisfy \(A^T y = c - R^T s\) for some \(s\). (b) says that \(y\) should belong to the *dual affine plane* \(\{ y : \exists s : A^T y + R^T s = c \}\), which is the shift of the parallel linear subspace \(\widetilde{\mathcal{L}} = \{ y : \exists s : A^T y + R^T s = 0 \}\) by a (whatever) vector from the dual affine plane, e.g., the vector \(\bar{y}\). Elementary Linear Algebra says that \(\widetilde{\mathcal{L}} = \mathcal{L}^\perp\). Indeed,

\[ [\widetilde{\mathcal{L}}]^\perp = \{ z : z^T y = 0 \ \forall y : \exists z : A^T y + R^T z = 0 \} = \{ z : z^T y + 0^T z = 0 \ \text{whenever} \ A^T y + R^T z = 0 \} = \{ z : \exists x : [z^T, 0] = x^T [A^T, R^T] \} = \{ z : \exists x : Ax = z, Rx = 0 \} = \mathcal{L}. \]
Opt\((P)\) = \(\min_x \{c^T x : Ax - b \in K, R x = r\}\) \hspace{1cm} (P)
Opt\((D)\) = \(\max_{y,s} \{b^T y + r^T s : y \in K^*, A^T y + R^T s = c\}\) \hspace{1cm} (D)

\(\clubsuit\) **Bottom line:** Problems \((P), (D)\) are equivalent, respectively, to

\[\text{Opt}(P) = \min_{\eta} \{\bar{y}^T \eta : \eta \in [\mathcal{L} - \bar{\eta}] \cap K\}\]  \hspace{1cm} (P)
\[\text{Opt}(D) = \max_y \{\bar{\eta}^T y : y \in [\mathcal{L}^\perp + \bar{y}] \cap K^*\}\]  \hspace{1cm} (D)
\[\mathcal{L} = \{Ax : R x = 0\}, \ R \bar{x} = r, \ \bar{\eta} = b - A \bar{x}, \ A^T \bar{y} + R^T \bar{s} = c\]

**Note:** When \(x\) is feasible for \((P)\), and \([y;s]\) is feasible for \((D)\), the vectors \(\eta = Ax - b\), \(y\) are feasible for \((P)\), resp., \((D)\), and

\[\text{DualityGap}(x; [y,s]) = c^T x - b^T y - r^T s = [A^T y + R^T s]^T x - b^T y - r^T s = [Ax - b]^T y = \eta^T y\]

\(\Rightarrow\) Geometrically, \((P), (D)\) are as follows: "geometric data" of the problems are the pair of linear subspaces \(\mathcal{L},\ \mathcal{L}^\perp\) in the space where \(K, K^*\) live, the subspaces being orthogonal complements to each other, and pair of vectors \(\bar{\eta}, \ \bar{y}\) in this space.

- \((P)\) is equivalent to minimizing \(f(\eta) = \bar{y}^T \eta\) over the intersection of \(K\) and the primal feasible plane \(\mathcal{M}_P\) which is the shift of \(\mathcal{L}\) by \(-\bar{\eta}\)
- \((D)\) is equivalent to maximizing \(g(y) = \bar{\eta}^T y\) over the intersection of \(K^*\) and the dual feasible plane \(\mathcal{M}_D\) which is the shift of \(\mathcal{L}^\perp\) by \(\bar{y}\)
- taken together, \((P)\) and \((D)\) form the problem of minimizing the duality gap over feasible solutions to the problems, which is exactly the problem of finding pair of vectors in \(\mathcal{M}_P \cap K\) and \(\mathcal{M}_D \cap K^*\) as close to orthogonality as possible.

Pay attention to the ideal geometrical primal-dual symmetry we observe.
**Conic Duality Theorem**

♠ **Definition.** A conic problem of optimizing a linear objective under the constraints

\[ Ax - b \in K, \; Rx = r \]

is called *strictly feasible*, if there exists a feasible solution \( \bar{x} \) which *strictly* satisfies the conic constraint:

\[ \exists \bar{x} : R\bar{x} = r \; \& \; A\bar{x} - b \in \text{int} K. \]

Assuming that the conic constraint is split into "general" and "polyhedral" parts, so that the feasible set is given by

\[ Ax - b \in K, \; Px - p \geq 0, \; Rx = r \]

the problem is called *essentially strictly feasible*, if there exists a feasible solution \( \bar{x} \) which strictly satisfies the "general" conic constraint:

\[ \exists \bar{x} : R\bar{x} = r, \; P\bar{x} - p \geq 0, \; A\bar{x} - b \in \text{int} K. \]
\textbf{Note:} When the conic constraint in the primal problem allows for splitting into "general" and "polyhedral" parts:

\[
\text{Opt}(P) = \min_x \{ c^T x : Ax - b \in K, Px - p \geq 0, Rx = r \} \quad (P)
\]

then the dual problem reads

\[
\text{Opt}(D) = \max_{y,z,s} \{ b^T y + p^T z + r^T s : y \in K^*, z \geq 0, A^T y + P^T z + R^T s = c \} \quad (D)
\]

so that its conic constraint also is split into "general" and "polyhedral" parts.
Conic Duality Theorem \textit{Consider conic program along with its dual:}

\begin{align*}
\text{Opt}(P) &= \min_x \{ c^T x : Ax - b \in K, Rx = r \} \quad (P) \\
\text{Opt}(D) &= \max_{y,s} \{ b^T y + r^T s : y \in K_*, A^T y + R^T s = c \} \quad (D)
\end{align*}

Then
\begin{itemize}
\item \textbf{Primal-Dual Symmetry:} The duality is symmetric: (D) is conic along with (P)m and the problem dual to (D) is (equivalent to) (P).
\item \textbf{Weak Duality:} One has \(\text{Opt}(D) \leq \text{Opt}(P)\).
\item \textbf{Strong Duality:} Assume that one of the problems (P), (D) is strictly feasible and bounded, boundedness meaning on the feasible set the objective is bounded from below in the minimization and from above - in the maximization case. Then the other problem in the pair is solvable, and

\[ \text{Opt}(P) = \text{Opt}(D). \]
\end{itemize}

In particular, if both problems are strictly feasible (and thus both are bounded by Weak Duality), then both problems are solvable with equal optimal values.

In addition, if one of the problems is strictly feasible, then \(\text{Opt}(P) = \text{Opt}(D)\).
Refinement

Let the conic constraints in \((P), (D)\) be split into "general" and "polyhedral" parts, so that the problems read

\[
\begin{aligned}
\text{Opt}(P) &= \min_x \left\{ c^T x : Ax - b \in K, Px \geq p, Rx = r \right\} \\
\text{Opt}(D) &= \max_{y,z,s} \left\{ b^T y + p^T z + r^T s : y \in K^*, z \geq 0, A^T y + P^T z + R^T s = c \right\}
\end{aligned}
\]  

\((P)\) \hspace{1cm} \((D)\)

Then Strong Duality can be strengthened to the following claim: If one of the problems is essentially strictly feasible and bounded, then the other problem is solvable, and

\[\text{Opt}(P) = \text{Opt}(D).\]

In particular, if both problems are essentially strictly feasible, both are solvable with equal optimal values.

In addition, if one of the problems is essentially strictly feasible, then \(\text{Opt}(P) = \text{Opt}(D).\)
Note:
A. When no "general" conic constraint is present (i.e., in the LP situation) Refined Conic Duality Theorem is equivalent to LP Duality Theorem.
B. In general, the difference between the Strong Duality part of Conic duality Theorem and LP Duality Theorem is that the former requires (essential) strict feasibility, while the latter requires just feasibility. This difference "reflects reality" – when at least one of the primal-dual pair of problems is not essentially strictly feasible, various "pathologies" can arise. It can be shown by examples that it is possible that in a primal-dual pairs \((P), (D)\) of conic programs,
— one of the problems is strictly feasible and bounded (implying that the other problem is solvable and \(\text{Opt}(P) = \text{Opt}(D)\)), but is not solvable;
— one of the problems is solvable, and the other one is infeasible,
— both problems are solvable, but with different optimal values: \(\text{Opt}(D) < \text{Opt}(P)\).
Corollary [Optimality Conditions in Conic Programming] Consider primal-dual pair of conic problems

\[
\text{Opt}(P) = \min_x \{ c^T x : Ax - b \in K, Rx = r \} \quad (P)
\]
\[
\text{Opt}(D) = \max_{y,s} \{ b^T y + r^T s : y \in K^*, A^T y + R^T s = c \} \quad (D)
\]

and assume that both problems are strictly feasible. A pair \( x, [y; s] \) of primal and dual feasible solutions is comprised of optimal solutions to the respective problems

— [Zero Duality Gap] iff \( \text{DualityGap}(x, [y; s]) = c^T x - [b^T y + r^T s] \) is zero, and
— [Complementary Slackness] iff \( y^T [Ax - b] = 0 \).

Proof: We are in the situation when \( \text{Opt}(P) = \text{Opt}(D) \) by Strong Duality part of Conic Duality Theorem. Consequently, for primal-dual feasible \( x, [y; s] \) it holds

\[
\text{DualityGap}(x, [y; s]) = [c^T x - \text{Opt}(P)] + [\text{Opt}(D) - b^T y - r^T s]
\]

By primal-dual feasibility, both brackets are nonnegative, and their sum can be 0 iff \( c^T x = \text{Opt}(P) \) and \( b^T y + r^T s = \text{Opt}(D) \), as claimed in Zero Duality Gap.

Next, we have

\[
\text{DualityGap}(x, [y; s]) = c^T x - b^T y - r^T s = [A^T y + R^T s]^T x - b^T y - r^T s
\]
\[
= [Ax - b]^T y + [Rx - r]^T s = [Ax - b]^T y,
\]

implying that Zero Duality Gap is equivalent, for primal-dual feasible \( x, [y; s] \), to Complementary Slackness.
Example: Dual to the Steiner sum problem

♣ Steiner sum problem:

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} \|x - a_i\|_2.$$  \[ m > 1, a_1, ..., a_m \text{ are distinct points in } \mathbb{R}^n \]

“Cover story” (n = 2): There are m oil wells located at points $a_1, ..., a_m \in \mathbb{R}^2$. Where should one place an oil collector in order to minimize the total length of pipelines connecting the wells to the collector?

♣ The problem can be reformulated as conic:

$$\min_{t_1, ..., t_m, x} \left\{ \sum_{i=1}^{m} t_i : \left[ x - a_i ; t_i \right] \in \mathbb{L}^{n+1}, i = 1, ..., m \right\}$$  \[ (P) \]

Lorentz cones are self-dual, so that the problem dual to (S) is obtained by — assigning the constraints $[x - a_i; t_i]$ by Lagrange multipliers $[y_i; z_i] \in \mathbb{L}^{n+1}$ giving rise to the aggregated constraint

$$[\sum_i y_i^T]x + \sum_i z_i t_i \geq \sum_i y^T_i a_i$$

— imposing on the multipliers the restriction that the left hand side in the aggregated constraint is, identically in the primal variables $x, t_i$, equal to the primal objective $\sum_i t_i$, which amounts to

$$\sum_i y_i = 0, \quad z_1 = ... = z_m = 1$$

and maximizing under this restriction the right hand side of the aggregated constraint. Thus, the dual problem reads

$$\max_{y_1, ..., y_m} \left\{ \sum_i a_i^T y_i : \sum_i y_i = 0, \|y_i\|_2 \leq 1, i \leq m \right\}$$  \[ (D) \]
\[
\text{Opt}(P) = \min_{t_1, \ldots, t_m, x} \left\{ \sum_{i=1}^{m} t_i : [x - a_i; t_i] \in \mathbb{L}^{n+1}, i = 1, \ldots, m \right\} \quad (P)
\]
\[
\text{Opt}(D) = \max_{y_1, \ldots, y_m} \left\{ \sum_i a_i^T y_i : \sum_i y_i = 0, \|y_i\|_2 \leq 1, i \leq m \right\} \quad (D)
\]

- \((P)\) clearly is solvable and strictly feasible \(\Rightarrow\) \((D)\) is solvable and \(\text{Opt}(P) = \text{Opt}(D)\).
- From optimality conditions it is easily seen that
  - A point \(x\) distinct from \(a_1, \ldots, a_m\) is an optimal solution to the Steiner sum problem \(\text{iff}\)
    \[
    \sum_i \frac{a_i - x}{\|a_i - x\|_2} = 0.
    \]
  - Point \(x = a_\ell\) is an optimal solution \(\text{iff}\)
    \[
    \| \sum_{i \neq \ell} \frac{a_i - x}{\|a_i - x\|_2} \|_2 \leq 1.
    \]
In the simplest case of 3 points \( a_1 = A, a_2 = B, a_3 = C \) in 2D plane, the optimal solution is
— either the point from which all 3 sides of the triangle \( \triangle ABC \) are seen at the angle 120°

(such a point exists if angles of the triangle are < 120°),
— or the vertex of the triangle corresponding to the angle \( \geq 120° \), if such an angle is present:


Proof of Conic Duality Theorem

\[ \text{Opt}(P) = \min_x \{ c^T x : Ax - b \in K, Rx = r \} \quad (P) \]
\[ \text{Opt}(D) = \max_{y,s} \{ b^T y + r^T s : y \in K^*, A^T y + R^T s = c \} \quad (D) \]

Primal-Dual Symmetry: \((D)\) is a conic problem. To write down its dual, we rewrite it as a minimization problem

\[ -\text{Opt}(D) = \min_{y,s} \{ -b^T y - r^T s : y \in K^*, A^T y + R^T s = c \} \]

denoting the Lagrange multipliers for the constraints \(y \in K^*\) and \(A^T y + R^T s = c\) by \(z\) and \(-x\), the dual to dual problem reads

\[ \max_{z,x} \left\{ -c^T x : -Ax + z = -b, z \in (K^*)_*[=K], -Rx = -r \right\} \]

says that \(Ax - b \in K\)

Eliminating \(z\), we arrive at \((P)\). \hfill \square

Weak Duality: By construction of the dual. \hfill \square
Opt\( (P) \) = \( \min_x \{ c^T x : A x - b \in K, R x = r \} \) \hfill (P)
Opt\( (D) \) = \( \max_{y,s} \{ b^T y + r^T s : y \in K^*, A^T y + R^T s = c \} \) \hfill (D)

**Strong Duality:** We should prove that if one of the problems \( (P) \), \( (D) \) is strictly feasible and bounded, then the other problem is solvable with \( \text{Opt}(P) = \text{Opt}(D) \), or, which is the same by Weak Duality, with \( \text{Opt}(D) \geq \text{Opt}(P) \). By Primal-Dual Symmetry, we lose nothing when assuming that \( (P) \) is strictly feasible and bounded.

**Step 0:** Let us reduce the situation to the one when a strictly feasible solution to \( (P) \) is the origin. Specifically, denoting by \( \bar{x} \) a strictly feasible solution to \( (P) \) and passing in \( P \) from variable \( x \) to \( z = x - \bar{x} \), we arrive at the problem

\[
[\text{Opt}(P) - c^T \bar{x} =] \quad \text{Opt}(P') = \min_z \{ c^T z : A z - [b - A \bar{x}] \in K, R z = 0 \} \quad (P')
\]

with strictly feasible solution 0 and with the dual problem

\[
\text{Opt}(D') = \max_{y,s} \{ [b - A \bar{x}]^T y : y \in K^*, A^T y + R^T z = c \} \quad (D')
\]

Note that the feasible sets of \( (D) \) and \( (D') \) are the same, and on this feasible set, due to \( R \bar{x} = r \), we have

\[
[b - A \bar{x}]^T y = b^T y + r^T s - \bar{x}^T [A^T y + R^T s] = b^T y + r^T s - c^T \bar{x},
\]

implying that \( (D) \) and \( (D') \) simultaneously are solvable/unsolvable, and their optimal values, same as those of \( (P) \) and \( (P') \), differ by \( c^T \bar{x} \), so that \( \text{Opt}(P) = \text{Opt}(D) \) is equivalent to \( \text{Opt}(P') = \text{Opt}(D') \).

Thus, it suffices to prove Strong Duality in the case when \( \bar{x} = 0 \).
\[
\text{Opt}(P) = \min_x \{ c^T x : Ax - b \in K, Rx = 0 \} \quad (P)
\]
\[
\text{Opt}(D) = \max_{y, s} \{ b^T y : y \in K^*, A^T y + R^T s = c \} \quad (D)
\]
x = 0 is strictly feasible solution to (P), that is
\[-b \in \text{int} K.\]

**Step 1.** Let \( L = \{ x : Rx = 0 \} \). It may happen that \( c \) is orthogonal to \( L \) ("trivial case"). In this case the primal objective vanishes on the primal feasible set, that is, \( \text{Opt}(P) = 0 \), and \( c = A^T s_\ast \) for some \( s_\ast \), implying that \([y = 0; s_\ast]\) is a feasible solution to (D) with zero value of the dual objective. Thus, \( \text{Opt}(D) \geq 0 = \text{Opt}(P) \), implying that \( \text{Opt}(D) = \text{Opt}(P) \) and the solution \([0; s_\ast]\) is optimal for (D), so that Strong Duality holds true in the trivial case.

**Step 2.** Now let the projection \( \bar{c} \) of \( c \) on \( L \) be nonzero, implying that the set
\[
L_- = \{ x \in L : \bar{c}^T x < \text{Opt}(P) \} = \{ x \in L : c^T x < \text{Opt}(P) \}
\]
is nonempty. Note that the convex set \( M = \{ Ax - b : x \in L_- \} \) is nonempty and does not intersect \( K \). Consequently, \( M \) and \( K \) can be separated:
\[
\exists f \neq 0 : \inf_{z \in K} f^T z \geq \sup_{z \in M} f^T z.
\]
\( \overline{c}^T x \) is nonconstant on \( L = \{ x : Rx = 0 \} \) \hspace{1cm} (a)

\( f \neq 0 \) \hspace{1cm} (b)

\[ \inf_{z \in K} f^T z \geq \sup_{x} \{ f^T[Ax - b] : Rx = 0, \overline{c}^T x < \text{Opt}(P) \} \] \hspace{1cm} (c)

- \( K \) is a cone and \( \inf \) in (c) is finite \( \Rightarrow \) this inf is zero and \( f \in K^* \)

\( \Rightarrow \) \( \sup \) in (b) is \( \leq 0 \), so that (b) reads

\[ 0 \geq \sup_{x} \{ [A^T f]^T x : Rx = 0, \overline{c}^T x < \text{Opt}(P) \} - f^T b. \] \hspace{1cm} (d)

The maximization domain here is cut off linear space \( L = \{ x : Rx = 0 \} \) by strict linear inequality \( \overline{c}^T x < \text{Opt}(P) \) with nonconstant on \( L \) left hand side

\( \Rightarrow \) (d) implies that the orthogonal projection of \( A^T f \) onto \( L \) is \( \alpha \overline{c} \) with some \( \alpha \geq 0 \)

\( \Rightarrow \) (d) reads

\[ 0 \geq \sup_{x} \{ \alpha \overline{c}^T x : Rx = 0, \overline{c}^T x < \text{Opt} \} - f^T b = \alpha \text{Opt}(P) - f^T b. \] \hspace{1cm} (e)

Now, we have seen that \( f \in K^* \) and \( f \neq 0 \) by (b), while \(-b \in \text{int } K \Rightarrow f^T b > 0\), implying by (e) that \( \alpha > 0 \).

Setting \( y = \alpha^{-1} f \), we get \( y \in K^* \), and (e) reads \( y^T b \geq \text{Opt}(P) \). Besides this, the orthogonal projection of \( A^T y \) onto \( L \) is exactly the orthogonal projection \( \overline{c} \) of \( e \) onto \( L \)

\( \Rightarrow A^T y - c \) is orthogonal to \( L = \{ x : Rx = 0 \} \) \( \Rightarrow A^T y + R^T s = c \) for properly selected \( s \)

\( \Rightarrow [y; s] \) is dual feasible with the value of dual objective \( \text{Opt}(D) = \text{Opt}(P) \).
It remains to prove the if one of the problems \((P), (D)\) is strictly feasible, then \(\text{Opt}(P) = \text{Opt}(D)\). Indeed, by Primal-Dual Symmetry we lose nothing when assuming that \((P)\) is strictly feasible. The case when \((P)\) is also bounded has been considered; when \((P)\) is unbounded, \((D)\) is infeasible by Weak Duality; thus, in this case \(\text{Opt}(P) = \text{Opt}(D) = -\infty\). □
Consequences of Conic Duality Theorem

Question: When a linear vector inequality

\[ Ax \geq_K b \]  \hspace{1cm} (I)

has no solutions?

Sufficient condition for infeasibility: By “admissible aggregation” of (I) one can obtain a contradictory scalar inequality:

\[ \exists \lambda \geq_{K*} 0 : \ A^T \lambda = 0, \ \lambda^T b > 0. \]  \hspace{1cm} (II)

Indeed, assuming that \( Ax \geq_K b \) for some \( x \), we would get

\[ 0 \leq [\lambda]_T [Ax - b] = [A^T \lambda]_T x - \lambda^T b = -\lambda^T b < 0 \ - \ contradiction! \]
Conic Theorem on Alternative:

A. If (II) has a solution, then (I) has no solutions.

B. If (II) has no solutions, then (I) is "almost solvable," meaning that for every $\epsilon > 0$, you may perturb $b$ by no more than $\epsilon$ to get a solvable system (I):

$$\forall \epsilon > 0 \exists b' : \|b - b'| \leq \epsilon \text{ & } Ax \succeq_K b'$$

is solvable.

C. (II) has no solutions iff (I) is almost solvable.
\[ Ax \geq_K b \]
\[ \lambda \geq_K 0, A^T\lambda = 0, \lambda^T b > 0 \quad \text{(I)} \]

\[ \lambda \geq_K 0, A^T\lambda = 0, \lambda^T b > 0 \quad \text{(II)} \]

**Proof of CTA:** Let us fix \( f > K 0 \), and consider the conic program

\[
\text{Opt} = \min_{t,x} \{t : Ax \geq_K b - tf\} \quad (P)
\]

Since \( f > K 0 \), all pairs \([x = 0; t]\) with large enough positive \( t \) are strictly feasible solutions to \((P)\) *(since for large \( t > 0 \) we have \( tf - b = t(f - t^{-1}b) > K 0 \)).

**Claim:** \((I)\) is almost solvable iff Opt \( \leq 0 \).

**One direction:** If Opt \( \leq 0 \), then for every \( \delta > 0 \) \((P)\) has a feasible solution with \( t \leq \delta \), and, in addition, \((P)\) has a feasible solution with some nonnegative \( t \). Since the feasible set of \((P)\) is convex, for every \( \delta > 0 \) \((P)\) has a feasible solution \( x_\delta, t_\delta \) with \( t_\delta \in [0, 2\delta] \Rightarrow b_\delta := b - t_\delta f \) is such that \( Ax_\delta \geq_K b_\delta \). Since \( ||b_\delta - b|| = t_\delta ||f|| \leq 2\delta ||f|| \) and \( \delta \) can be made arbitrarily small, \((I)\) is almost solvable.

**Opposite direction:** If \((I)\) is almost solvable, then for every \( \delta > 0 \) there exist \( b_\delta, x_\delta \) such that \( Ax_\delta \geq_K b_\delta \) and \( ||b - b_\delta|| \leq \delta \). Since \( f > K 0 \), \( K \) contains a ball of radius \( r > 0 \) centered at \( f \), or, equivalently,

\[
\frac{||d||}{r} f \geq_K d \forall d.
\]

In particular, \( Ax_\delta \geq_K b_\delta \Rightarrow Ax_\delta \geq_K b + [b_\delta - b] \geq_K b - \frac{||b - b_\delta||}{r} f \geq_K b - \frac{\delta}{r} f \), whence Opt \( \leq \delta/r \) for all \( \delta > 0 \), that is, Opt \( \leq 0 \).

**Claim \Rightarrow CTA:** \((P)\) is strictly feasible, so that by Conic Duality Theorem Opt \( \leq 0 \) iff the optimal value in the problem

\[
\max_\lambda \{b^T\lambda : A\lambda = 0, \lambda \in K^*_+ \}
\]

\((D)\) dual to \((P)\) is \( \leq 0 \). The latter is the case iff \( b^T\lambda \leq 0 \) for every nonzero \( \lambda \in K^*_+ \) such that \( A\lambda = 0 \) (since for such \( \lambda \) it holds \( f^T\lambda > 0 \), so that after multiplying \( y \) by a positive scalar it becomes feasible for \((D))\), which is exactly the same as to say that \((II)\) has no solutions. \( \square \)
\[ \begin{align*}
Ax & \geq K b \\
\lambda & \geq K, 0, A^T \lambda = 0, \lambda^T b > 0 \quad \text{(I)}
\end{align*} \]

**CTA vs. GTA:** "Polyhedral analogy" of CTA is General Theorem on Alternative restricted to the situation where the system of (scalar) linear inequalities for which we want to certify insolvability contains nonstrict inequalities only. In this situation GTA is stronger than item C in CTA – in GTA "almost solvable" is simply "solvable."

♠ In the general conic case, "almost solvable" cannot be strengthened to "solvable," as is seen from the following example: the linear vector inequality

\[
Ax - b := [x + 1; x - 1; \sqrt{2}x] \geq_L 3 0
\]

\[
A = [1; 1; \sqrt{2}], \quad b = [-1; 1; 0]
\]

reads \(2x^2 + 2 \leq 2x^2\) and has no solutions. The associated system (II) reads

\[
\begin{align*}
\lambda_1 + \lambda_2 + \sqrt{2}\lambda_3 &= 0, \\
\sqrt{\lambda_1^2 + \lambda_2^2} &\leq \lambda_3, \quad \lambda_2 - \lambda_1 > 0.
\end{align*}
\]

that is,

\[
\|[−1; −1]\|_2\|[λ_1; λ_2]\|_2 = \sqrt{2}\sqrt{\lambda_1^2 + \lambda_2^2} \leq \sqrt{2}\lambda_3 = [−1; −1]^T[λ_1; λ_2]
\]

1.89
By Cauchy Inequality, the only possibility for this chain is for the vector \([\lambda_1; \lambda_2]\) to be proportional, with nonnegative coefficient, to \([-1; -1]\), which contradicts \(\lambda_1 - \lambda_2 > 0\). Thus, in our example both (I) and (II) have no solutions!
\[ Ax \geq_{K} b \]
\[ \lambda \geq_{K} 0, \ A^T \lambda = 0, \lambda^T b > 0 \] (II)

**What is going on:** The set of those \( b \)'s for which (I) is solvable is the convex set

\[ \{ b = Ax - u, \ x \in \mathbb{R}^n, \ u \in K \} \],

and the set \( B_* \) of those \( b \)'s for which (I) is almost solvable is the set of \( b \)'s which can be approximated to whatever high accuracy by points from \( B \), that is, \( B_* \) is the closure of \( B \).

By item C of CTA, (II) is solvable whenever \( b \) is outside of \( B_* \). **When \( B \) is closed**, to be outside of \( B \) and of \( B_* \) is one and the same

\[ \Rightarrow \text{When the set of those } b \text{'s for which (I) is solvable is not just convex, but is also closed, (II) is solvable whenever (I) is unsolvable.} \]

However, \( B \) is not necessarily closed, so that **in general solvability of (II) is only sufficient, but not necessary, condition for insolvability of (I).**

When \( K \) is a polyhedral cone, \( B \) is polyhedral (as the arithmetic sum of two polyhedral sets, \( B \) admits an immediate polyhedral representation)

\[ \Rightarrow B \text{ is automatically closed.} \]
Question: When a scalar inequality

\[ c^T x \geq d \]  \hspace{1cm} (S)

is a consequence of a vector inequality

\[ Ax \succeq_K b \]  \hspace{1cm} (V)

Answer: A. If (S) can be obtained from (V) and the trivial inequality

0 \geq -1 by "admissible linear aggregation:")

\[ \exists y \succeq_{K^*} 0 : A^T y = c \& y^T b \geq d, \]  \hspace{1cm} (*)

then (S) is a consequence of (V).

B. If (S) is a consequence of (V) and (V) is strictly feasible, then (S) can be obtained from (V) by admissible linear aggregation.

Both claims are immediate consequences of the Conic Duality Theorem as applied to the conic problem

\[ \text{Opt}(P) = \min_x \{ c^T x : Ax \succeq_K b \} \]

— (S) is nothing but the claim that \text{Opt}(P) \geq d, and A, B is what Weak, respectively, Strong, Duality says.
II. CONIC QUADRATIC PROGRAMMING
The $m$-dimensional Lorentz cone is

$$L^m = \{ x = [x_1; \ldots; x_m] \in \mathbb{R}^m : x_m \geq \sqrt{x_1^2 + \ldots + x_{m-1}^2} \}$$

By definition, $L^1 = \mathbb{R}_+$ ("empty sum equals zero").

A conic quadratic problem is a conic problem

$$\min_{x} \left\{ c^T x : A x - b \geq_{K} 0 \right\} \quad \text{(CP)}$$

for which the cone $K$ is a direct product of Lorentz cones:

$$K = L^{m_1} \times L^{m_2} \times \ldots \times L^{m_k} = \left\{ y = \begin{bmatrix} y[1] \\ y[2] \\ \vdots \\ y[k] \end{bmatrix} : y[i] \in L^{m_i}, \ i = 1, \ldots, k \right\}.$$  

• Thus, a conic quadratic problem is an optimization problem with linear objective and finitely many “conic quadratic constraints”:

$$\min_{x} \left\{ c^T x : A_i x - b_i \geq_{L^{m_i}} 0, \ i = 1, \ldots, k \right\}. \quad (\ast)$$
\[
\min_x \{ c^T x : A_i x - b_i \geq L^0_i, \ i = 1, \ldots, k \}. \tag{*}
\]

Representing

\[
[A_i, b_i] = \begin{bmatrix} D_i & d_i \\ p_i^T & q_i \end{bmatrix}
\]

\((q^i\) is a real), we may rewrite \((*)\) as

\[
\min_x \left\{ c^T x : \frac{\|D_i x - d_i\|_2}{p_i^T x - q_i}, \ i = 1, \ldots, k \right\}. \tag{CQ}
\]

\(\bullet\) A scalar linear inequality \(a^T x - b \geq 0\) is the same as the conic quadratic inequality \(a^T x - b \in \mathbb{L}^1\), so that adding to \((CQ)\) finitely many scalar linear inequalities, we do not vary the structure of the problem.
**Problem dual to Conic Quadratic Problem**

\[
\min_x \left\{ c^T x : \|D_i x - d_i\|_2 \leq p_i^T x - q_i, i = 1, \ldots, k \right\}.
\]  \hspace{1cm} (CQ)

\[
[D_i; p_i^T] x - [d_i; q_i] \geq_{L^m} 0
\]

**Fact:** Lorentz cones are self-dual: \((L^m)^* = L^m\).

Indeed,

\[
(L^m)^* = \{[y; s] : [y; s]^T [x; t] \geq 0 \forall (x; t : \|x\|_2 \leq t)\} = \{[y; s] : [y; s]^T [x; 1] \geq 0 \forall (x : \|x\|_2 \leq 1)\} = \{[y; s] : s \geq \max_{\|x\|_2 \leq 1} [-y^T x] = \{[y; s] : s \geq \|y\|_2\}.
\]

⇒ The problem dual to (CQ) reads

\[
\max_{[y_i; s_i], i \leq k} \left\{ \sum_i [y_i^T d_i + s_i q_i] : \|y_i\|_2 \leq s_i, i \leq k, \sum_i [D_i^T y_i + s_i p_i] = c \right\}
\]
Examples of CQP’s, I
Stable Grasp

♣ When an $N$-finger robot is capable to hold rigid body?
This is what happens at $i$-th contact point:

$\bullet$ [Coulomb’s Law] The friction force $F^i$ caused by the contact force $f^i$
is tangent to the surface of the body at $p^i$:

$$(F^i)^T \nu^i = 0,$$

and its magnitude is bounded by constant times the normal component of the external force:

$$\|F^i\|_2 \leq \mu (f^i)^T \nu^i$$

[$\mu > 0$: friction coefficient]
Assume that the body is affected by additional external forces (e.g., the gravity ones). From the viewpoint of Mechanics, all these forces can be represented by a single external force $F^\text{ext}$ (the sum of actual external forces) – and a torque $T^\text{ext}$ (the sum of vector products of the actual external forces and the points where the forces are applied).

The body can be in static equilibrium iff the total force acting at the body and the total torque are zero:

$$\sum_{i=1}^{N} (f^i + F^i) + F^\text{ext} = 0$$
$$\sum_{i=1}^{N} p^i \times (f^i + F^i) + T^\text{ext} = 0$$

$u \times v$: vector product of $u, v \in \mathbb{R}^3$

Assume $f^i, F^\text{ext}, T^\text{ext}$ are given. The nature will try to adjust the friction forces $F^i$ to satisfy the equilibrium constraints (1) along with the ”friction constraints”

$$[\nu^i]^T F^i = 0, \|F^i\|_2 \leq \mu[\nu^i]^T f^i, \ i = 1, \ldots, N$$

If it is possible, the body will be held by the robot ("stable grasp"), otherwise it will somehow move.
Conclusion: Possibility of stable grasp is equivalent to solvability of system of conic quadratic constraints

\[
\begin{align*}
\sum_{i=1}^{N} (f_i + F_i) + F^\text{ext} &= 0, \\
\sum_{i=1}^{N} p_i \times (f_i + F_i) + T^\text{ext} &= 0, \\
[\nu_i]^T F_i &= 0, \\
\|F_i\|_2 &\leq \mu [\nu_i]^T f_i
\end{align*}
\]

with variables \( F_i, i = 1, \ldots, N \).

\( \Rightarrow \) Various grasp-related optimization problems, like

Given
— external force \( F^\text{ext} \),
— the direction \( e^\text{ext} \) of external torque,
— the directions \( u^i \) of forces exerted by robot's fingers,
— ranges \([0, f^i_{\text{max}}]\) of magnitudes of the forces exerted by robot's fingers:

\[
f^i = \lambda_i u^i, \quad \lambda_i \in [0, f^i_{\text{max}}],
\]

find the largest possible magnitude \( T \) of the external torque still allowing for stable grasp.

can be posed as conic quadratic problems.
**Example.** A 4-finger robot should hold a cylinder:

![Diagram of a cylinder with fingers](image)

**Perspective, front and side views**

The external torque is directed along the cylinder axis. What is the largest magnitude of the torque still allowing for stable grasp? This is the conic quadratic problem

\[
\begin{align*}
\max_{T,F^i,\lambda_i} \left\{ T : \begin{array}{l}
\sum_i (\lambda_i u^i + F^i) + F^\text{ext} = 0 \\
\sum_i p^i \times (\lambda_i u^i + F^i) + T^\text{ext} = 0 \\
\|F^i\|_2 \leq \mu [\nu^i]^T u^i \lambda_i, [\nu^i]^T F^i = 0, i \leq N \\
0 \leq \lambda_i \leq f^i_{\max},, i \leq N
\end{array} \right\}.
\end{align*}
\]
What can be expressed via conic quadratic constraints?

♣ Normally, an initial form of an optimization model is

\[
\min \{ f(x) : x \in X \}, \quad X = \bigcap_{i=1}^{m} X_i \quad \text{[usually } X_i = \{ x : g_i(x) \leq 0 \} \]  

We can always make the objective linear:

\[
\min f(x) \Leftrightarrow \min_{y=[x;t]\in Y} t \quad \text{[} Y = \{ [x;t] : x \in X, t \geq f(x) \} \]  

From now on, assume that the objective is linear, so that the original problem is

\[
\min_x \{ c^T x : x \in X \} \quad [X = \bigcap_{i=1}^{m} X_i] \quad \text{(Ini)}
\]

♣ Question: When (Ini) can be reformulated as a conic quadratic problem?
\[
\min_x \{ c^T x : x \in X \} \quad [X = \bigcap_{i=1}^m X_i] \quad \text{(Ini)}
\]

**Question:** When (Ini) can be reformulated as a conic quadratic problem?

**Answer:** This is the case when \( X \) is a **Conic Quadratic representable** (CQr) set.

**Definition.** Let \( X \subset \mathbb{R}^n \). We say that \( X \) is CQr, if \( X \) admits Conic Quadratic Representation (CQR)

\[
X = \{ x \in \mathbb{R}^n : \exists u \in \mathbb{R}^m : Px + Qu - r \in K \}, \quad \text{(CQR)}
\]

where \( K \) is a direct product of Lorentz cones, that is, \( X \) can be represented as a projection onto the plane of \( x \)-variables of the solution set of a conic constraint in \((x,u)\)-variables, the cone being a direct product of Lorentz cones.

**Equivalently:** \( X \subset \mathbb{R}^n \) is CQr, if \( x \in X \) if and only if \( x \) can be extended, by properly selected "certificate" \( u \in \mathbb{R}^m \), to a solution to a system of conic quadratic inequalities in variables \( x, u \). Every system with this property is a Conic Quadratic Representation of \( X \).
\[ X = \{ x \in \mathbb{R}^n : \exists u \in \mathbb{R}^m : Px + Qu - r \in K \}, \]  
(CQR)

**Immediate observation:** Given Conic Quadratic Representation (CQR) of \( X \), the problem \( \min_{x \in X} c^T x \) is equivalent to the conic quadratic program

\[ \min_{x,u} \left\{ c^T x : Px + Qu - r \in K \right\}, \]

equivalence meaning that \( x \) is feasible for the former problem iff \( x \) can be extended to a feasible solution to the latter problem. Note that this extension preserves the value of the objective.
**Example:** Consider the program

\[
\min_x \{ x : x^2 + 2x^4 \leq 1 \} \tag{Ini}
\]

A CQR for \( X = \{ x : x^2 + 2x^4 \leq 1 \} \) can be obtained as follows:

\[
x^2 + 2x^4 \leq 1 \iff \exists t_1, t_2 : \begin{cases} x^2 \leq t_1 \\ t_1^2 \leq t_2 \\ t_1 + 2t_2 \leq 1 \end{cases}
\]

and

\[
s^2 \leq r \iff 4s^2 + (r - 1)^2 \leq (r + 1)^2 \iff \begin{bmatrix} 2s \\ r - 1 \\ r + 1 \end{bmatrix} \geq_{L^3} 0,
\]

\[
\Rightarrow X = \left\{ x : \exists t_1, t_2 : \begin{bmatrix} 2x \\ t_1 - 1 \\ t_1 + 1 \end{bmatrix} \geq_{L^3} 0, \quad \begin{bmatrix} 2t_1 \\ t_2 - 1 \\ t_2 + 1 \end{bmatrix} \geq_{L^3} 0, \quad t_1 + 2t_2 \leq 1 \right\},
\]

“says” that \( x^2 \leq t_1 \)

“says” that \( t_1^2 \leq t_2 \)

and (Ini) is the conic quadratic program

\[
\min_{x, t_1, t_2} \left\{ x : \begin{bmatrix} 2x \\ t_1 - 1 \\ t_1 + 1 \end{bmatrix} \geq_{L^3} 0, \quad \begin{bmatrix} 2t_1 \\ t_2 - 1 \\ t_2 + 1 \end{bmatrix} \geq_{L^3} 0, \quad t_1 + 2t_2 \leq 1 \right\}.
\]

2.11
Definition. Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a function. We say that $f$ is \textit{Conic Quadratic representable} (CQr), if its epigraph

$$ \text{Epi}\{f\} = \{[x; t] \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq t\} $$

is a CQr set. Every CQR of Epi$\{f\}$ is called a Conic Quadratic Representation (CQR) of $f$. Thus, CQR of $f$ is the equivalence

$$ t \geq f(x) \Leftrightarrow \exists u : Px + tp + Qu - r \in K, $$

where $K$ is a direct product of Lorentz cones.

Example: The function $f(x) = x^2 + 2x^4 : \mathbb{R} \to \mathbb{R}$ is CQr:

$$ t \geq x^2 + 2x^4 \Leftrightarrow \exists t_1, t_2 : \begin{bmatrix} 2t_1 \\ t_1 - 1 \\ t_1 + 1 \end{bmatrix} \geq_{L^3} 0, \begin{bmatrix} 2t_1 \\ t_2 - 1 \\ t_2 + 1 \end{bmatrix} \geq_{L^3} 0, t_1 + 2t_2 \leq t $$

Immediate Observation: Level sets $\{x : f(x) \leq a\}$ of a CQr function $f : \mathbb{R}^n \to \mathbb{R}$ are CQr sets with CQR's readily given by a CQR of $f$:

$$ t \geq f(x) \Leftrightarrow \exists u : Px + pt + Qu - r \in K $$

\[\implies \{x : f(x) \leq a\} = \{x : \exists u : Px + pa + Qu - r \in K\} \]
**Immediate Observation:** Given CQR’s of a CQr function $f$ and a CQr set $X$, minimization of $f$ over $X$ reduces straightforwardly to a conic quadratic problem:

\[
\begin{align*}
  t \geq f(x) & \iff \exists u : P_f x + t p_f + Q_f u - r_f \in K_f \\
  x \in X & \iff \exists v : P_X x + Q_X v - r_X \in K_X
\end{align*}
\]

\[
\min_{x \in X} f(x) \iff \min_{t,x,u,v} \left\{ t : \begin{array}{c}
  P_f x + t p_f + Q_f u - r_f \in K_f \\
  P_X x + Q_X v - r_X \in K_X
\end{array} \right\}
\]
Calculus of CQr functions/sets

**Fact:** CQr functions/sets admit a fully algorithmic calculus: basic *convexity-preserving* operations with functions/sets as applied to CQr operands, produce CQr results, and CQR’s of results are readily given by CQR’s of operands.

**Note:** "Convexity-preserving" is crucial here: convexity is built-in property of CQr functions/sets, so that operations which do not preserve convexity (like taking union of two sets) do not preserve, in general, conic quadratic representability.
Calculus of CQR’s: Raw Materials. The following functions/sets are CQR with explicit CQR’s:

1. Closed half-spaces and affine functions

\[ X = \{ x : a^T x - b \geq 0 \} \] — this is CQR

\[ \text{Epi}\{a^T x + b\} = \{ [x; t] : t - a^T x - b \geq 0 \} \] — this is CQR

2. Euclidean norm \( f(x) = \|x\|_2 : \mathbb{R}^n \to \mathbb{R} \):

\[ \text{Epi}\{f\} := \{ [x; t] : t \geq \|x\|_2 \} = \{ [x; t] \in \mathbb{L}^{n+1} \} \]

3. Squared Euclidean norm \( f(x) = x^T x : \mathbb{R}^n \to \mathbb{R} \):

\[ t \geq x^T x \iff (t + 1)^2 \geq (t - 1)^2 + 4x^T x \iff [2x; t - 1; t + 1] \in \mathbb{L}^{n+2} \]

4. Fractional-quadratic function \( f(x, s) = \begin{cases} \frac{x^T x}{s}, & s > 0 \\ 0, & x = 0, s = 0 \\ +\infty, & \text{all remaining cases} \end{cases} \) [\( x \in \mathbb{R}^n, s \in \mathbb{R} \)]:

\[ \text{Epi}\{f\} = \{ [x; s; t] : [2x; t - s; t + s] \in \mathbb{L}^{n+2} \} \]

5. Branch of hyperbola \( \{(t, s) \in \mathbb{R}^2 : ts \geq 1, t, s \geq 0\} : \)

\[ \{(t, s) : ts \geq 1, t, s \geq 0 \} = \{(t, s) : [2; t - s; t + s] \in \mathbb{L}^3 \} \]

2.15
6. Rotated Lorenz cone $X = \left\{ [x; t; s] : x^T x \leq ts, t, s \geq 0 \right\} \subset \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}$:

$$\left\{ [x; t; s] : x^T x \leq ts, t, s \geq 0 \right\} = \left\{ [x; t; s] : [2x; t - s; t + s] \in L^{n+2} \right\}$$
Operations preserving CQ-representability of sets

**S.A. Taking finite intersections:** Intersection of CQr sets \( X_i, i \leq N \), is CQr:

\[
X_i = \left\{ x \in \mathbb{R}^n : \exists u^i : P_ix + Q_iu^i - r_i \in K_i \right\}, i \leq N
\]

\[
\bigcap_{i \leq N} X_i = \{ x : \exists u = [u^1; \ldots; u^N] : P_ix + Q_iu^i - r_i \in K_i, i \leq N \}
\]

In particular, a polyhedral set \( \{ x : Ax - b \geq 0 \} \) is CQr (as the intersection of closed half-spaces, which are CQr), and intersecting a CQr set with the solution set of a finite system of nonstrict linear inequalities preserves CQ-representability.

**S.B. Taking direct products.** Direct product of CQr sets \( X_i \subset \mathbb{R}^{n_i}, i \leq N \), is CQr:

\[
X_i = \left\{ x^i \in \mathbb{R}^{n_i} : \exists u^i : P_ix^i + Q_iu^i - r_i \in K_i \right\}, i \leq N
\]

\[
X_1 \times \ldots \times X_N := \{ [x^1; \ldots; x^N] : x^i \in X_i \} = \{ [x^1; \ldots; x^N] : \exists u = [u^1; \ldots; u^N] : P_ix^i + Q_iu^i - r_i \in K_i, i \leq N \}
\]

2.17
**S.C. Taking affine images:** If $X \subset \mathbb{R}^n$ is CQr and $x \mapsto Ax + b : \mathbb{R}^n \to \mathbb{R}^k$ is an affine mapping, then the set $AX + b := \{y = Ax + b : x \in X\}$ is CQr:

\[
X = \{x : \exists u : Px + Qu - r \in K\}
\]

\[
AX + b = \{y : \exists [x; u] : y = Ax - b, Px + Qu - r \in K\}
\]

\[
y - [Ax - b] \in \mathbb{R}_+^k, [Ax - b] - y \in \mathbb{R}_+^k
\]

and all cones involved are direct products of Lorentz cones.

**Corollary:** Let $S$ be a finite system of conic quadratic inequalities in variables $(x, u)$. Then the set

\[
X = \{x : \exists u : (x, u) \text{ solves } S\}
\]

is CQr.

Indeed, the solution set $Y$ of $(S)$ clearly is CQr with CQR given by $(S)$, and $X$ is the linear image of $Y$.

**S.D. Taking inverse affine images.** If $X \subset \mathbb{R}^n$ is CQr and $y \mapsto A(x) = Ax + b : \mathbb{R}^k \to \mathbb{R}^n$ is an affine mapping, then the set $A^{-1}(X) := \{y : Ay + b \in X\}$ is CQr:

\[
X = \{x : \exists u : Px + Qu - r \in K\}
\]

\[
A^{-1}(X) = \{y : \exists u : P[Ay + b] + Qu - r \in K\}
\]
**S.E. Taking arithmetic sums:** If sets $X_i \subset \mathbb{R}^n$, $i = 1, ..., N$, are CQr, so is their arithmetic sum $X = X_1 + ... + X_N := \{x = x_1 + ... + x_N : x_i \in X_i, i = 1, ..., N\}$:

$$X_i = \left\{ x : \exists u^i : P_i x + Q_i u^i - r_i \in K_i \right\}, i \leq N$$

$$X_1 + ... + X_N = \{x : \exists x^i, u^i, i \leq N : P_i x^i + Q_i u^i - r_i \in K_i, i \leq N, x = \sum_i x^i\}$$

Alternatively: $X$ is the image of the direct product $Y = X_1 \times ... \times X_N$ under the linear mapping

$$y \equiv (x_1, ..., x_N) \mapsto x_1 + ... + x_N,$$

and both operations preserve CQ representability.
Several more advanced convexity-preserving operations "behave well" on CQr sets under mild regularity assumptions:

**S.F*. Passing from a set to its support function and polar.** Let $X \subset \mathbb{R}^n$ be a nonempty convex set. Its *support function* is defined as

$$
\phi_X(y) = \sup_x \{y^T x : x \in X\} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}.
$$

The support function of $X$ is the same as the support function of the closure of $X$, and the function "remembers" this closure: if $X, X'$ are nonempty convex sets, then $\phi_X \equiv \phi_{X'}$ iff $\text{cl}X = \text{cl}X'$.

**Fact:** If $X \subset \mathbb{R}^n$ is a nonempty convex set given by essentially strictly feasible CQR, then $\phi_X(\cdot)$ is CQr:

$$
X = \left\{ x : \exists u : Px + Qu - r \in K \right\}
$$

$$
t \geq \phi_X(y) \iff t \geq \sup_{x,u} \{y^T x : Px + Qu \geq_K r\}
$$

$$
\iff t \geq \min_\lambda \left\{ -r^T \lambda : P^T \lambda + y = 0, Q^T \lambda = 0, \lambda \in K_* \right\}
$$

$$
\iff \left\{ [y; t] : \exists \lambda : P^T \lambda + y = 0, Q^T \lambda = 0, t + r^T \lambda \geq 0, \lambda \in K_* \right\} \subseteq [= K]
$$

where the second and the third $\iff$ are due to (refined) Strong Duality.
**Corollary:** When \( X \) is CQr with essentially strictly feasible CQR, the polar of \( X \)

\[
Polar(X) = \{y : y^T x \leq 1 \forall x \in X\}
\]

is CQr.

Indeed, \( Polar(X) = \{y : \phi_X(y) \leq 1\} \), and a level set of CQr function is CQr with CQR readily given by a CQR of the function.

**Fact:** Polar (\( X \)) always is closed, convex, and contains the origin.

**Fact:** When \( X \) is a closed convex set containing the origin, so is Polar (\( X \)), and the polar of the polar is \( X \).

**Fact:** The larger is a set, the smaller is its polar:

\[X \subset Y \Rightarrow 0 \in Polar(Y) \subset Polar(X)\]
S.G*. Passing from a set to its recessive cone. Let $X$ be a nonempty closed convex set. Its *recessive cone* is defined as

$$\text{Rec}(X) = \{d : \exists \bar{x} \in X : \bar{x} + td \in X \ \forall t \geq 0\}.$$}

i.e., $\text{Rec}(X)$ is comprised of directions $d$ of all rays (treating a point as a ray with zero direction) contained in $X$. It is easily seen that

- If $X$ contains a ray, directed by $d$, then the parallel ray emanating from whatever point of $X$, is contained in $X$:

$$X = X + \text{Rec}(X)$$

- $\text{Rec}(X)$ is closed convex cone.
- $\text{Rec}(X) = \{0\}$ iff $X$ is bounded.
- For a polyhedral set $X = \{x : Ax \leq b\}$ it holds

$$\text{Rec}(X) = \{x : Ax \leq 0\}.$$
Fact: Let a CQr set \( X = \{x : \exists u : Px + Qu - r \in K\} \) be nonempty. Then

A. The CQr set \( R = \{x : \exists u : Px + Qu \in K\} \) is a convex cone contained in the recessive cone of \( \text{cl}X \).

B. Let the intersection of the image space of \( Q \) and \( K \) be trivial – the origin: \( Qu \in K \Rightarrow Qu = 0 \). Then \( X \) is closed and \( R = \text{Rec}(X) \).

Proof of Lemma. Let \( X \ni x_i \to \bar{x}, i \to \infty \). By Lemma, the sequence \( u = u_{x_i} \) is bounded; passing to subsequence, we can assume that \( u_i \to u, i \to \infty \). Since \( Px_i + Qu_i - r \in K \), we get \( Px + Qu - r \in K \), that is, \( x \in X \), Thus, \( X \) is closed. Next, \( d \in \text{Rec}(X) \) \& \( \bar{x} \in X \) \& \( t > 1 \Rightarrow \exists u_t : P(x + td) + Qu^t - r \in K \Rightarrow P[x + td] + Qu_t - r \in K \) with \( u_t = u_{P[x+td]-r} \Rightarrow Pd + Qt^{-1}u_t + [Px - r]/t \in K \), and \( v_t = t^{-1}u_t \) remain bounded as \( t \to \infty \) by Lemma. Selecting \( t_j \to \infty, j \to \infty \), such that \( v_{t_j} \to \bar{v} \) as \( j \to \infty \), we have
\[
Pd + Qv = \lim_{j \to \infty} [Pd + Qt^{-1}v_t + (Px - r)/t_j] \in K,\]
Thus, \( d \in R \), and therefore \( \text{Rec}(X) \subset R \), which combines with A to imply \( R = \text{Rec}(X) \). \( \square \)

Proof of Lemma. Let \( Z = \{z : \exists u : Qu + z \in K\} \). For \( z \in Z \), let \( u_z \) be the \( \|\cdot\|_2 \)-smallest vector \( u \) such that \( Qu + z \in K \); clearly, \( u_z \) exists, \( u_0 = 0, u_z \in [\text{Ker}Q]^{\perp} \), and \( u_{t_z} = tu_z \) when \( t > 0 \). It suffices to prove that \( \|u_z\|_2 \leq C\|z\|_2 \) for some \( C < \infty \). Assuming the opposite, there exists a sequence \( z_i \in Z \) such that \( \|u_z\|_2 > i\|z\|_2 \Rightarrow u_z, i \neq 0 \). Setting \( \zeta_i = z_i/\|u_z\|_2, u_i = u_{\zeta_i} = u_{z_i}/\|u_{z_i}\| \), we get \( u_i \in [\text{Ker}Q]^{\perp}, \|u_i\|_2 = 1, Qu_i + \zeta_i \in K \) and \( \zeta_i \to 0, i \to \infty \). For properly selected \( i_1 < i_2 < \ldots \) we have \( u_{i_j} \to u, j \to \infty \), implying \( \|u\|_2 = 1, u \in [\text{Ker}Q]^{\perp} \) and \( Qu \in K \). Since \( 0 \neq u \in [\text{Ker}Q]^{\perp} \), we have also \( Qu \neq 0 \), which under the premise of B is impossible. \( \square \)
Note: When our sufficient condition $Qu \geq_K \Rightarrow Qu = 0$ for the validity of the implication

$$X = \{ x : \exists u : Px + Qu - r \in K \} \Rightarrow X \text{ is closed} \quad \& \quad \text{Rec}(X) = R := \{ d : \exists v : Pd + Qv \in K \}$$

is violated, the implication may fail to be true.

However: when the condition is ”severely violated:” $\exists u : Qu >_K 0$, the implication holds true by trivial reasons – in this case $X = R$ is the entire space!
S.G*. Taking conic hull. The conic hull of a nonempty convex set $X \subset \mathbb{R}^n$ is CQr is defined as

$$X^+ := \{[x; t] : t > 0, x/t \in X\}$$

To get $X^+$, we lift $X \subset \mathbb{R}^n$ to get the set $X_+ = \{[x; 1] : x \in X\} \subset \mathbb{R}^{n+1}$; $X^+$ is the union of all (open) rays in $\mathbb{R}^{n+1}$ emanating from the origin and crossing $X_+$, i.e., $X^+ \cup \{0\}$ is the smallest cone containing $X_+$.

**Fact:** The conic hull $X^+$ of CQr set $X$ is CQr:

$$X = \{x : \exists u : Px + Qu - r \in K\}, \quad X^+ = \{[x; t] : t > 0, x/t \in X\}$$

$$\Downarrow$$

$$X^+ = \{[x; t] : \exists u, s : Px + Qu - tr \in K, t \geq 0, s \geq 0, ts \geq 1\} = \{[2; t-s; t+s] \in L^3\}.$$

Indeed, $\{[x; t] : t > 0, x/t \in X\} = \{[x; t] : \exists u : t > 0, P[x/t] + Qu - r \in K\} = \{[x; t] : \exists u : t > 0, Px + Qu - tr \in K, s \geq 0, t \geq 0, st \geq 1\}.$
\[ X^+ = \{[x; t] : t > 0, t^{-1}x \in X\} \text{ [conic hull of } X]\]

**Note:** If nonempty CQr set \( X = \{x : \exists u : Px + Qu - r \in K\} \) is closed, then the CQr set

\[ \hat{X}^+ = \{[x; t] : \exists u : Px + Qu - tr \in K, t \geq 0\} \]

is "in-between" the complete conic hull \( \bar{X}^+ = X^+ \cup \{0\} \) of \( X \) and the closed conic hull \( \text{cl}X^+ = \text{cl} \bar{X}^+ \) of \( X \):

\[ \bar{X}^+ := X^+ \cup \{0\} \subset \hat{X}^+ \subset \text{cl}X^+ = \text{cl} \bar{X}^+. \]

If \( X \) is closed and bounded, then \( \bar{X}^+ \) is closed, so that in this case

\[ \bar{X}^+ = \hat{X}^+ = \text{cl} \bar{X}^* \]

is CQr.

**Proof.** \( \hat{X}^+ \) clearly contains the origin and we already known that it contains the conic hull \( X^+ = \{[x; t] \in \bar{X}^+ : t > 0\} \) of \( X \Rightarrow \bar{X}^+ \subset \hat{X}^+ \). On the other hand, let \([x; t] \in \hat{X}^+ \) and \( \bar{x} \in X \), so that \( t \geq 0 \), \( Px + Qu - tr \in K \), and \( P\bar{x} + Qv - r \in K \) for some \( u, v \). Then for every \( \epsilon \in (0, 1) \) we have

\[
P\left[ x + \epsilon \bar{x} \right] + Q[u + \epsilon v] - [t + \epsilon] r \in K \Rightarrow [x_\epsilon; t_\epsilon] \in X^+.\]
Since \([x_\epsilon; t_\epsilon] \to [x; t]\) as \(\epsilon \to +0\), we get \([x; t] \in \text{cl} X^+\). Thus, \(\tilde{X}^+ \subset \text{cl} X^+\).

The fact that \(\bar{X}^+\) is closed whenever \(X\) is bounded and closed is immediate. Let \(\bar{X}^+ \ni [x_i; t_i] \to [x; t], i \to \infty\); we should prove that \([x; t] \in \bar{X}^+\). If infinitely many of \(t_i\) are zeros, then \([x; t]\) is the origin (since \([x; 0] \in \bar{X}^+\) iff \(x = 0\)), and the origin does belong to \(\bar{X}^+\). When only finitely many of \(t_i\) are zeros, then the vectors \(y_i = x_i/t_i\) are well defined for all large enough \(i\) and belong to \(X\), and thus form a bounded sequence. Passing to a subsequence, we can assume that \(y_i \to y\) as \(i \to \infty\), and \(y \in X\) since \(X\) is closed. We see that \([x_i; t_i] = t_i[y_i; 1]\) with \(y_i \to y \in X, i \to \infty\), implying that \([x; t] = \lim_{i \to \infty} [x_i; t_i] = \lim_{i \to \infty} t_i[y_i; 1] = t[y; 1]\).

Since \(t \geq 0\) and \(y \in X\), we see that \([s; t] \in \bar{X}^+\). \(\square\)
S.H*. Taking convex hulls of finite unions. Let $X_i \subset \mathbb{R}^n$, $i = 1, \ldots, N$, be nonempty closed CQr sets: $X_i = \{ x : \exists u^i : P_i x + Q_i u^i - r_i \in K_i \}$, and $\hat{X}$ be the convex hull of their union:

$$\hat{X} = \text{Conv}(X_1 \cup \ldots \cup X_N).$$

Then the CQr set

$$\tilde{X} = \left\{ x : \exists y^i, u^i, \lambda_i, i \leq N : \lambda_i \geq 0, \sum_i \lambda_i = 1, x = \sum_i y^i \mid P_i y^i + Q_i u^i - \lambda_i r_i \in K_i, i \leq N \right\}$$

is in-between $\hat{X}$ and $\text{cl}\hat{X}$: $\hat{X} \subset \tilde{X} \subset \text{cl}\hat{X}$. In particular, when $\hat{X}$ is closed (which definitely is the case, e.g., when all $X_i$ are bounded), then $\hat{X} = \tilde{X}$ is Cr.

Proof. When $x \in \hat{X}$, we have $x = \sum_i \lambda_i x^i$ with $\lambda_i \geq 0$, $\sum_i \lambda_i = 1$ and $x^i \in X_i$, that is, $P_i x^i + Q_i v^i - r_i \in K_i$ for some $v^i$. Setting $y_i = \lambda_i x^i$, $u^i = \lambda_i v^i$, we get $P_i y^i + Q_i u^i - \lambda_i r_i \in K_i$ and $x = \sum_i y^i$, whence $x \in \tilde{X}$. Thus, $\hat{X} \subset \tilde{X}$. Now let $x \in \tilde{X}$ and $\tilde{y}^i$ be such that $N \tilde{y}^i \in X_i$, so that

$$\exists(y^i, u^i, \tilde{u}_i, \lambda_i) : \lambda_i \geq 0, \sum_i \lambda_i = 1, x = \sum_i y^i, P_i y^i + Q_i u^i - \lambda_i r_i \in K_i, P_i \tilde{y}^i + Q_i \tilde{u}_i - N^{-1} p_i \in K_i.$$

For $\epsilon \in (0, 1]$ it holds $P_i [(1 - \epsilon) y^i + \epsilon \tilde{y}^i] + Q_i [(1 - \epsilon) u^i + \epsilon \tilde{u}_i] - [(1 - \epsilon) \lambda_i + \epsilon N^{-1}] r_i \in K_i, i \leq N$, whence $z^i := y^i / \lambda_i \in X_i, i \leq N$, and since $\sum_i \lambda_i = 1$ and $\lambda_i \geq 0$, we get $x_\epsilon := \sum_i y^i = \sum_i \lambda_i z^i \in \hat{X}$. Then $\epsilon \to +0$, $x_\epsilon \to x = \sum_i y^i$, whence $x \in \text{cl}\hat{X}$. Thus, $\tilde{X} \subset \text{cl}\hat{X}$. □
Operations preserving CQ-representability of functions

F.A. Restricting onto CQr set. If \( f(x) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \) is CQr function and \( X \subset \mathbb{R}^n \) is CQr set, then the restriction \( f_X(x) = \begin{cases} f(x), & x \in X \\ +\infty, & \text{otherwise} \end{cases} \) is CQr:

\[
\begin{align*}
 t \geq f(x) & \iff \exists u : P_f x + tp + Q_f u - r_f \in K_f \\
 X & = \{ x : \exists v : P_X x + Q_X v - r_X \in K_X \}
\end{align*}
\]

\[
\downarrow
\]

\[
 t \geq f_X(x) \iff \exists u, v : P_f x + tp + Q_f u - r_f \in K_f, P_X x + Q_X v - r_X \in K_X
\]

F.B. Taking finite maxima. If \( f_i : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}, \ i = 1, \ldots, N, \) are CQr, then so is their maximum \( f(x) = \max_i f_i(x) \).

Indeed, \( \text{Epi}\{f\} = \bigcap_i \text{Epi}\{f_i\} \), and intersection of finitely many CQr sets is CQr.
**F.C. Summation with nonnegative weights.** If functions \( f_i : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}, \ i = 1, \ldots, N, \) are CQr and \( \alpha_i \geq 0, \) then the function

\[
f(x) = \sum_{i=1}^{n} \alpha_i f_i(x)
\]
is CQr. Indeed, assuming w.l.o.g. that \( \alpha_i > 0, \ i \leq N, \) we have

\[
t \geq f_i(x) \iff \exists u^i : P_i x + t p_i + Q_i u^t - r_i \in K_i, i \leq N
\]

\[
t \geq \sum_i \alpha_i f_i(x) \iff \exists t_i, u_i, i \leq N : P_i x + t_i p_i + Q_i u^i - r_i \in K_i \ \forall i, \ t \geq \sum_i \alpha_i t_i.
\]

**F.D. Direct summation.** If \( f_i : \mathbb{R}^{n_i} \to \mathbb{R} \cup \{+\infty\}, \ i = 1, \ldots, N, \) are CQr, so is

\[
f(x^1, \ldots, x^N) = \sum_{i=1}^{N} f_i(x^i) : \mathbb{R}^{n_1}_{x^1} \times \ldots \times \mathbb{R}^{n_N}_{x^N} \to \mathbb{R} \cup \{+\infty\}:
\]

\[
t \geq f_i(x^i) \iff \exists u^i : P_i x^i + t p_i + Q_i u^t - r_i \in K_i, i \leq N
\]

\[
t \geq \sum_i f_i(x^i) \iff \exists t_i, u_i, i \leq N : P_i x^i + t_i p_i + Q_i u^i - r_i \in K_i \ \forall i, \ t \geq \sum_i t_i.
\]
F.E. Affine substitution of argument. If $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is CQr and $y \mapsto Ay + b : \mathbb{R}^k \to \mathbb{R}^n$ is an affine mapping, then the superposition $g(y) = f(Ay + b)$ is CQr:

$t \geq f(x) \iff \exists u : Px + tp + Qu - r \in K$

$t \geq g(y) \iff \exists u : P[Ay + b] + tp + Qu - r \in K$

2.30
F.F. Taking superposition. Let $F(y): \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ and $f_i(x): \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}, \ i = 1, \ldots, m,$ be CQr. Assume that $F(y)$ is nondecreasing in every one of $y_i$. Then the superposition

$$G(x) = \begin{cases} 
F(f_1(x), \ldots, f_m(x)), & f_i(x) < +\infty, \ i \leq m \\
+\infty, & \text{otherwise}
\end{cases}$$

is CQr:

$$t \geq F(y) \iff \exists u : Py + tp + Qu - r \in K$$
$$t \geq f_i(x) \iff \exists u^i : P.ix + tp_i + Qi.u^i - r_i \in K_i, i \leq N$$

$t \geq G(x) \iff \exists \tau = [\tau_1; \ldots; \tau_m], v^i : P\tau + Qu - r \in K, P_ix + \tau_ip_i + Qi.u^i - r_i \in K_i, i \leq N, \tau_i = f_i(x), i \leq k$

Refinement I. Let $f_1, \ldots, f_k$ be affine. Then the conclusion of Superposition Theorem remains true when $F$ is nondecreasing in arguments $y_{k+1}, \ldots, y_m$, CQr of $G$ being

$$t \geq G(x) \iff \exists u, \tau = [\tau_1; \ldots; \tau_m], v^i : P\tau + Qu - r \in K, P_ix + \tau_ip_i +Qi.u^i - r_i \in K_i, i \leq N, \tau_i = f_i(x), i \leq k$$

Illustration: The functions $F(y) = y^2$ and $f(x) = x^2 - 1$ are CQr; however, $F(f(x)) = (x^2 - 1)^2$ is nonconvex and thus is not CQr. In contrast, square of affine function is CQr.
Refinement II: Let $F(y): \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ and $f_i(x): \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, $i = 1, \ldots, m$, be CQr, with $f_1, \ldots, f_k$ affine. Assume that for some CQr set $Y \subseteq \mathbb{R}^m$ $F$ is nondecreasing on $Y$:

$$\forall (y' \in Y, y \in Y, y' \geq y) : F(y') \geq F(y)$$

and that for every $x$ such that $f_i(x) < +\infty$, $i \leq m$, it holds $f(x) := [f_1(x); \ldots; f_m(x)] \in Y$. Then the superposition

$$G(x) = \begin{cases} F(f_1(x), \ldots, f_m(x)), & f_i(x) < +\infty, i \leq m \\ +\infty, & \text{otherwise} \end{cases}$$

is CQr:

$$t \geq F(y) \iff \exists u : Py + tp + Qu - r \in K$$
$$t \geq f_i(x) \iff \exists u^i : P_i x + tp_i + Q_i u^i - r_i \in K_i, i \leq N$$
$$Y = \{y : \exists w : Ry + Sw - s \in K_Y\}, f(x) \in \mathbb{R}^m \Rightarrow f(x) \in Y$$

$$t \geq G(x) \iff \exists u, \tau = [\tau_1; \ldots; \tau_m], v^i, w : \begin{cases} P\tau + tp + Qu - r \in K [\Rightarrow F(\tau) \leq t] \\ P_i x + \tau_i p_i + Q_i u^i - r_i \in K_i, i \leq N [\Rightarrow \tau_i \geq f_i(x) \forall i] \\ R\tau + Sw - s \in K_Y, f_i(x) = \tau_i, i \leq k [\Rightarrow \tau \in Y \& f_i(x) = \tau_i, i \leq k] \end{cases}$$

Illustration: The functions $F(y) = y^2$ and $f(x) = x^2$ are CQr, and $F$ is nondecreasing on the CQr set $Y = \mathbb{R}_+$ where $f$ takes its values $\Rightarrow F'(f(x)) = x^4$ is CQr.

2.32
F.G. Projective transformation. Let $f(x) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function. It is known that then the projective transformation

$$F(x, \alpha) = \begin{cases} \alpha f(x/\alpha), & \alpha > 0 \\ +\infty, & \text{otherwise} \end{cases}$$

is convex as well. When $f$ is CQr, so is its projective transformation:

$$t \geq f(x) \iff \exists u : Px + tp + Qu - r \in K$$

$$\Downarrow$$

$$t \geq F(x, \alpha) \iff \exists u, s : \begin{cases} Px + tp + Qu - \alpha r \in K \ [\text{when } \alpha > 0, \text{ says that } t/\alpha \geq f(x/\alpha)] \\ [2; \alpha - s; \alpha + s] \in L^3 \ [\text{enforces } \alpha > 0] \end{cases}$$
Several more advanced convexity-preserving operations "behave well" on CQr functions under mild regularity assumptions:

**F.H*. Partial minimization.** Let $f(x, y) : \mathbb{R}^{nx} \times \mathbb{R}^{ny} \to \mathbb{R} \cup \{+\infty\}$ be CQr, $X \in \mathbb{R}^{nx}$ be a CQr set, and let parametric problem

$$
\min_y f(x, y)
$$

with $x \in X$ be solvable whenever it is feasible. Then the function

$$
g(x) = \begin{cases} 
\min_y f(x, y), & x \in X \\
+\infty, & x \notin X
\end{cases}
$$

is CQr:

$$
\begin{align*}
& t \geq f(x, y) \iff \exists u : P_f[x; y] + tp_f + Q_f u - r_f \in K_f \\
& X = \{x : \exists v : P_X x + Q_X v - r_X \in K_X\} \; \& \; \text{min}_y f(x, y) \text{ is achieved whenever it is } < +\infty
\end{align*}
$$

$$
t \geq g(x) \iff \exists y, u, v : P_f[x; y] + tp_f + Q_f u - r_f \in K_f, P_X x + Q_X v - r_X \in K_X
$$

says that $t \geq f(x, y)$ says that $x \in X$
**F.I.** Taking Legendre transformation: If \( f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \) is CQr with a strictly feasible CQR

\[
\{(t,x) : t \geq f(x)\} = \{(t,x) : \exists u : Px + tp + Qu - r \in K\}
\]

then the Legendre transformation of \( f \)

\[
f^*(\xi) = \sup_x \left[ \xi^T x - f(x) \right]
\]

is CQr:

\[
\{[\xi,\tau] : \tau \geq f^*(\xi)\} = \{[\xi,\tau] : \tau \geq \xi^T x - t \forall (t,x) \in \text{Epi}\{f\}\}
\]

\[
= \{[\xi,\tau] : \tau \geq \sup_{(t,x) \in \text{Epi}\{f\}} \left[ \xi^T x - t \right]\}
\]

\[
= \left\{[\xi,\tau] : \tau \geq \sup_{x,t,u} \{\xi^T x - t : Px + tp + Qu - r \in K\}\right\}
\]

\[
= \left\{[\xi,\tau] : \tau \geq \min_y \left\{-r^T y : P^T y + \xi = 0, Q^T y = 0, p^T y = 1, y \in K^* \right\} \right\} \quad (a)
\]

\[
= \left\{[\xi,\tau] : \exists y : p^T y = 1, P^T y + \xi = 0, \tau + r^T y \geq 0, y \in K^* \right\} \quad (b)
\]

where (a), (b) are due to Strong Duality.

2.35
More examples of CQr functions/sets

7. Convex quadratic form \( f(x) = x^T Q^T Q x + q^T x + r \) is CQr, since it can be obtained from the squared Euclidean norm and affine function (both are CQr) by affine substitution of argument and addition. Here is an explicit CQR for \( f \):

\[
\text{Epi}\{f\} = \{(x, t) : \begin{bmatrix} 2Qx \\ t - q^T x - r - 1 \\ t - q^T x - r + 1 \end{bmatrix} \geq_{L^{m+2}} 0\}\quad [Q : m \times n]
\]

Observation: Let \( m \) be nonnegative integer, and let \( M = 2^m \). The set

\[
\mathcal{X}_m = \{(t, x_1, x_2, ..., x_M) \in \mathbb{R}_+^{M+1} : t^M \leq x_1...x_M\}
\]

is CQr. Indeed,

\[
\mathcal{X}_m = \left\{(t, x_1, ..., x_M) \geq 0 : \exists y_{ij} \geq 0 : \begin{cases} y_{1,1}^2 \leq x_1 x_2, y_{1,2}^2 \leq x_3 x_4, ..., y_{1,M/2}^2 \leq x_{M-1} x_M \\ y_{2,1}^2 \leq y_{1,1} y_{1,2}, ..., y_{2,M/4}^2 \leq y_{1,M/2-1} y_{1,M/2} \\ \vdots \\ t^2 \leq y_{m-1,1} y_{m-1,2} \end{cases} \right\}
\]

Observation implies CQr’s of power functions.
8.1. Convex increasing power function \( f(x) = (x)^{\pi_+}, \ a_+ = \max[x,0], \) with rational degree \( \pi = \frac{p}{q} \geq 1 \) is CQr.

Indeed, let \( \mu \in \mathbb{N} \) be such that \( M \equiv 2^\mu \geq p + q \). We have

\[
Y \equiv \{(\tau, x_1, \ldots, x_M) \geq 0 : \tau^M \leq x_1 \ldots x_M\} \text{ is CQr}
\]
\[
\Rightarrow \{(t, \xi) \geq 0 : \xi^M \leq t^q \xi^{M-p} 1^{p-q}\} = \{(t, \xi) : \mathcal{A}(t, \xi) \in Y\} \text{ is CQr}
\]

[affine substitution of variables \( \mathcal{A}(t, \xi) = (\xi, t^{\frac{M}{M-p}}, t^{\frac{M}{p-q}}, 1, \ldots, 1) \)]

\[
\Rightarrow \{(t, \xi) \geq 0 : t \geq \xi^\frac{p}{q}\} \text{ is CQr}
\]
\[
\Rightarrow \text{Epi}\{f\} = \{(t, x) : \exists \xi : (t, \xi) \geq 0, t \geq \xi^\frac{p}{q}, \xi \geq x\} \text{ is CQr}
\]

Illustration:

\[
t \geq (x)^{7/3}_+ \iff \exists (z : z \geq 0, z \geq x) : t \geq z^{7/3}
\]
\[
\iff \exists (z : z \geq 0, z \geq x) : t \geq 0, z^{16} \leq t^3 z^9 1^4 = t \cdot t \cdot t \cdot z \cdot z \cdot z \cdot z \cdot z \cdot z \cdot 1 \cdot 1 \cdot 1 \cdot 1
\]
\[
\iff \exists (z, u_i : z \geq 0, z \geq x, u_i \geq 0) : \left\{ \begin{align*}
&u_1^2 \leq t^2, u_2^2 \leq t z, u_3^2 \leq z^2, u_4^2 \leq z^2, u_5^2 \leq z^2, u_6^2 \leq z^2, u_7^2 \leq 1, u_8^2 \leq 1 \\
z^8 \leq u_1 u_2 u_3 u_4 u_5 u_6 u_7 u_8, t \geq 0
\end{align*} \right\
\]
\[
\iff \exists (z, u_i, v_i : z \geq 0, z \geq x, u_i \geq 0, v_i \geq 0) : \left\{ \begin{align*}
&u_1^2 \leq t^2, u_2^2 \leq t x, u_3^2 \leq x^2, u_4^2 \leq x^2, u_5^2 \leq x^2, u_6^2 \leq x^2, u_7^2 \leq 1, u_8^2 \leq 1 \\\nu_1^2 \leq u_1 u_2, u_2^2 \leq u_3 u_4, u_3^2 \leq u_5 u_6, u_4^2 \leq u_7 u_8 \\
z^4 \leq v_1 v_2 v_3 v_4, t \geq 0
\end{align*} \right\
\]
\[
\iff \exists \left( z, u_i, v_i, w_i : \begin{array}{l} z \geq 0, z \geq x, \\
u_i \geq 0, v_i \geq 0, \\
w_i \geq 0 \end{array} \right) : t \geq 0, \left\{ \begin{align*}
&u_1^2 \leq t^2, u_2^2 \leq t x, u_3^2 \leq x^2, u_4^2 \leq x^2, u_5^2 \leq x^2, u_6^2 \leq x^2, u_7^2 \leq 1, u_8^2 \leq 1 \\
v_1^2 \leq u_1 u_2, v_2^2 \leq u_3 u_4, v_3^2 \leq u_5 u_6, v_4^2 \leq u_7 u_8 \\
&v_1^2 \leq w_1 w_2, v_2^2 \leq v_3 v_4 \\
z^2 \leq w_1 w_2, t \geq 0
\end{align*} \right\
\]
8.2. **Convex piecewise power function**

\[ f(x) = \begin{cases} 
  x^{\pi^+}, & x \geq 0 \\
  |x|^{\pi^-}, & x \leq 0 
\end{cases} \]

with rational degrees \( \pi_{\pm} \geq 1 \) is CQr.

Indeed, the function is obtained from CQR function \((x)^{\pi_+}\) by summation and affine substitution of variables:

\[ f(x) = (x^+)^{\pi^+} + (-x)^{\pi^-} \]

8.3. **Decreasing power function**

\[ f(x) = \begin{cases} 
  x^{-\pi}, & x > 0 \\
  +\infty, & x \leq 0 
\end{cases} \]

of rational degree \(-\pi < 0\) is CQr.

Indeed, when \( \pi = p/q \) with positive integers \( p, q \) and \( \mu \in \mathbb{N} \) is such that \( M = 2^\mu \geq p + q \) we have

\[ \text{Epi}\{f\} = \{(t, x) : t \geq 0, x \geq 0, x^{pt^q} \geq 1\} = \{(t, x) : 1 \leq x^{pt^q}1^{M-p-q}\} \]

which is the inverse affine image of the CQr set

\[ \{(\tau, x_1, \ldots, x_m) \geq 0 : \tau^M \leq x_1 \cdots x_M\} \]

under the affine mapping \((t, x) \mapsto (1, \underbrace{x, \ldots, x}_p, t, \ldots, t, \underbrace{1, \ldots, 1}_{M-p-q})\)
8.4. The hypograph of a concave power monomial. When $\pi_i > 0$ are rational and $\sum_i \pi_i \leq 1$, the convex monomial

$$f(x) = \begin{cases} -x_1^{\pi_1} \cdots x_m^{\pi_m}, & x \geq 0 \\ +\infty, & \text{otherwise} \end{cases}$$

is CQr.

Indeed, let $\pi_i = p_i/q$ with positive integers $p_i$ and positive integer $q$, and let $\mu \in \mathbb{N}$ be such that $M = 2^\mu \geq q$. Then

$$\text{Epi}\{f\} = \{(x, t) : \exists \tau : \tau \geq 0, t + \tau \geq 0, (\tau, x) \in \mathcal{M}\},$$

$$\mathcal{M} = \{ (\tau, x_1, \ldots, x_m) \geq 0 : \tau^q \leq x_1^{p_1} x_2^{p_2} \cdots x_m^{p_m} \}$$

$$= \{ (\tau, x_1, \ldots, x_m) \geq 0 : \tau^M \leq x_1^{p_1} x_2^{p_2} \cdots x_m^{p_m} \tau^{M-q} 1^q - \sum_i p_i \}$$

that is, $\text{Epi}\{f\}$ is the intersection of a polyhedral set and the inverse image of the CQr set

$$\{(s, y_1, \ldots, y_M) \geq 0 : s^M \leq y_1 \cdots y_M \}$$

under the affine mapping

$$\left( \tau, x_1, \ldots, x_m \right) \mapsto \left( \tau, \underbrace{x_1, \ldots, x_1}_{p_1}, \ldots, \underbrace{x_m, \ldots, x_m}_{p_m}, \tau, \ldots, \tau, \underbrace{1, \ldots, 1}_{M-q \underbrace{q - \sum_i p_i}} \right).$$
8.5. The epigraph of a convex power monomial. When \( \pi_i > 0 \) are rational, the function

\[
f(x) = \begin{cases} 
    x_1^{-\pi_1} \cdots x_m^{-\pi_m}, & x > 0 \\
    +\infty, & \text{otherwise}
\end{cases}
\]

is CQr.

Indeed, when \( p_1, \ldots, p_m, q \) are positive integers such that \( \pi_i = p_i/q \) and \( \mu \in \mathbb{N} \) is such that \( M = 2^\mu \geq p_1 + \ldots + p_m + q \), we have

\[
\text{Epi}\{f\} = \{(t, x_1, \ldots, x_m) \geq 0 : t^q x_1^{p_1} \cdots x_m^{p_m} \geq 1\},
\]

that is, \( \text{Epi}\{f\} \) is the intersection of a polyhedral set and the inverse image of the CQr set

\[
\{(s, y_1, \ldots, y_M) \geq 0 : s^M \leq y_1 \cdots y_M\}
\]

under the affine mapping

\[
(t, x_1, \ldots, x_m) \mapsto (1, t, \ldots, t, x_1, \ldots, x_1, \ldots, x_m, \ldots, x_m, 1, \ldots, 1)\].
\]
8.6. The epigraph of the $\| \cdot \|_\pi$-norm. When $\pi \geq 1$ is rational (or $\pi = \infty$), the function $f(x) = \|x\|_\pi : \mathbb{R}^m \to \mathbb{R}$ is CQr. Indeed, the case of $\pi = \infty$ is trivial — in this case $\text{Epi}\{f\}$ is a polyhedral set. Now let $\pi = p/q$ with positive integer $p \geq q$. It is immediately seen that

$$\|x\|_p \leq t \iff t \geq 0 \& \exists v_1, \ldots, v_m \geq 0 : |x_i| \leq t^{(\pi - 1)/\pi} v_i^{1/\pi}, \ i = 1, \ldots, m, \sum_{i=1}^n v_i \leq t. \quad (*)$$

As we have seen in 8.5, the set $Z = \{(\tau, \xi, \sigma) : \tau \geq 0, \sigma \geq 0, \xi \leq \frac{p-q}{p} \frac{q}{\sigma p}\}$ is CQr. Consequently, so are the sets

$$X_i = \{(x, v, t) \in \mathbb{R}^{2m+1} : t \geq 0, v \geq 0, |x_i| \leq t^{(\pi - 1)/\pi} v_i^{1/\pi}\} = \{(x, v, t) \in \mathbb{R}^{2m+1} : t \geq 0, v \geq 0 \pm x_i \leq t^{p-q/p} v_i^{q/p}\}$$

— each of these sets is the intersection of two inverse affine images of $Z$ under affine mappings. By $(*)$, $\text{Epi}\{f\}$ is the image, under the linear mapping $(x, t, v) \mapsto (x, t)$, of the CQr set

$$\{(x, t, v) : \sum_i v_i \leq t\} \cap \bigcap_i X_i,$$

so that $\text{Epi}\{f\}$ is a CQr set, $\Rightarrow f$ is CQr.
Consider an LP program

\[ \min_x \{ c^T x : Ax + b \geq 0 \} \quad \text{(LP)} \]

In applications, the data \((c, A, b)\) of the program not always are known exactly. In LP practice, however, “small” data uncertainties (like 0.1% or less) are usually ignored, and the problem is processed as if the data were exact.

(!) \textit{It turns out that ignoring small data uncertainties can make the optimal solution meaningless.}
Example 1: Synthesis of Antenna array

♣ The diagram of an antenna. Consider a (monochromatic) antenna placed at the origin. The electric field generated by the antenna at a remote point $r\delta$ ($\delta$ is a unit direction) is

$$E = a(\delta)r^{-1}\cos(\phi(\delta) + tw - 2\pi r/\lambda) + o(r^{-1})$$

- $t$: time
- $\omega$: frequency
- $\lambda$: wavelength

It is convenient to aggregate $a(\delta)$ and $\phi(\delta)$ into a single complex-valued function — the diagram of the antenna

$$D(\delta) = a(\delta)(\cos(\phi(\delta)) + i\sin(\phi(\delta))).$$

- The directional density of the energy sent by the antenna is proportional to $|D(\cdot)|^2$

- The diagram $D(\cdot)$ of a complex antenna comprised of several antenna elements is the sum of the diagrams $D_i(\cdot)$ of the elements:

$$D(\delta) = D_1(\delta) + \ldots + D_N(\delta)$$
Synthesis of Array of Antennae: Given a target diagram $D_*(\cdot)$ along with $N$ “building blocks” – antenna elements with diagrams $D_1(\cdot), \ldots, D_N(\cdot)$ – find “weights” $z_j \in \mathbb{C}$ such that the function

$$\sum_{j=1}^{N} z_j D_j(\cdot)$$

is as close as possible to the target diagram $D_*(\cdot)$.

- Physically, multiplication of a diagram $D_j(\cdot)$ by a complex weight $z_j$ means that the corresponding standard “building block” is preceded by appropriate amplification and delay devices.

- Choosing a fine grid $\Delta$ of directions $\delta$, we may pose the Antenna Synthesis problem as a discrete approximation problem with complex-valued data and design variables:

$$\min_{\tau, x} \left\{ \tau : \left| D_*(\delta) - \sum_{j=1}^{N} z_j D_j(\delta) \right| \leq \tau \ \forall \delta \in \Delta \right\},$$

which is a CQP.

- Sometimes the diagrams of the elements and the target diagram are real-valued. In this case, we lose nothing when restricting $z_j$ to be real, and thus end up with an LP program.
Antenna synthesis: Example

Let a planar antenna be comprised of a central circle and 9 concentric rings of the same area placed in the $XY$-plane ("Earth’s surface"):

The radius of the antenna is 1m
The diagram of a ring \( \{a \leq r \leq b\} \) in the \( XY \)-plane is real-valued and depends on direction’s altitude \( \theta \) only:

\[
D_{a,b}(\theta) = \frac{1}{2} \int_a^b \left[ \int_0^{2\pi} \rho \cos(2\pi \rho \lambda^{-1} \cos(\theta) \cos(\phi)) d\phi \right] d\rho :
\]

Diagrams of 10 rings as functions of altitude \( \theta \in [0, \pi/2] \), \( \lambda = 0.5 \text{m} \)
• Assume the target diagram to be real-valued function of the altitude “concentrated” in the angle $\frac{\pi}{2} - \frac{\pi}{12} \leq \theta \leq \frac{\pi}{2}$:

![The target diagram]

• With 120-point discretization of altitudes, the Antenna Synthesis problem becomes an LP program with 11 variables and 240 linear constraints:

$$\min_{x, \tau} \left\{ \tau : -\tau \leq D_\ast(\theta_\ell) - \sum_{j=1}^{10} x_j D_j(\theta_\ell) \leq \tau, \; \theta_\ell = \frac{\pi}{2\ell}, \; 1 \leq \ell \leq 120 \right\}$$
• The resulting diagram approximates the target within absolute inaccuracy 0.0621:

The target diagram (dashed) and the synthesized diagram (solid)

• The optimal weights (rounded to 5 digits) are

<table>
<thead>
<tr>
<th>element #</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>weight</td>
<td>1624.4</td>
<td>-14700</td>
<td>55383</td>
<td>-107247</td>
<td>95468</td>
<td>19221</td>
<td>-138620</td>
<td>144870</td>
<td>-69303</td>
<td>13311</td>
</tr>
</tbody>
</table>

2.48
The optimal weights $x^*_j$, $j = 1, \ldots, 10$, are characteristics of physical devices. In reality, they somehow drift around their computed values. What happens when the weights are affected by small (just 0.1%) random perturbations:

$$x_j = (1 + \epsilon_j)x^*_j$$

$$\left[\{\epsilon_j \sim \text{Uniform}\left[-0.001, 0.001\right]\}_{j=1}^{10}\right]$$

The results of 0.1% “implementation errors” are disastrous:

“Dream and reality”

“Nominal” diagram   Actual diagram
[ dashed: the target diagram ]

- The target diagram is of the uniform norm 1, and its uniform distance from the nominal diagram is $\approx 0.06$.
- The realization of “actual diagram” shown on the picture is at the uniform distance 7.8 from the target diagram!
Example 2: NETLIB Case Study: Diagnosis

♣ NETLIB is a collection of about 100 not very large LPs, mostly of real-world origin. To motivate the methodology of our “case study”, here is constraint # 372 of the NETLIB problem PILOT4:

\[
\begin{align*}
    a^T x & \equiv -15.79081 x_{826} - 8.598819 x_{827} - 1.88789 x_{828} - 1.362417 x_{829} - 1.526049 x_{830} \\
    & \quad -0.031883 x_{849} - 28.725555 x_{850} - 10.792065 x_{851} - 0.19004 x_{852} - 2.757176 x_{853} \\
    & \quad -12.290832 x_{854} + 717.562256 x_{855} - 0.057865 x_{856} - 3.785417 x_{857} - 78.30661 x_{858} \\
    & \quad -122.163055 x_{859} - 6.46609 x_{860} - 0.48371 x_{861} - 0.615264 x_{862} - 1.353783 x_{863} \\
    & \quad -84.644257 x_{864} - 122.459045 x_{865} - 43.15593 x_{866} - 1.712592 x_{870} - 0.401597 x_{871} \\
    & \quad + x_{880} - 0.946049 x_{898} - 0.946049 x_{916} \\
    \geq & \quad b \equiv 23.387405
\end{align*}
\]

(C)

The related nonzero coordinates in the optimal solution \(x^*\) of the problem, as reported by CPLEX, are:

\[
\begin{align*}
    x_{826}^* &= 255.6112787181108 & x_{827}^* &= 6240.488912232100 & x_{828}^* &= 3624.613324098961 \\
    x_{829}^* &= 18.20205065283259 & x_{849}^* &= 174397.0389573037 & x_{870}^* &= 14250.00176680900 \\
    x_{871}^* &= 25910.00731692178 & x_{880}^* &= 104958.3199274139
\end{align*}
\]

This solution makes (C) an equality within machine precision.

♣ Most of the coefficients in (C) are “ugly reals” like -15.79081 or -84.644257. We definitely may believe that these coefficients characterize technological devices/processes, and as such hardly are known to high accuracy. Thus, “ugly coefficients” may be assumed to be uncertain and to coincide with the “true” data within accuracy of 3-4 digits. The only exception is the coefficient 1 of \(x_{880}\), which perhaps reflects the structure of the problem and is exact.
\[ a^T x \equiv -15.79081 x_{826} - 8.598819 x_{827} - 1.88789 x_{828} - 1.362417 x_{829} - 1.526049 x_{830} \\
-0.031883 x_{849} - 28.725555 x_{850} - 10.792065 x_{851} - 0.19004 x_{852} - 2.757176 x_{853} \\
-12.290832 x_{854} + 717.562256 x_{855} - 0.057865 x_{856} - 3.785417 x_{857} - 78.30661 x_{858} \\
-122.163055 x_{859} - 6.46609 x_{860} - 0.48371 x_{861} - 0.615264 x_{862} - 1.353783 x_{863} \\
-84.644257 x_{864} - 122.459045 x_{865} - 43.15593 x_{866} - 1.712592 x_{870} - 0.401597 x_{871} \\
+ x_{880} - 0.946049 x_{898} - 0.946049 x_{916} \geq b \equiv 23.387405 \]

\[ x^*_{826} = 255.6112787181108 \quad x^*_{827} = 6240.488912232100 \quad x^*_{828} = 3624.613324098961 \]
\[ x^*_{829} = 18.20205065283259 \quad x^*_{849} = 174397.0389573037 \quad x^*_{870} = 14250.00176680900 \]
\[ x^*_{871} = 25910.00731692178 \quad x^*_{880} = 104958.3199274139 \]

Assume that the uncertain entries of \( a \) are 0.1%-accurate approximations of unknown entries in the “true” data \( \tilde{a} \), how would this uncertainty affect the validity of the constraint evaluated at the nominal solution \( x^* \)?

- The worst case, over all 0.1%-perturbations of uncertain data, violation of the constraint is as large as 450% of the right hand side!
- In the case of random and independent 0.1% perturbations of the uncertain coefficients, the statistics of the “relative constraint violation”

\[
V = \frac{\max[b - \tilde{a}^T x^*, 0]}{b} \times 100%
\]

also is disastrous:

<table>
<thead>
<tr>
<th>Prob{V &gt; 0}</th>
<th>Prob{V &gt; 150%}</th>
<th>Mean(V)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50</td>
<td>0.18</td>
<td>125%</td>
</tr>
</tbody>
</table>

Relative violation of constraint # 372 in PILOT4

(1,000-element sample of 0.1% perturbations of the uncertain data)
We see that quite small (just 0.1%) perturbations of “obviously uncertain” data coefficients can make the “nominal” optimal solution $x^*$ heavily infeasible and thus – practically meaningless.
In our Case Study, we choose a “perturbation level” $\epsilon$ (taking values 1%, 0.1%, 0.01%), and, for every one of the NETLIB problems, measure the “reliability index” of the nominal solution at this perturbation level, specifically, as follows.

- We compute the optimal solution $x^*$ of the program by CPLEX.
- For every one of the inequality constraints $a^T x \leq b$ --- we split the right hand side coefficients $a_j$ into “certain” (rational fractions $p/q$ with $|q| \leq 100$) and “uncertain” (all the rest). Let $J$ be the set of all uncertain coefficients of (*)
--- we define the reliability index of (*) as
\[
\frac{a^T x^* + \epsilon \sqrt{\sum_{j \in J} a_j^2 (x_j^*)^2 - b}}{\max[1,|b|]} \times 100%
\]

Note that the reliability index is of order of typical violation (measured in percents of the right hand side) of the constraint, as evaluated at $x^*$, under independent random perturbations, of relative magnitude $\epsilon$, of the uncertain coefficients.

- We treat the nominal solution as unreliable, and the problem - as bad, the level of perturbations being $\epsilon$, if the worst, over the inequality constraints, reliability index is worse than 5%.
The results of the Diagnosis phase of our Case Study are as follows.

From the total of 90 NETLIB problems we have processed,
• in 27 problems the nominal solution turned out to be unreliable at the largest ($\epsilon = 1\%$) level of uncertainty;
• 19 of these 27 problems are already bad at the 0.01%-level of uncertainty, and in 13 of these 19 problems, 0.01% perturbations of the uncertain data can make the nominal solution more than 50%-infeasible for some of the constraints.
<table>
<thead>
<tr>
<th>Problem</th>
<th>Size(^a))</th>
<th>(\epsilon = 0.01%)</th>
<th>(\epsilon = 0.1%)</th>
<th>(\epsilon = 1%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#bad(^b))</td>
<td>Index(^c))</td>
<td>#bad</td>
<td>Index</td>
</tr>
<tr>
<td>80BAU3B</td>
<td>2263 x 9799</td>
<td>37</td>
<td>84</td>
<td>177</td>
</tr>
<tr>
<td>25FV47</td>
<td>822 x 1571</td>
<td>14</td>
<td>16</td>
<td>28</td>
</tr>
<tr>
<td>ADLITTLE</td>
<td>57 x 97</td>
<td>2</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>APIRO</td>
<td>28 x 32</td>
<td>1</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>BNL2</td>
<td>2325 x 3489</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRANDY</td>
<td>221 x 249</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAPRI</td>
<td>272 x 353</td>
<td>10</td>
<td>39</td>
<td>14</td>
</tr>
<tr>
<td>CYCLE</td>
<td>1904 x 2857</td>
<td>2</td>
<td>110</td>
<td>5</td>
</tr>
<tr>
<td>D2Q06C</td>
<td>2172 x 5167</td>
<td>107</td>
<td>1,150</td>
<td>134</td>
</tr>
<tr>
<td>E226</td>
<td>224 x 282</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FFFFF800</td>
<td>525 x 854</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FINNIS</td>
<td>498 x 614</td>
<td>12</td>
<td>10</td>
<td>63</td>
</tr>
<tr>
<td>GREENBEA</td>
<td>2393 x 5405</td>
<td>13</td>
<td>116</td>
<td>30</td>
</tr>
<tr>
<td>KB2</td>
<td>44 x 41</td>
<td>5</td>
<td>27</td>
<td>6</td>
</tr>
<tr>
<td>MAROS</td>
<td>847 x 1443</td>
<td>3</td>
<td>6</td>
<td>38</td>
</tr>
<tr>
<td>NESM</td>
<td>751 x 2923</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PEROLD</td>
<td>626 x 1376</td>
<td>6</td>
<td>34</td>
<td>26</td>
</tr>
<tr>
<td>PILOT</td>
<td>1442 x 3652</td>
<td>16</td>
<td>50</td>
<td>185</td>
</tr>
<tr>
<td>PILOT4</td>
<td>411 x 1000</td>
<td>42</td>
<td>210,000</td>
<td>63</td>
</tr>
<tr>
<td>PILOT87</td>
<td>2031 x 4883</td>
<td>86</td>
<td>130</td>
<td>433</td>
</tr>
<tr>
<td>PILOTJA</td>
<td>941 x 1988</td>
<td>4</td>
<td>46</td>
<td>20</td>
</tr>
<tr>
<td>PILOTNOV</td>
<td>976 x 2172</td>
<td>4</td>
<td>69</td>
<td>13</td>
</tr>
<tr>
<td>PILOTWE</td>
<td>723 x 2789</td>
<td>61</td>
<td>12,200</td>
<td>69</td>
</tr>
<tr>
<td>SCFXM1</td>
<td>331 x 457</td>
<td>1</td>
<td>95</td>
<td>3</td>
</tr>
<tr>
<td>SCFXM2</td>
<td>661 x 914</td>
<td>2</td>
<td>95</td>
<td>6</td>
</tr>
<tr>
<td>SCFXM3</td>
<td>991 x 1371</td>
<td>3</td>
<td>95</td>
<td>9</td>
</tr>
<tr>
<td>SHARE1B</td>
<td>118 x 225</td>
<td>1</td>
<td>257</td>
<td>1</td>
</tr>
</tbody>
</table>

\(^a\) \# of linear constraints (excluding the box ones) plus 1 and \# of variables
\(^b\) \# of constraints with index > 5%
\(^c\) The worst, over the constraints, reliability index, in %
Conclusions:

- In real-world applications of Linear Programming one cannot ignore the possibility that a small uncertainty in the data (intrinsic for the majority of real-world LP programs) can make the usual optimal solution of the problem completely meaningless from a practical viewpoint.

Consequently,

- In applications of LP, there exists a real need of a technique capable of detecting cases when data uncertainty can heavily affect the quality of the nominal solution, and in these cases to generate a “reliable” solution, one which is immune against uncertainty.
Consider an LP program

\[
\min_x \left\{ c^T x : Ax + b \geq 0 \right\}
\]  

(LP)

Assume that the data \((c, A, b)\) of the program are not known exactly; all we know is an uncertainty set \(U\) the “true data” belong to.

A natural way to process an LP program with uncertain data is to build the \textit{robust counterpart} of the program, where we impose on candidate solutions the requirement to be \textit{robust feasible} – to satisfy \textit{all} realizations of the inequality constraints. Among these robust feasible solutions, we are seeking for the “best” – with the smallest possible \textit{guaranteed} value of the objective. Thus, the robust counterpart of (LP) is the problem

\[
f(x) = \min_x \left\{ \max_{c \in U_{\text{obj}}} c^T x : Ax + b \geq 0 \; \forall (A, b) \in U_{\text{cons}} \right\}
\]  

(RC)

where

\[
U_{\text{obj}} = \{ c : \exists (A, b) : (c, A, b) \in U \},

U_{\text{cons}} = \{ (A, b) : \exists c : (c, A, b) \in U \}
\]

are the projections of the uncertainty set on the spaces of the data of the objective and the constraints, respectively.
Robust counterpart is a *semi-infinite* convex optimization program – one with infinitely many linear inequality constraints. Possibilities to process such a problem depend on the geometry of the uncertainty set $\mathcal{U}$.

If the uncertainty set $\mathcal{U}$ is an ellipsoid (or an intersection of ellipsoids), (RC) can be converted to a conic quadratic program.
Uncertain LP with “simple” ellipsoidal uncertainty sets

\[
\min_x \left\{ c^T x : Ax + b \geq 0 \right\}, \quad (c, A, b) \in U; \quad A = \begin{bmatrix} a_1^T ; \ldots ; a_m^T \end{bmatrix} : m \times n
\]

\[
\downarrow
\]

\[
\min_{t, x} \left\{ t : c^T x \leq t, Ax + b \geq 0 \ \forall (c, A, b) \in U \right\}. \quad (RC)
\]

Assume that the projections \( U_{\text{obj}} \) and \( U_i \) of the uncertainty set on the space of the objective data and the data of \( i \)-th constraint, \( i = 1, \ldots, m \), are ellipsoids:

\[
U_{\text{obj}} = \left\{ c = c^0 + P_0 u : u \in \mathbb{R}^{k_0}, u^T u \leq 1 \right\};
\]

\[
U_i = \left\{ [a_i; b_i] = [a_i^0; b_i^0] + P_i u : u \in \mathbb{R}^{k_i}, u^T u \leq 1 \right\}
\]
\[
\min_{x} \{ c^T x : Ax + b \geq 0 \}, (c, A, b) \in \mathcal{U} \quad \text{(ULP)}
\]

\[
\Rightarrow \min_{t, x} \left\{ \begin{array}{l}
t : (1) \quad c^T x \leq t, \\
i = 1, \ldots, m \end{array} \right\} \quad \text{(RC)}
\]

\[
\mathcal{U}_i = \{ [a_i; b_i] = [a_i^0; b_i^0] + P_i u : u \in \mathbb{R}^{k_i}, u^T u \leq 1 \}
\]

- A candidate solution \((t, x)\) satisfies all realizations of \((2_i)\) iff

\[
[a_i^0]^T x + b_i^0 + [P_i u]^T [x; 1] \geq 0 \quad \forall u : u^T u \leq 1
\]

\[
\Leftrightarrow \min_{u : u^T u \leq 1} \left\{ [a_i^0]^T x + b_i^0 + [P_i u]^T [x; 1] \right\} \geq 0
\]

\[
\Leftrightarrow [a_i^0]^T x + b_i^0 - \|P_i^T [x; 1]\|_2 \geq 0
\]

\[
\Leftrightarrow \|P_i^T [x; 1]\|_2 \leq [a_i^0]^T x + b_i^0 \quad \text{c.q.i.}
\]

Similarly, \((t, x)\) satisfies all realizations of \((1)\) iff \(\|P_0^T x\|_2 \leq t - [c^0]^T x\).

- Thus, \((RC)\) is the conic quadratic program

\[
\min_{t, x} \left\{ t : \|P_0^T x\|_2 \leq t - [c^0]^T x, \|P_i^T [x; 1]\|_2 \leq [a_i^0]^T x + b_i^0, i \leq m \right\}
\]
Theorem. Consider an uncertain LP

\[
\begin{aligned}
&\min \left\{ c^T x : A x \geq b \right\} : (c, A, b) \in U \\
&\text{ (ULP)}
\end{aligned}
\]

and assume that the uncertainty set \( U \) is CQr with an essentially strictly feasible CQR. Then the set of robust feasible solutions to (ULP) is CQr with explicitly given CQR, so that the Robust Counterpart of (ULP) is (equivalent to) an explicit conic quadratic problem.

If \( U \) is LP-representable:

\[
U = \{ \zeta = (c, A, b) : \exists u : P\zeta + Qu + r \geq 0 \},
\]

then the RC of (ULP) is (equivalent to) an explicit LP problem.

\[\star\] Example: The Robust Counterpart of uncertain LP with interval uncertainty:

\[
U_{\text{obj}} = \{ c : |c_j - c^0_j| \leq \delta c_j, j = 1, \ldots, n \}
\]

\[
U_i = \{ (a_{i1}, \ldots, a_{im}, b_i) : |a_{ij} - a^0_{ij}| \leq \delta a_{ij}, |b_i - b^0_i| \leq \delta b_i \}
\]

is the LP program

\[
\begin{align*}
\min_{x,y,t} \quad & \delta c_j y_j \leq t \\
\text{subject to} \quad & \sum_j c^0_j x_j + \sum_j \delta c_j y_j \leq t \\
& t : \sum_j a^0_{ij} x_j + \sum_j \delta a_{ij} y_j \leq b_i - \delta b^0_i \\
& -y_j \leq x_j \leq y_j
\end{align*}
\]
Theorem is an immediate consequence of the following Observation: Let $\mathcal{Z} \subset \mathbb{R}^{n+1}$ be a nonempty CQr set given by essentially strictly feasible CQR:

$$\mathcal{Z} = \left\{ z \in \mathbb{R}^n : \exists u : \begin{cases} Pz + Qu - r \in \mathcal{K} \\ Rx + Su - s = 0 \end{cases} \right\}$$

$$\exists (\bar{z}, \bar{u}) : P\bar{z} + Q\bar{u} - r \in \text{int} \mathcal{K}, R\bar{z} + S\bar{u} = s$$

($\mathcal{K}$: direct product of Lorentz cones). Then the set

$$\mathcal{X} = \{ x : z^T[x; 1] \leq 0 \forall z \in \mathcal{Z} \}$$

of robust feasible solutions to the uncertain linear constraint $z^T[x; 1] \leq 0$, the uncertain data running through $\mathcal{Z}$, is CQr with explicitly given CQR. Indeed,

$$x \in \mathcal{X} \iff \sup_{z \in \mathcal{Z}} [x; 1]^Tz \leq 0 \iff 0 \geq \sup_{z,u} \{ [x; 1]^Tz : Pz + Qu - r \in \mathcal{K}, Rz + Su = s \}$$

$$\iff 0 \geq \min_{y,v} \{-r^Ty - s^Tv : y \in \mathcal{K}^* [= \mathcal{K}], P^Ty + R^Tv + [x; 1] = 0, Q^Ty + S^Tv = 0 \}$$

$$\iff \exists y, v : y \in \mathcal{K}^* [= \mathcal{K}], P^Ty + R^Tv + [x; 1] = 0, Q^Ty + S^Tv = 0, r^Ty + s^Tv \geq 0$$

with (a), (b) given by Strong Duality.

2.62
How it works? – Antenna Example

\[
\min_{x, \tau} \left\{ \tau : -\tau \leq D_*(\theta) - \sum_{j=1}^{10} x_j D_j(\theta) \leq \tau, \; \ell = 1, \ldots, L \right\}
\]

\[\Leftrightarrow \min_{x, \tau} \left\{ \tau : Ax + \tau a + b \geq 0 \right\} \quad \text{(LP)}\]

- The influence of “implementation errors”

\[x_j \mapsto (1 + \epsilon_j)x_j, \; |\epsilon_j| \leq \rho,\]

is as if there were no implementation errors, but the part \(A\) of the constraint matrix were uncertain and known “up to multiplication by a diagonal matrix with diagonal entries from \([1 - \rho, 1 + \rho]\)”: \(U_{\text{ini}} = \{A = A^{\text{nom}}\text{Diag}\{1 + \epsilon_1, \ldots, 1 + \epsilon_{10}\} : |\epsilon_j| \leq \rho\}\) (U)

Note that as far as a particular constraint is concerned, the uncertainty is an interval one with \(\delta A_{ij} = \rho|A_{ij}|\). The remaining coefficients (and the objective) are certain.

♣ To improve reliability of our design, we could replace the uncertain LP program (LP), (U) with its robust counterpart, which is nothing but an explicit LP program.

2.63
However, to work with interval uncertainty set $U_{ini}$ would be “too conservative” — the implementation errors are random and independent $\Rightarrow$ the probability for all of them to take simultaneously the “most unfavourable” values is negligibly small.

Let us try to define the uncertainty set in a smarter way.

Consider a linear constraint
\[
\sum_{j=1}^{n} a_j x_j + b \geq 0 \tag{L}
\]
and let $a_j$ be randomly perturbed: $a_j \mapsto (1 + \epsilon_j) a_j$ $\epsilon_j$ being independent symmetrically distributed and bounded random variables:
\[
\epsilon_j \sim -\epsilon_j \text{ and } |\epsilon_j| \leq \sigma_j.
\]

What is a “reliable version” of (L)?

**Note:** When assuming $a_j$ fixed and $x_j$ randomly perturbed: $x_j \mapsto (1 + \epsilon_j) x_j$, we are in exactly the same situation as when $a_j$ are randomly perturbed and $x_j$ are fixed!
\[ \sum_{j=1}^{n} a_j x_j + b \geq 0 \quad \text{(L)} \]

- With randomly perturbed \( a_j \), the left hand side in (L) becomes a random variable:

\[ \zeta = \sum_{j=1}^{n} a_j (1 + \epsilon_j) x_j + b \]

\[
\begin{align*}
\text{Mean}\{\zeta\} & \equiv \mathcal{E}\{\zeta\} = \sum_{j=1}^{n} a_j x_j + b, \\
\text{StD}\{\zeta\} & \equiv (\mathcal{E}\{(\zeta - \text{Mean}\{\zeta\})^2\})^{1/2} \leq \sqrt{\sum_{j=1}^{n} \sigma_j^2 a_j^2 x_j^2}.
\end{align*}
\]

- Let us choose a “safety parameter” \( \kappa \) and ignore all events where

\[ \zeta < \text{Mean}\{\zeta\} - \kappa \text{StD}\{\zeta\}, \]

taking full responsibility for all remaining events.

With this “common sense” approach, a “reliable” version of (L) becomes the conic quadratic inequality

\[ \sum_{j=1}^{n} a_j x_j + b - \kappa \sqrt{\sum_{j=1}^{n} \sigma_j^2 a_j^2 x_j^2} \geq 0 \quad \text{(L}_{\text{rel}}) \]
\[
\sum_{j=1}^{n} a_j (1 + \epsilon_j) x_j + b \geq 0 \quad \text{(L)}
\]

\[E\{\epsilon_j\} = 0; \quad |\epsilon_j| \leq \sigma_j\]

\[
\sum_{j=1}^{n} a_j x_j + b - \kappa \sqrt{\sum_{j=1}^{n} \sigma_j^2 a_j^2 x_j^2} \geq 0 \quad \text{(L}_{\text{rel}}\text{)}
\]

- Note: (L_{rel}) is exactly the robust counterpart of (L) associated with the ellipsoidal uncertainty set

\[
\mathcal{U}_\kappa = \left\{ a' = a + \kappa \text{Diag}(\sigma_1 a_1, \ldots, \sigma_n a_n) u : u^T u \leq 1 \right\} \quad \text{(Ell)}
\]

Thus, ignoring “rare events” is equivalent to replacing the actual box

\[
\mathcal{U}_{\text{true}} = \left\{ a' : |a'_j - a_j| \leq \sigma_j |a_j|, \quad j = 1, \ldots, n \right\}
\]

of values of the perturbed coefficient vector

\[
a' = ((1 + \epsilon_1) a_1, \ldots, (1 + \epsilon_n) a_n)^T
\]

with ellipsoid (Ell).
• It is easily seen that

\[
\text{Prob}\left\{ \zeta < \text{Mean}\{\zeta\} - \kappa \sqrt{\sum_{j=1}^{n} \sigma_j^2 a_j^2 x_j^2} \right\} \leq \exp\left\{ -\frac{\kappa^2}{2} \right\}
\]

The probability of the “rare event” we are ignoring when replacing $U_{\text{true}}$ with $U_{5.26}$ is $< 10^{-6}$. Note that for $n$ large and all $\sigma_j$ are of the same order of magnitude, the ellipsoid $U_{5.26}$ is a “negligible part” of the box $U_{\text{true}}$!
Proof of the Probability Bound

**Theorem** [Hoeffding’s Inequality] Let $c_i$, $\sigma_i$ be deterministic reals, and $\xi_i$ be independent random variables with zero mean such that $|\xi_i| \leq \sigma_i$. Then for every $\kappa > 0$ one has

$$p(\kappa) = \text{Prob}\left\{ \sum c_i \xi_i > \kappa \sqrt{\sum c_i^2 \sigma_i^2} \right\} \leq \exp\left\{ -\kappa^2 / 2 \right\}.$$

**Proof.** For $\gamma > 0$ we have

$$\exp\{\gamma \kappa \sigma\} p(\kappa) \leq \mathbb{E}\left\{ \exp\{\gamma \sum c_i \xi_i\} \right\} = \prod_i \mathbb{E}\{\exp\{\gamma c_i \xi_i\}\}$$

$$= \prod_i \mathbb{E}\left\{ \exp\{\gamma c_i \xi_i\} - \sinh(\gamma c_i \sigma_i)\sigma_i^{-1} \xi_i \right\} \quad [\text{since } \mathbb{E}\{\xi_i\} = 0]$$

$$\leq \prod_i \max_{-\sigma_i \leq s_i \leq \sigma_i} \left[ \exp\{\gamma c_i s_i\} - \sinh(\gamma c_i \sigma_i)\sigma_i^{-1} s_i \right]$$

$$g_i(s_i), g_i(\cdot): \text{convex}$$

$$g_i(\pm \sigma_i) = \cosh(\gamma c_i \sigma_i)$$

$$= \prod_i \cosh(\gamma c_i \sigma_i) = \prod_i \left[ \sum_{k=0}^{\infty} \frac{[\gamma^2 c_i^2 \sigma_i^2]^k}{(2k)!} \right] \leq \prod_i \left[ \sum_{k=0}^{\infty} \frac{[\gamma^2 c_i^2 \sigma_i^2]^k}{2^k k!} \right]$$

$$= \prod_i \exp\left\{ \frac{\gamma^2 c_i^2 \sigma_i^2}{2} \right\} = \exp\{\gamma^2 \sigma_i^2\}.$$

Thus,

$$p(\kappa) \leq \min_{\gamma > 0} \exp\{\gamma^2 \sigma_i^2 / 2 - \gamma \kappa \sigma\} = \exp\left\{ -\kappa^2 / 2 \right\}.$$
Applying the outlined methodology to our Antenna example:

\[
\min_{x, \tau} \left\{ \tau : -\tau \leq D_\star(\theta_\ell) - \sum_{j=1}^{10} x_j D_j(\theta_\ell) \leq \tau, 1 \leq \ell \leq 120 \right\} \tag{LP}
\]

\[
\downarrow
\]

\[
\begin{align*}
\min_{x, \tau} & \quad \tau \\
D_\star(\theta_\ell) - \sum_{j=1}^{10} x_j D_j(\theta_\ell) + \kappa \sigma \sqrt{\sum_{j=1}^{10} x_j^2 D_j^2(\theta_\ell)} & \leq \tau \\
D_\star(\theta_\ell) - \sum_{j=1}^{10} x_j D_j(\theta_\ell) - \kappa \sigma \sqrt{\sum_{j=1}^{10} x_j^2 D_j^2(\theta_\ell)} & \geq -\tau
\end{align*} \tag{RC}
\]

\[
[\sigma = 0.001]
\]

we get a robust design.
The results of “Robust Antenna Design” ($\kappa = 1$) are as follows:

- A typical “robust” diagram

- The diagram shown on the picture is at uniform distance 0.0822 from the target (just by 30% larger than the “nominal optimal value” is 0.0622 given by “nominal design” which ignores the implementation errors)

- As a compensation, robust design is incomparably more stable than the nominal one: in a sample of 40 realizations of “robust diagrams”, the uniform distance to the target varies from 0.0814 to 0.0830.

- When implementation errors become 10 times larger (1% instead of 0.1%), the “robust design” remains nearly as good as in the case of 0.1%-perturbations: now in a sample of 40 realizations of “robust diagrams”, the uniform distance to the target varies from 0.0834 to 0.116.

2.70
Why the “nominal design” is that sensitive to implementation errors? The basic diagrams $D_j(\cdot)$ are “nearly linearly dependent”. As a result, the nominal problem is “ill-posed” – it possesses a huge domain comprised of “nearly optimal” solutions. Indeed, look what are the optimal values in the nominal Antenna Design LP with added box constraints $|x_j| \leq L$ on the variables:

<table>
<thead>
<tr>
<th>$L$</th>
<th>1</th>
<th>10</th>
<th>$10^2$</th>
<th>$10^3$</th>
<th>$10^4$</th>
<th>$10^5$</th>
<th>$10^6$</th>
<th>10'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opt.Val</td>
<td>0.09449</td>
<td>0.07994</td>
<td>0.07358</td>
<td>0.06955</td>
<td>0.06588</td>
<td>0.06272</td>
<td>0.06215</td>
<td>0.06215</td>
</tr>
</tbody>
</table>

The “exactly optimal” solution to the nominal problem is very large, and therefore even small relative implementation errors may completely destroy the corresponding design.

In the robust counterpart, magnitudes of candidate solutions are penalized, so that RC implements a smart trade-off between the optimality and the magnitude (i.e., the stability) of the solution.

<table>
<thead>
<tr>
<th>$j$</th>
<th>$x_{j,\text{nom}}$</th>
<th>$x_{j,\text{rob}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.6e3</td>
<td>-0.30</td>
</tr>
<tr>
<td>2</td>
<td>-1.4e4</td>
<td>5.0</td>
</tr>
<tr>
<td>3</td>
<td>5.5e4</td>
<td>-3.4</td>
</tr>
<tr>
<td>4</td>
<td>-1.1e5</td>
<td>-5.1</td>
</tr>
<tr>
<td>5</td>
<td>9.5e4</td>
<td>6.9</td>
</tr>
<tr>
<td>6</td>
<td>1.9e4</td>
<td>5.5</td>
</tr>
<tr>
<td>7</td>
<td>-1.3e5</td>
<td>5.3</td>
</tr>
<tr>
<td>8</td>
<td>1.4e6</td>
<td>-7.5</td>
</tr>
<tr>
<td>9</td>
<td>-6.9e4</td>
<td>-8.9</td>
</tr>
<tr>
<td>10</td>
<td>1.3e4</td>
<td>13</td>
</tr>
</tbody>
</table>
How it works? NETLIB Case Study

♣ We solved the Robust Counterparts of the bad NETLIB problems, assuming interval uncertainty in “ugly coefficients" of inequality constraints and no uncertainty in equations. It turns out that

- Reliable solutions do exist, except for 4 cases corresponding to the highest ($\epsilon = 1\%$) perturbation level.
- The “price of immunization” in terms of the objective value is surprisingly low: when $\epsilon \leq 0.1\%$, it never exceeds 1% and it is less than 0.1% in 13 of 23 cases. Thus, passing to the robust solutions, we gain a lot in the ability of the solution to withstand data uncertainty, while losing nearly nothing in optimality.
<table>
<thead>
<tr>
<th>Problem</th>
<th>Nominal optimal value</th>
<th>$\epsilon = 0.01%$</th>
<th>$\epsilon = 0.1%$</th>
<th>$\epsilon = 1%$</th>
</tr>
</thead>
<tbody>
<tr>
<td>80BAU3B</td>
<td>987224.2</td>
<td>987311.8 (+ 0.01%)</td>
<td>989084.7 (+ 0.19%)</td>
<td>1009229 (+ 2.23%)</td>
</tr>
<tr>
<td>25FV47</td>
<td>5501.846</td>
<td>5501.862 (+ 0.00%)</td>
<td>5502.191 (+ 0.01%)</td>
<td>5505.653 (+ 0.07%)</td>
</tr>
<tr>
<td>ADLITTLE</td>
<td>225495.0</td>
<td>225594.2 (+ 0.04%)</td>
<td>228061.3 (+ 1.14%)</td>
<td></td>
</tr>
<tr>
<td>AFIR0</td>
<td>-464.7531</td>
<td>-464.7500 (+ 0.00%)</td>
<td>-464.2613 (+ 0.11%)</td>
<td></td>
</tr>
<tr>
<td>BNL2</td>
<td>1811.237</td>
<td>1811.237 (+ 0.00%)</td>
<td>1811.338 (+ 0.01%)</td>
<td></td>
</tr>
<tr>
<td>BRANDY</td>
<td>1518.511</td>
<td></td>
<td>1518.581 (+ 0.00%)</td>
<td></td>
</tr>
<tr>
<td>CAPRI</td>
<td>1912.621</td>
<td>1912.738 (+ 0.01%)</td>
<td>1913.958 (+ 0.07%)</td>
<td></td>
</tr>
<tr>
<td>CYCLE</td>
<td>1913.958</td>
<td>1913.958 (+ 0.00%)</td>
<td>1913.958 (+ 0.00%)</td>
<td></td>
</tr>
<tr>
<td>D2Q06C</td>
<td>122784.2</td>
<td>122793.1 (+ 0.01%)</td>
<td>122893.8 (+ 0.09%)</td>
<td></td>
</tr>
<tr>
<td>E226</td>
<td>-18.75193</td>
<td></td>
<td>-18.75173 (+ 0.00%)</td>
<td></td>
</tr>
<tr>
<td>FINNIS</td>
<td>172791.1</td>
<td>172808.8 (+ 0.01%)</td>
<td>173269.4 (+ 0.28%)</td>
<td>178448.7 (+ 3.27%)</td>
</tr>
<tr>
<td>GREENBEA</td>
<td>-72555250</td>
<td>-72526140 (+ 0.04%)</td>
<td>-72192920 (+ 0.50%)</td>
<td>-68869430 (+ 5.08%)</td>
</tr>
<tr>
<td>KB2</td>
<td>-1749.900</td>
<td>-1749.877 (+ 0.00%)</td>
<td>-1749.638 (+ 0.01%)</td>
<td>-1746.613 (+ 0.19%)</td>
</tr>
<tr>
<td>MAROS</td>
<td>-58063.74</td>
<td>-58063.45 (+ 0.00%)</td>
<td>-58011.14 (+ 0.09%)</td>
<td>-57312.23 (+ 1.29%)</td>
</tr>
<tr>
<td>NESM</td>
<td>14076040</td>
<td></td>
<td>14172030 (+ 0.68%)</td>
<td></td>
</tr>
<tr>
<td>PEROLD</td>
<td>-9380.755</td>
<td>-9380.755 (+ 0.00%)</td>
<td>-9362.653 (+ 0.19%)</td>
<td>Infeasible</td>
</tr>
<tr>
<td>PILOT</td>
<td>-557.4875</td>
<td>-557.4538 (+ 0.01%)</td>
<td>-555.3021 (+ 0.39%)</td>
<td>Infeasible</td>
</tr>
<tr>
<td>PILOT4</td>
<td>-64195.51</td>
<td>-64149.13 (+ 0.07%)</td>
<td>-63584.16 (+ 0.95%)</td>
<td>-58113.67 (+ 9.47%)</td>
</tr>
<tr>
<td>PILOT87</td>
<td>301.7109</td>
<td>301.7188 (+ 0.00%)</td>
<td>302.2191 (+ 0.17%)</td>
<td>Infeasible</td>
</tr>
<tr>
<td>PILOTJA</td>
<td>-6113.136</td>
<td>-6113.059 (+ 0.00%)</td>
<td>-6104.153 (+ 0.15%)</td>
<td>-5943.937 (+ 2.77%)</td>
</tr>
<tr>
<td>PILOTNOV</td>
<td>-4497.276</td>
<td>-4496.421 (+ 0.02%)</td>
<td>-4488.072 (+ 0.20%)</td>
<td>-4405.665 (+ 2.04%)</td>
</tr>
<tr>
<td>PILOTWE</td>
<td>-2720108</td>
<td>-2719502 (+ 0.02%)</td>
<td>-2713356 (+ 0.25%)</td>
<td>-2651786 (+ 2.51%)</td>
</tr>
<tr>
<td>SCFXM1</td>
<td>18416.76</td>
<td>18417.09 (+ 0.00%)</td>
<td>18420.66 (+ 0.02%)</td>
<td>18470.51 (+ 0.29%)</td>
</tr>
<tr>
<td>SCFXM2</td>
<td>36660.26</td>
<td>36660.82 (+ 0.00%)</td>
<td>36666.86 (+ 0.02%)</td>
<td>36764.43 (+ 0.28%)</td>
</tr>
<tr>
<td>SCFXM3</td>
<td>54901.25</td>
<td>54902.03 (+ 0.00%)</td>
<td>54910.49 (+ 0.02%)</td>
<td>55055.51 (+ 0.28%)</td>
</tr>
<tr>
<td>SHARE1B</td>
<td>-76589.32</td>
<td>-76589.32 (+ 0.00%)</td>
<td>-76589.32 (+ 0.00%)</td>
<td>-76589.29 (+ 0.00%)</td>
</tr>
</tbody>
</table>

Objective values for nominal and robust solutions to bad NETLIB problems.
The rationale behind the Robust Optimization paradigm as applied to LP is based on two assumptions:

1. Constraints are a “must”: a meaningful solution should satisfy all realizations of the constraints allowed by the uncertainty set.

2. All decision variables should be specified (get numeric values) before the true data becomes known and thus should be independent of the true data.

In many cases, Assumption 2 is too conservative:

A. In dynamical decision-making, only part of decision variables correspond to “here and now” decisions, while the remaining variables represent “wait and see” decisions which are to be made when certain part of the true data is already revealed. A “wait and see” decision can – and should! – depend on the corresponding part of the true data.

B. Some of the decision variables do not correspond to actual decisions at all; they are artificial “analysis variables” introduced to convert the problem into the LP form. These variables can – and should! – depend on the entire true data.
**Example:** Consider the problem of finding the best $\| \cdot \|_1$-approximation

$$
\min_{x,t} \left\{ t : \sum_i |b_i - \sum_j a_{ij}x_j| \leq t \right\}. \tag{P}
$$

When the data are certain, this problem is equivalent to the LP program

$$
\min_{x,y,t} \left\{ t : \sum_i y_i \leq t, -y_i \leq b_i - \sum_j a_{ij}x_j \leq y_i \forall i \right\}. \tag{LP}
$$

With uncertain data, the Robust Counterpart of (P) becomes the semi-infinite problem

$$
\min_{x,t} \left\{ t : \sum_i |b_i - \sum_j a_{ij}x_j| \leq t \forall (b_i, a_{ij}) \in \mathcal{U} \right\},
$$

or, which is the same, the problem

$$
\min_{x,t} \left\{ t : \forall (b_i, a_{ij}) \in \mathcal{U} \exists y : \sum_i y_i \leq t, -y_i \leq b_i - \sum_j a_{ij}x_j \leq y_i \right\}, \tag{RCP}
$$

while the RC of (LP) is the much more conservative problem

$$
\min_{x,t} \left\{ t : \exists y : \forall (b_i, a_{ij}) \in \mathcal{U} : \sum_i y_i \leq t, -y_i \leq b_i - \sum_j a_{ij}x_j \leq y_i \right\}. \tag{RCLP}
$$
Adjustable Robust Counterpart of an Uncertain LP

Consider an uncertain LP. W.l.o.g., we may assume that the data of this LP are affinely parameterized by a “perturbation vector” $\zeta$ running through a given perturbation set $\mathcal{Z}$:

$$\mathcal{LP} = \left\{ \min_x \left\{ c^T[\zeta]x : A[\zeta]x - b[\zeta] \geq 0 \right\} : \zeta \in \mathcal{Z} \right\}$$

$c_j[\zeta], A_{ij}[\zeta], b_i[\zeta]$ are affine in $\zeta$

Assume that every decision variable may depend on a given “portion” of the true data. Since the latter is affine in $\zeta$, this assumption says that $x_j$ may depend on $P_j\zeta$, where $P_j$ are given matrices.

- $P_j = 0 \Rightarrow x_j$ is non-adjustable: this is an independent of the true data “here and now” decision;

- $P_j \neq 0 \Rightarrow x_j$ is adjustable: this is a “wait and see” decision or an analysis variable which may adjust itself – fully or partially, depending on $P_j$ – to the true data.
\( \mathcal{LP} = \{ \min_x \{ c^T[\zeta]x : A[\zeta]x - b[\zeta] \geq 0 \} : \zeta \in \mathcal{Z} \} \)

[\( c_j[\zeta], A_{ij}[\zeta], b_i[\zeta] \) are affine in \( \zeta \)]

♣ In our now circumstances, a natural Robust Counterpart of \( \mathcal{LP} \) is the problem

Find \( t \) and functions \( \phi_j(\cdot) \) such that the decision rules \( x_j = \phi_j(P_j\zeta) \) make all the constraints feasible for all perturbations \( \zeta \in \mathcal{Z} \), while minimizing the guaranteed value \( t \) of the objective:

\[
\min_{t,\{\phi_i(\cdot)\}} \left\{ t : \begin{array}{c}
\sum_j c_j[\zeta] \phi_j(P_j\zeta) \leq t \forall \zeta \in \mathcal{Z} \\
\sum_j \phi_j(P_j\zeta) A_j[\zeta] - b[\zeta] \geq 0 \forall \zeta \in \mathcal{Z}
\end{array} \right\} \quad (ARC)
\]
Very bad news: The **Adjustable Robust Counterpart**

$$\begin{align*}
\min_{t,\{\phi_i(\cdot)\}} \left\{ t : \begin{array}{l}
\sum_j c_j[\zeta] \phi_j(P_j\zeta) \leq t \forall \zeta \in \mathbb{Z} \\
\sum_j \phi_j(P_j\zeta) A_j[\zeta] - b[\zeta] \geq 0 \forall \zeta \in \mathbb{Z}
\end{array} \right\} \quad \text{(ARC)}
\end{align*}$$

of uncertain LP is an *infinite-dimensional* optimization program and as such typically is absolutely intractable: How could we represent efficiently general-type functions of many variables, not speaking about how to optimize with respect to these functions?

Remedy (perhaps?): Let us restrict the decision rules $x_j = \phi_j(P_j\zeta)$ to be easily representable – specifically, *affine* – functions:

$$\phi_j(P_j\zeta) \equiv \mu_j + \nu_j^T P_j\zeta.$$ 

With this dramatic simplification, (ARC) becomes a *finite-dimensional* (still semi-infinite) optimization problem in new non-adjustable variables $\mu_j, \nu_j$

$$\begin{align*}
\min_{t,\{\mu_j,\nu_j\}} \left\{ t : \begin{array}{l}
\sum_j c_j[\zeta](\mu_j + \nu_j^T P_j\zeta) \leq t \forall \zeta \in \mathbb{Z} \\
\sum_j (\mu_j + \nu_j^T P_j\zeta) A_j[\zeta] - b[\zeta] \geq 0 \forall \zeta \in \mathbb{Z}
\end{array} \right\} \quad \text{(AARC)}
\end{align*}$$

2.78
We have associated with uncertain LP
\[ \mathcal{LP} = \left\{ \min_x \left\{ c^T[\zeta]x : A[\zeta]x - b[\zeta] \geq 0 \right\} : \zeta \in \mathcal{Z} \right\} \]
\[ \left[ c_j[\zeta], A_{ij}[\zeta], b_i[\zeta] \right. \text{ are affine in } \zeta \]
and the “information matrices” \( P_1, \ldots, P_n \) the Affinely Adjustable Robust Counterpart
\[
\min_{t, \{\mu_j, \nu_j\}} \left\{ t : \sum_j c_j[\zeta](\mu_j + \nu_j^T P_j[\zeta]) \leq t \forall \zeta \in \mathcal{Z} \right. \\
\sum_j (\mu_j + \nu_j^T P_j[\zeta])A_{ij}[\zeta] - b[i][\zeta] \geq 0 \forall \zeta \in \mathcal{Z} \right\} \quad \text{(AARC)}
\]
Relatively good news:
A. AARC is by far more flexible than the usual (non-adjustable) RC of \( \mathcal{LP} \).
B. As compared to ARC, AARC has much more chances to be computationally tractable:
   — With “fixed recourse”, where the coefficients of adjustable variables are certain, AARC has the same tractability properties as RC: If the perturbation set \( \mathcal{Z} \) is CQr (or polyhedrally representable), (AARC) is equivalent to an explicit CQ (resp., LP) program.
   — In the general case, (AARC) may be computationally intractable; however, under mild assumptions on the perturbation set, (AARC) admits “tight” computationally tractable approximation.

2.79
Example: Simple Inventory Model. There is a single-product inventory system with
• a single warehouse which should at any time store at least $V_{\text{min}}$ and at most $V_{\text{max}}$ units of the product;
• uncertain demands $d_t$ of periods $t = 1, \ldots, T$ known to vary within given bounds:
  $$d_t \in [d_t^*(1 - \theta), d_t^*(1 + \theta)], \quad t = 1, \ldots, T$$
($\theta$ is the uncertainty level). No backlogged demand is allowed!
• $I$ factories from which the warehouse can be replenished:
  — at the beginning of period $t$, you may order $p_{t,i}$ units of product from factory $i$. Your orders should satisfy the constraints
    $$0 \leq p_{t,i} \leq P_i(t) \quad [\text{bounds on orders per period}]$$
    $$\sum_t p_{t,i} \leq Q_i \quad [\text{bounds on cumulative orders}]$$
  — there is no delivery delay
  — order $p_{t,i}$ costs you $c_i(t)p_{t,i}$.

The goal is to minimize the total cost of the orders.
With certain demand, the problem can be modelled as the LP program

\[
\min_{p_{t,i}, i \leq I, t \leq T, \sum_{t,i} c_i(t) p_{t,i}} \sum_{t, i} c_i(t) p_{t,i} \quad \text{[total cost]}
\]

\[
v_{t+1} - v_t - \sum_i p_{t,i} = d_t, \quad t = 1, \ldots, T \quad \text{[state equations. } v_t: \text{ inventory level at the end of day } t \text{ (} v_1 \text{ is given)}\]
\[
V_{\min} \leq v_t \leq V_{\max}, \quad 2 \leq t \leq T + 1 \quad \text{[bounds on states]}
\]
\[
0 \leq p_{t,i} \leq P_i(t), \quad i \leq I, t \leq T \quad \text{[bounds on orders]}
\]
\[
\sum_{t} p_{t,i} \leq Q_i, \quad i \leq I \quad \text{[cumulative bounds on orders]}
\]

With uncertain demand, it is natural to assume that the orders \( p_{t,i} \) may depend on the demands of the preceding periods \( 1, \ldots, t - 1 \). The analysis variables \( v_t \) are allowed to depend on the entire true data; in fact, it suffices to allow for \( v_t \) to depend on \( d_1, \ldots, d_{t-1} \).

- Applying the AARC methodology, we make \( p_{t,i} \) and \( v_t \) affine functions of past demands:

\[
p_{t,i} = \phi_{t,i}^0 + \sum_{1 \leq \tau < t} \phi_{t,i}^\tau d_\tau
\]
\[
v_t = \psi_t^0 + \sum_{1 \leq \tau < t} \psi_t^\tau d_\tau
\]
The AARC is the following semi-infinite LP in non-adjustable design variables $\phi$’s and $\psi$’s:

$$\begin{align*}
\text{min}_{C, \phi_{t,i}, \psi_i} & \quad C \\
\text{s.t.} & \quad \sum_{t,i} c_i(t) \left[ \phi_{t,i}^0 + \sum_{1 \leq \tau < t} \phi_{t,i}^\tau d_\tau \right] \leq C \\
& \quad \left[ \psi_{t+1}^0 + \sum_{\tau=1}^t \psi_{t+1}^\tau d_\tau \right] - \left[ \psi_{t}^0 + \sum_{\tau=1}^{t-1} \psi_{t}^\tau d_\tau \right] - \sum_i \left[ \phi_{t,i}^0 + \sum_{\tau=1}^{t-1} \phi_{t,i}^\tau d_\tau \right] = d_t \\
& \quad V_{\text{min}} \leq \left[ \psi_{t}^0 + \sum_{\tau=1}^{t-1} \psi_{t}^\tau d_\tau \right] \leq V_{\text{max}} \\
& \quad 0 \leq \left[ \phi_{t,i}^0 + \sum_{\tau=1}^{t-1} \phi_{t,i}^\tau d_\tau \right] \leq P_i(t) \\
& \quad \sum_t \left[ \phi_{t,i}^0 + \sum_{\tau=1}^{t-1} \phi_{t,i}^\tau d_\tau \right] \leq Q_i
\end{align*}$$

The constraints should be valid for all values of “free” indices and all demand realizations $d = \{d_t\}_{t=1}^T$ from the “demand uncertainty box”

$$\mathcal{D} = \{d : d_t^*(1 - \theta) \leq d_t \leq d_t^*(1 + \theta), 1 \leq t \leq T\}.$$

The AARC can be straightforwardly converted to a usual LP and easily solved.
In the numerical illustration to follow:

- the planning horizon is $T = 24$
- there are $I = 3$ factories with per period capacities $P_i(t) = 567$ and cumulative capacities $Q_i = 13600$
- the nominal demand $d_t^*$ is seasonal:

$$d_t^* = 1000 \left(1 + 0.5 \sin \left(\frac{\pi (t-1)}{12}\right)\right)$$

- the production costs also are seasonal:

$$c_i(t) = c_i \left(1 + 0.5 \sin \left(\frac{\pi (t-1)}{12}\right)\right), \ c_1 = 1, c_2 = 1.5, c_3 = 2$$

- $v_1 = V_{\text{min}} = 500, \ V_{\text{max}} = 2000$
- demand uncertainty $\theta = 20\%$

2.83
Results:

- The AARC optimal value is 35542.
  **Note:** The non-adjustable RC is infeasible even at 5% uncertainty level!

- With uniformly distributed in the range ±20% demand perturbations, the average, over 100 simulations, AARC management cost is 35121.
  **Note:** Over the same 100 simulations, the average “utopian” management cost (optimal for a priori known demand trajectories) is 33958, i.e., is by just 3.5% (!) less than the average AARC management cost.
Comparison with Dynamic Programming. When applicable, DP is the technique for dynamical decision-making under uncertainty – in (worst-case-oriented) DP, one solves the Adjustable Robust Counterpart of uncertain LP in question, with no ad hoc simplifications like “let us restrict ourselves with affine decision rules”.

Unfortunately, DP suffers from “curse of dimensionality” – with DP, the computational effort blows up rapidly as the state dimension of the dynamical process grows. Usually state dimension 4 is already “too big”.

Note: There is no “curse of dimensionality” in AARC!

- In our toy Inventory model, the state dimension is 4 (what matters for the future, is the current amount of product at the warehouse and 3 remaining cumulative capacities of the 3 factories). Thus, DP is hardly applicable.
- However, reducing the number of factories to 1, increasing the per period capacity of the remaining factory to 1800 and making its cumulative capacity $+\infty$, we reduce the state dimension to 1 and make DP easily implementable. With this setup, — the DP (that is, the “absolutely best”) optimal value is 31270 — the computed AARC optimal value is 31514 – just by 0.8% worse! In fact, 0.8% is due to rounding errors — it was shown [Bertsimas,Iancu,Parrilo’09] that in the case in question the ARC and the AARC optimal values are the same!

2.85
Whether Conic Quadratic Programming exists?
Fast Polyhedral Approximation of Lorentz Cone

♠ Fact: The canonical polyhedral representation \( X = \{x \in \mathbb{R}^n : Ax \leq b\} \) of the projection

\[
X = \{x : \exists u : Px + Qu \leq r\}
\]
of a polyhedral set \( X^+ = \{[x; u] : Px + Qu \leq r\} \) given by a moderate number of linear inequalities in variables \( x, u \) can require a huge number of linear inequalities in variables \( x \).

Question: Can we use this phenomenon in order to approximate to high accuracy a non-polyhedral set \( X \subset \mathbb{R}^n \) by projecting onto \( \mathbb{R}^n \) a higher-dimensional polyhedral and simple (given by a moderate number of linear inequalities) set \( X^+ \)?
The outlined possibility does exist when $X$ is the Lorentz cone.

**Theorem:** For every $n$ and every $\epsilon$, $0 < \epsilon < 1/2$, one can point out a polyhedral set $L^+$ given by an explicit system of homogeneous linear inequalities in variables $x \in \mathbb{R}^n$, $t \in \mathbb{R}$, $w \in \mathbb{R}^k$:

$$L^+ = \{ [x; t; w] : Px + tp + Qw \leq 0 \} \quad (!)$$

such that

- the number of inequalities in the system ($\approx 2n \ln(1/\epsilon)$) and the dimension of the slack vector $w$ ($\approx 0.7n \ln(1/\epsilon)$) do not exceed $O(1)n \ln(1/\epsilon)$
- the projection $L = \{ [x; t] : \exists w : Px + tp + Qw \leq 0 \}$ of $L^+$ on the space of $x, t$-variables is in-between the Second Order Cone and $(1 + \epsilon)$-extension of this cone:

$$L^{n+1} := \{ [x; t] \in \mathbb{R}^{n+1} : \|x\|_2 \leq t \} \subset L \subset L^{n+1}_\epsilon := \{ [x; t] \in \mathbb{R}^{n+1} : \|x\|_2 \leq (1 + \epsilon)t \}.$$

In particular, we have

$$B^1_n \subset \{ x : \exists w : Px + p + Qw \leq 0 \} \subset B^{1+\epsilon}_n$$

$$B^r_n = \{ x \in \mathbb{R}^n : \|x\|_2 \leq r \}$$

2.87
**Note:** When $\epsilon = 1.e-17$, a usual computer does not distinguish between $r = 1$ and $r = 1 + \epsilon$. Thus, *for all practical purposes*, the $n$-dimensional Euclidean ball admits polyhedral representation with $\approx 28n$ variables $w$ and $\approx 79n$ linear inequality constraints.

**Note:** A straightforward representation $X = \{x : Ax \leq b\}$ of a polyhedral set $X$ satisfying

$$B_n^1 \subset X \subset B_n^{1+\epsilon}$$

requires at least $N = O(1)\epsilon^{-\frac{n-1}{2}}$ linear inequalities. With $n = 100$, $\epsilon = 0.01$, we get

$$N \geq 3.0e85 \approx 300,000 \times [\# \text{ of atoms in universe}]$$

With “fast polyhedral approximation” of $B_n^1$, a 0.01-approximation of $B_{100}^1$ requires just 922 linear inequalities on 100 original and 325 additional variables.
With fast polyhedral approximation of the cone $L^{n+1} = \{[x; t] \in \mathbb{R}^{n+1} : \|x\|_2 \leq t\}$, Conic Quadratic Optimization programs “for all practical purposes” become LO programs. For example, by what we know about CQr functions/sets, the program

$$\text{minimize } c^T x \text{ subject to }$$

$$Ax = b$$

$$x \geq 0$$

$$\left(\sum_{i=1}^{8} |x_i|^3\right)^{1/3} \leq x_2^{1/7} x_3^{2/7} x_4^{3/7} + 2x_1^{1/5} x_5^{2/5} x_6^{1/5}$$

$$5x_2 \geq \frac{1}{x_1^{1/2} x_2} + \frac{2}{x_2^{1/3} x_3 x_4^{5/8}}$$

can be in a systematic fashion converted to Conic Quadratic Programming and thus “for all practical purposes” is just and LP program.
Building Fast Polyhedral Approximation

♣ Goal: To nearly represent by linear inequalities the set

\[ L^{n+1} = \{ [x_1; \ldots; x_n; t] : \sqrt{x_1^2 + \ldots + x_n^2} \leq t \} \]

that is, to find a polyhedrally represented set

\[ \hat{L} = \{ [x = [x_1; \ldots; x_n; t] : \exists w : P x + t p + Q w \leq 0 \} \]

such that

\[ L^{n+1} \subset \hat{L} \subset L_{\epsilon}^{n+1}, \]

\[ L_{\epsilon}^{n+1} = \{ [x_1; \ldots; x_n; t] : \sqrt{x_1^2 + \ldots + x_n^2} \leq (1 + \epsilon) t \} \]

• \( \epsilon > 0 \): given tolerance.

♠ Observation: It suffices to solve our problem when \( n = 2 \).

Reason: Inequality \( \sqrt{x_1^2 + \ldots + x_n^2} \leq t \) can be represented by a system of similar inequalities with 3 variables in each.
Example: To represent the set 
\[ L^6 = \{ [x; t] \in \mathbb{R}^6 : \sqrt{x_1^2 + x_2^2 + \ldots + x_5^2} \leq t \}, \]
by a system of constraints of the form \( \sqrt{p^2 + q^2} \leq r \), we
♠ add to \( x, t \) variable \( w_1 \) and write down the system
\[ \sqrt{x_4^2 + x_5^2} \leq w_1, \sqrt{x_1^2 + x_2^2 + x_3^2 + w_1^2} \leq t \]
• the system does represent \( L^6 \) — the projection of its solution set on the space of \( x, t \)-variables is exactly \( L^6 \)
• the “sizes” (\# of variables involved) of the constraints in the system are \( \leq 5 \), while the size of the constraint in the original description of \( L^6 \) was 6.
♠ add to \( x, t, w_1 \) variable \( w_2 \) and write down the system
\[ \sqrt{x_4^2 + x_5^2} \leq w_1, \sqrt{x_3^2 + w_1^2} \leq w_2, \sqrt{x_1^2 + x_2^2 + w_2^2} \leq t \]
This system still represents \( L^6 \), and the maximal size of its constraints is 4.
♠ add to \( x, t, w_1, w_2 \) variable \( w_3 \) and write down the system
\[ \sqrt{x_4^2 + x_5^2} \leq w_1, \sqrt{x_3^2 + w_1^2} \leq w_2, \sqrt{x_2^2 + w_2^2} \leq w_3, \sqrt{x_1^2 + w_3^2} \leq t \]
This system represents \( L^6 \), and all its constraints are of the form \( \sqrt{p^2 + q^2} \leq r \). We are done.
Note: The above recipe clearly extends from the 6-dimensional case to the general one. Representing $L^{n+1}$ via constraints of the form $\sqrt{p^2 + q^2} \leq r$ requires $n - 2$ additional variables and $n - 1$ constraints.

Note: The number of steps in the latter procedure can be reduced from $n - 2$ to $\text{Ceil}(\log_2(n)) - 1$ by using the same construction as when building CQR of the set $\{(t, x_1, \ldots, x_{2\mu}) \geq 0 : t \leq (x_1, \ldots, x_{2\mu})^{1/2\mu}\}$; the resulting number of constraints of the form $\sqrt{p^2 + q^2} \leq r$ and of additional variables still are (at most) $n - 1$ and $n - 2$ respectively.

♠ Conclusion: In order to find a tight polyhedral approximation of $L^{n+1} = \{[x_1; \ldots; x_n; t] : \sqrt{x_1^2 + \ldots + x_n^2} \leq t\}$, we can

• represent the constraint $\sqrt{x_1^2 + \ldots + x_n^2} \leq t$ by a system of inequalities of the form $\sqrt{p^2 + q^2} \leq r$

• replace every one of the resulting constraints by its tight polyhedral approximation.

Note: We should account for “accumulation of errors.” This is an easy task...
Fast polyhedral approximation of 
\[ L^3 = \{ [p; q; r] : \sqrt{p^2 + q^2} \leq r \} \]

“Ice-cream” cone \( L^3 \)

♣ Given variables \( p, q, r \), we choose a positive integer \( K \), and consider \( K + 1 \) points \( P_1, \ldots, P_{K+1} \) on the 2D plane as follows.

- The first points \( P_1 = [u_1; v_1] \) satisfies
  \[ u_1 \geq |p|, \ v_1 \geq |q| \]
  which can be represented by a system of 4 linear constraints in variables \( p, q, u_1, v_1 \).
The relation between $P_k = [u_k; v_k]$ and $P_{k+1} = [u_{k+1}; v_{k+1}]$ is as follows.

— we rotate $P_k$ clockwise by the angle $\phi_k = \frac{\pi}{2^{k+1}}$, thus getting a point $Q_k$.

— we reflect $Q_k$ w.r.t. the $u$-axis, thus getting point $Q'_k$.

— we impose on $P_{k+1} = [u_{k+1}; v_{k+1}]$ the restriction to belong to the vertical line passing through $Q_k$ and $Q'_k$ and to be not lower than $Q_k$ and $Q'_k$. 

\[ P_k \rightarrow P_{k+1} \]
Note: Relations between $P_k = [u_k; v_k]$ and $P_{k+1} = [u_{k+1}; v_{k+1}]$ amount to a system of linear constraints

\[
\begin{align*}
    u_{k+1} &= \cos(\phi_k) u_k + \sin(\phi_k) v_k \\
    v_{k+1} &= -\sin(\phi_k) u_k + \cos(\phi_k) v_k
\end{align*}
\]

right hand side: $u$-coordinate of $Q_k$ and $Q'_k$

\[
\begin{align*}
    v_{k+1} &\geq \sin(\phi_k) u_k - \cos(\phi_k) v_k
\end{align*}
\]

right hand side: $v$-coordinate of $Q_k$

in variables $u_k, v_k, u_{k+1}, v_{k+1}$. 

2.95
Let us write down all built so far constraints on original and additional variables

\[
\begin{align*}
    u_1 & \geq p \\
    u_1 & \geq -p \\
    v_1 & \geq q \\
    v_2 & \geq -q \\
    u_{k+1} & = \cos(\phi_k)u_k + \sin(\phi_k)v_k \\
    v_{k+1} & \geq -\sin(\phi_k)u_k + \cos(\phi_k)v_k \\
    v_{k+1} & \geq \sin(\phi_k)u_k - \cos(\phi_k)v_k \\
    k & = 1, \ldots, K
\end{align*}
\]

and augment this system by the requirement for \( P_{K+1} \) to be close to the segment \([0, r]\) of the \( u \)-axis:

\[
0 \leq u_{K+1} \leq r, \quad 0 \leq v_{K+1} \leq \tan(\phi_K) \cdot r
\]

**Observation 1:** When \( p, q, r \) can be augmented by properly selected \( u \)'s and \( v \)'s to satisfy the above constraints, we have

\[
\sqrt{p^2 + q^2} \leq r \sqrt{1 + \tan^2(\phi_K)}
\]

Indeed, by the above constraints on \( p, q, r \) and the additional variables, the points \( P_k = [u_k; v_k] \) satisfy

\[
\|[p; q]\|_2 \leq \|P_1\|_2 \leq \ldots \leq \|P_{K+1}\|_2 = \sqrt{u_{K+1}^2 + v_{K+1}^2} \leq r \sqrt{1 + \tan^2(\phi_K)}.
\]
<table>
<thead>
<tr>
<th>$u_1$</th>
<th>$\geq$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>$\geq$</td>
<td>$-p$</td>
</tr>
<tr>
<td>$v_1$</td>
<td></td>
<td>$q$</td>
</tr>
<tr>
<td>$v_2$</td>
<td></td>
<td>$-q$</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
    u_{k+1} &= \cos(\phi_k)u_k + \sin(\phi_k)v_k \\
    v_{k+1} &= -\sin(\phi_k)u_k + \cos(\phi_k)v_k \\
    v_{k+1} &= \sin(\phi_k)u_k - \cos(\phi_k)v_k \\
\end{align*}
\]

$k = 1, \ldots, K$

\[0 \leq u_{K+1} \leq r, \ 0 \leq v_{K+1} \leq \tan(\phi_K) \cdot r\]

**Observation 2:** When $\sqrt{p^2 + q^2} \leq r$, $p, q, r$ indeed can be augmented by $u$'s and $v$'s to satisfy our constraints.

This combines with Observation 1 to imply that the projection of the polyhedral set given by our constraints onto the space of $p, q, r$ variables is in-between the $L^3$ and $L^3_{\delta_K}$, with

\[
\delta_K = \sqrt{1 + \tan^2(\phi_K) - 1} = \sqrt{1 + \tan^2 \left( \frac{\pi}{2K+1} \right) - 1} \leq \frac{\pi^2}{2^{2K+2}}.
\]

$\Rightarrow$ To make $\delta_K \leq \epsilon$, we need just $O(1) \ln(1/\epsilon)$ additional variables and linear constraints!
To justify Observation 2, let us augment $p$, $q$ with $u$’s and $v$’s which “rigidly” satisfy the magenta constraints, specifically, let us set $u_1 = |p|$, $v_1 = |q|$, and let $P_{k+1}$ be the “highest” of the points $Q_k$, $Q'_k$:

Then

$$r \geq \sqrt{p^2 + q^2} = \|[p; q]\|_2 = \|P_1\|_2 = \ldots = \|P_{K+1}\|_2$$

and the angle between $P_{k+1}$ and the nonnegative ray of the $u$-axis does not exceed $\phi_k = \frac{\pi}{2k+1}$.

$\implies P_{K+1} = [u_{K+1}, v_{K+1}]$ indeed satisfies

$$0 \leq u_{K+1} \leq r \quad \text{and} \quad 0 \leq v_{K+1} \leq \tan(\phi_K) \cdot r.$$
To justify the claim on the angles, observe that with our “rigid” construction of $P_1, \ldots, P_{K+1}$,

- $P_1$ lives in the first quadrant, and $P_2$ is obtained from $P_1$ by rotating clockwise by the angle $\phi_1 = \pi/4$ (and, perhaps, reflecting the result w.r.t. the $u$-axis to bring it to the first quadrant). After rotation, the angle between the point and the $u$-axis does not exceed $\pi/4$, and reflection, if any, keeps this angle intact

$\Rightarrow P_2$ lives in the first quadrant and makes angle at most $\phi_1 = \pi/4$ with the $u$-axis

$\Rightarrow P_3$, which is obtained from $P_2$ by rotating clockwise by the angle $\phi_2 = \pi/8$ (and, perhaps, reflecting the result w.r.t. $u$-axis to bring it to the first quadrant), lives in the first quadrant and makes the angle at most $\phi_2 = \pi/8$ with the $u$-axis

$\Rightarrow \ldots \Rightarrow P_{K+1}$ lives in the first quadrant and makes angle at most $\phi_K = \frac{\pi}{2^{K+1}}$ with the $u$-axis.
The simplest way to build a polyhedral approximation of the Lorentz cone is to take the tangent planes along a “fine” finite grid of generators and to use, as the approximation, the resulting polyhedral cone:

This approach is a complete failure: the number of tangent planes required to get an 0.5-approximation of $L^m$ is at least

$$N = \sqrt{2\pi(m - 2)} \exp\{m/6\},$$

which is $> 429,481,377$ for $m = 100$.

With our approach, we approximate $L^m$ by a projection of a higher-dimensional polyhedron. When projecting an $N$-dimensional polyhedron onto a plane of dimension $<< N$, the number of facets may grow up exponentially, so that a low-dimensional projection of a “simple” high-dimensional polyhedron may have astronomically many facets. With our approach, we build a family of polyhedral cones $P^{m,k} \subset R^{O(mk)}$ given by just $O(mk)$ linear inequalities, while their projections $\hat{P}^{m,k}$ on $R^m$ have enough facets to approximate $L^m$ within accuracy $\exp\{-O(k)\}$:
Approximating sets by projections of higher-dimensional polyhedral sets, we can dramatically reduce the “size” of approximation. For example,

- When approximating the unit 2D circle by a projection of a higher-dimensional polytope $P$, we can get approximations as follows:
  - with $P$ given by 12 inequalities in 10 variables – accuracy $5.e-3$, as good as circumscribed polygon with 16 sides
  - with $P$ given by 18 inequalities in 13 variables – accuracy $3.e-4$, as good as circumscribed polygon with 127 sides
  - with $P$ given by 30 inequalities in 19 variables – accuracy $7.e-8$, as good as circumscribed polygon with 8,192 sides
  - with $P$ given by 54 inequalities in 31 variables – accuracy $4.e-15$, as good as circumscribed polygon with 34,200,933 sides
Polyhedral approximation of $L^m$ is basically the same as polyhedral approximation of $m$-dimensional Euclidean ball

$$B_m = \{ x \in \mathbb{R}^m : \|x\|_2 \leq 1 \}.$$ 

There is a less sophisticated way to approximate Euclidean balls by projections of polyhedral sets:

**Theorem [Lindenstrauss-Johnson]:** For two positive integers $N, n$ with $N \geq 10n$, random $n$-dimensional projection of $N$-dimensional unit box – the set

$$B = \{ x \in \mathbb{R}^n : \exists y \in \mathbb{R}^N : x = Ay, -1 \leq y_1, ..., y_N \leq 1 \}$$

[A: drawn at random from Gaussian distribution]

with probability approaching one as $N, n$ grow, is in-between two $n$-dimensional Euclidean balls with the ratio of radii $(1 + O(\sqrt{n/N}))$.

This result has tremendous theoretical implications. However, — no individual matrices $A$ yielding “nearly round” $B$ are known (pity! these matrices would be ideally suited for Compressed Sensing)

**Note:** Our fast polyhedral approximation is explicit!

— to make $B$ an $\epsilon$-approximation of $B_n$, you need $N = O(1/\epsilon^2)n$

**Note:** With fast polyhedral approximation, you need much smaller $N$: $N = O(\ln(1/\epsilon))n$
Open question: With fast polyhedral approximation, centrally symmetric ball $B_n$ is $\epsilon$-approximated by the projection of a highly asymmetric polyhedron of dimension $N = O(\ln(1/\epsilon))n$ given by $M = O(N)$ linear inequalities. Is it possible to make this higher-dimensional polyhedron centrally symmetric, preserving the type of dependence of $N, M$ on $n$ and $\epsilon$?
III. SEMIDEFINITE PROGRAMMING
Preliminaries

- The space $\mathbb{R}^{m \times n}$ of $m \times n$ matrices can be identified with $\mathbb{R}^{mn}$

  $$A = [a_{ij}]_{i=1,...,m} \leftrightarrow \text{Vec}(A) = (a_{11}, \ldots, a_{1n}, a_{21}, \ldots, a_{2n}, \ldots, a_{m1}, \ldots, a_{mn})^T$$

  The inner product of matrices induced by this representation is

  $$\langle A, B \rangle \equiv \sum_{i,j} A_{ij} B_{ij} = \text{Tr}(A^T B) = \text{Tr}(AB^T) \quad [A, B \in \mathbb{R}^{m \times n}]$$

  $$\left[ \text{Tr}(C) = \sum_{i=1}^n C_{ii}, C \in \mathbb{R}^{n \times n}, \text{is the trace of } C \right]$$

- In particular, the space $S^m$ of $m \times m$ symmetric matrices equipped with the inner product inherited from $\mathbb{R}^{m \times m}$:

  $$\langle A, B \rangle \equiv \sum_{i,j} A_{ij} B_{ij} = \text{Tr}(A^T B) = \text{Tr}(AB)$$

  is a Euclidean space ($\text{dim } S^m = \frac{m(m+1)}{2}$).

- The positive semidefinite symmetric $m \times m$ matrices form a cone (closed, convex, pointed and with a nonempty interior) in $S^m$:

  $$S^m_+ = \{ A \in S^m : \xi^T A \xi \geq 0 \ \forall \xi \in \mathbb{R}^m \}$$

3.1
\[ S^+_m = \{ A \in S^m : \xi^T A \xi \geq 0 \ \forall \xi \in \mathbb{R}^m \} \]

- **Equivalent descriptions of** \( S^+_m \): an \( m \times m \) matrix \( A \) is positive semidefinite
  - iff \( A \) is symmetric \( (A = A^T) \) and all its eigenvalues are nonnegative;
  - iff \( A \) can be decomposed as \( A = D^T D \)
  - iff \( A \) can be represented as a sum of symmetric dyadic matrices:
    \[ A = \sum_j d_j d_j^T; \]
  - iff \( A = U^T \Lambda U \) with orthogonal \( U \) and diagonal \( \Lambda \), the diagonal entries of \( \Lambda \) being nonnegative;
  - iff \( A \) is symmetric \( (A = A^T) \) and all principal minors of \( A \) are nonnegative.

- As every cone, \( S^+_m \) defines a “good” partial ordering on \( S^m \):
  \[
  A \succeq B \iff A - B \succeq 0 \iff \xi^T A \xi \geq \xi^T B \xi \quad \forall \xi \\
  \left[ A = A^T, B = B^T \text{ are of the same size} \right]
  \]
• **Useful observation:** Validity of $\succeq$ inequality is preserved when multiplying both sides by a matrix $Q$ from the left and by $Q^T$ from the right:

$$A \succeq B \Rightarrow Q^T AQ \succeq Q^T BQ \quad [A, B \in S^m, Q \in \mathbb{R}^{m \times k}]$$

Indeed,

$$\left\{ \xi^T A \xi \geq \xi^T B \xi \quad \forall \xi \right\} \Rightarrow \left\{ \eta^T Q^T A Q \eta \geq \eta^T Q^T B Q \eta \quad \forall \eta \right\}$$

• **Useful observation:** When $A$ and $B$ are rectangular matrices such that $\text{Tr}(AB)$ is well defined (i.e., $AB$ is well defined and square), we have

$$\text{Tr}(AB) = \text{Tr}(BA).$$
• **Observation**: The semidefinite cone is self-dual:

\[
(S^+_m)^* \equiv \left\{ A \in S^m : \text{Tr}(AB) \geq 0 \ \forall B \in S^+_m \right\} = S^+_m.
\]

Indeed,

\[
\xi^T A \xi = \text{Tr}(\xi^T A \xi) = \text{Tr}(A \xi \xi^T)
\]

It follows that if \( A \in S^m \) is such that \( \text{Tr}(AB) \geq 0 \) for all \( B \succeq 0 \), then \( A \succeq 0 \):

\[
\xi \in \mathbb{R}^m \Rightarrow B = \xi \xi^T \succeq 0 \Rightarrow \text{Tr}(AB) = \xi^T A \xi \geq 0
\]

Vice versa, if \( A \in S^m_+ \), then \( \text{Tr}(AB) \geq 0 \) for all \( B \succeq 0 \):

\[
B \succeq 0 \Rightarrow B = \sum_j d_j d_j^T \Rightarrow \text{Tr}(AB) = \sum_j \text{Tr}(A d_j d_j^T) = \sum_j d_j^T A d_j \geq 0.
\]
Semidefinite program

• A *semidefinite program* is a conic program associated with the semidefinite cone:

\[
\min_x \left\{ c^T x : Ax - B \succeq 0 \quad \left[ \iff Ax - B \succeq \mathbb{S}_+ ^m 0 \right] \right\}
\]

\[
[ Ax = \sum_{i=1}^{\dim x} x_i A_i, \quad A_i \in \mathbb{S}^m ]
\]

A constraint of the type

\[ x_1 A_1 + \ldots + x_n A_n \succeq B \]

with variables \( x_1, \ldots, x_n \) is called an *LMI* – Linear Matrix Inequality. Thus, a semidefinite program is to minimize a linear objective under an LMI constraint.

• **Observation:** A *system* of LMI constraints

\[
A_i(x) := \sum_j x_j A_{ij} - B_i \succeq 0, \quad i = 1, \ldots, m
\]

is equivalent to *single* LMI constraint

\[
\text{Diag}\{ A_1(x), \ldots, A_m(x) \} \succeq 0.
\]
Program dual to an SDP program

\[
\min_x \left\{ c^T x : Ax - B \equiv \sum_{j=1}^n x_j A_j - B \succeq 0 \right\} \quad \text{(SDPr)}
\]

According to our general scheme, the problem dual to (SDPr) is

\[
\max_Y \left\{ \langle B, Y \rangle : A^* Y = c, Y \succeq 0 \right\} \quad \text{(SDDl)}
\]

(recall that \( S^m_+ \) is self-dual!).

It is easily seen that the operator \( A^* \) conjugate to \( A \) is given by

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

Consequently, the dual problem is

\[
\max_Y \left\{ \text{Tr}(BY) : \text{Tr}(YA_i) = c_i, i = 1, \ldots, n, Y \succeq 0 \right\} \quad \text{(SDDl)}
\]
**SDP optimality conditions**

\[
\begin{align*}
\min_x \left\{ c^T x : A x - B \equiv \sum_{j=1}^{n} x_j A_j - B \succeq 0 \right\} & \quad \text{(SDPr)} \\
\max_Y \left\{ \text{Tr}(BY) : \text{Tr}(A_j Y) = c_j, \; j = 1, \ldots, n; \; Y \succeq 0 \right\} & \quad \text{(SDDl)}
\end{align*}
\]

- Assume that (!) both (SDPr) and (SDDl) are strictly feasible, so that by Conic Duality Theorem both problems are solvable with equal optimal values.

By Conic Duality, the necessary and sufficient condition for a primal-dual feasible pair \((x, Y)\) to be primal-dual optimal is that

\[
\text{Tr}( \underbrace{A x - B}_{\text{"primal slack" } X} Y ) = 0
\]

- For a pair of symmetric positive semidefinite matrices \(X\) and \(Y\), one has

\[
\text{Tr}(XY) = 0 \iff XY = YX = 0.
\]

3.7
\[
\min_x \left\{ c^T x : Ax - B \equiv \sum_{j=1}^{n} x_j A_j - B \geq 0 \right\} \quad \text{(SDPr)}
\]
\[
\max_Y \left\{ \text{Tr}(BY) : \text{Tr}(A_j Y) = c_j, \ j = 1, \ldots, n; \ Y \succeq 0 \right\} \quad \text{(SDDI)}
\]

(!) both (SDPr) and (SDDI) are strictly feasible,

- Thus, **under assumption (!) a primal-dual feasible pair \((x,Y)\) is primal-dual optimal iff**

\[
[Ax - B]Y = Y[Ax - B] = 0
\]

Cf. Linear Programming:

(P): \[
\min_x \left\{ c^T x : Ax - b \geq 0 \right\}
\]

(D): \[
\max_y \left\{ b^T y : A^T y = c, \ y \geq 0 \right\}
\]

\((x,y)\) primal-dual optimal

\(\Updownarrow\)

\((x,y)\) primal-dual feasible and \(y_j[Ax - b]_j = 0 \ \forall j\)

3.8
What can be expressed via SDP?

\[
\min_x \left\{ c^T x : x \in X \right\}
\] (Ini)

- A sufficient condition for (Ini) to be equivalent to an SD program is that \( X \) is a SDr (“SemiDefinite-representable”) set:

**Definition.** A set \( X \subset \mathbb{R}^n \) is called SDr, if it admits SDR (“SemiDefinite Representation”)

\[
X = \{ x : \exists u : A(x, u) \succeq 0 \}
\]

\[
A(x, u) = \sum_j x_j A_j + \sum_\ell u_\ell B_\ell + C : \mathbb{R}^n_x \times \mathbb{R}^k_u \rightarrow \mathbb{S}^m
\]

- Given a SDR of \( X \), we can write down (Ini) equivalently as the semidefinite program

\[
\min_{x, u} \left\{ c^T x : A(x, u) \succeq 0 \right\}.
\]
Same as in the case of Conic Quadratic Programming, we can

Define the notion of a SDr function

\[ f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\} \]

as a function with SDr epigraph:

\[ \{(t, x) : t \geq f(x)\} = \{(t, x) : \exists u : A(t, x, u) \succeq 0 \} \]

and verify that if \( f \) is a SDr function, then all its level sets

\[ \{x : f(x) \leq a\} \]

are SDr;

Develop a “calculus” of SDr functions/sets with exactly the same combination rules as for CQ-representability.
When a function/set is SDr?

**Proposition.** Every CQr set/function is SDr as well.

**Proof.**

1. **Lemma.** Every direct product of Lorentz cones is SDr.

2. **Lemma⇒Proposition:** Let \( X \subset \mathbb{R}^n \) be CQr:

\[
X = \{ x \mid \exists u : A(x,u) \in K \},
\]

\( K \) being a direct product of Lorentz cones and \( A(x,u) \) being affine. By Lemma,

\[
K = \{ y : \exists v : B(y,v) \succeq 0 \}
\]

with affine \( B(\cdot,\cdot) \). It follows that

\[
X = \left\{ x : \exists u,v : \underbrace{B(A(x,u),v)}_{\text{LMI}} \succeq 0 \right\},
\]

which is a SDR for \( X \).
Lemma. Every direct product of Lorentz cones is SDr.

Proof. It suffices to prove that a Lorentz cone $L^m$ is a SDr set (since SD-representability is preserved when taking direct products).

To prove that $L^m$ is SDr, let us make use of the following Lemma on Schur Complement. A symmetric block matrix

$$A = \begin{pmatrix} P & Q^T \\ Q & R \end{pmatrix}$$

with positive definite $R$ is positive (semi)definite iff the matrix

$$P - Q^T R^{-1} Q$$

is positive (semi)definite.
**LSC⇒Lemma:** Consider the linear mapping

\[
\begin{bmatrix}
x_1 \\
x_2 \\
... \\
x_m
\end{bmatrix}
\mapsto Ax =
\begin{pmatrix}
x_m & x_1 & x_2 & x_3 & ... & x_{m-1} \\
x_1 & x_m \\
x_2 & x_m \\
x_3 & x_m \\
... & ... \\
x_{m-1} & x_m
\end{pmatrix}
\]

We claim that

\[
L^m = \{ x : A(x) \succeq 0 \}.
\]

Indeed,

\[
L^m = \left\{ x \in \mathbb{R}^m : x_m \geq \sqrt{x_1^2 + ... + x_{m-1}^2} \right\}
\]

and therefore

- if \( x \in L^m \) is nonzero, then \( x_m > 0 \) and

  \[
x_m - (x_1^2 + x_2^2 + ... + x_{m-1}^2)/x_m \geq 0
  \]

  so that \( A(x) \succeq 0 \) by LSC. If \( x = 0 \), then \( A(x) = 0 \succeq 0 \).

- if \( A(x) \succeq 0 \) and \( A(x) \neq 0 \), then \( x_m > 0 \) and, by LSC,

  \[
x_m - (x_1^2 + x_2^2 + ... + x_{m-1}^2)/x_m \geq 0 \Rightarrow x \in L^m.
  \]

And if \( A(x) = 0 \), then \( x = 0 \in L^m \).
Lemma on Schur Complement. A symmetric block matrix

\[
A = \begin{bmatrix}
P & Q^T \\
Q & R
\end{bmatrix}
\]

with positive definite \(R\) is positive (semi)definite iff the matrix

\[
P - Q^T R^{-1} Q
\]

is positive (semi)definite.

Proof. \(A\) is \(\succeq 0\) if and only if

\[
\inf_v \begin{bmatrix} u \\ v \end{bmatrix}^T \begin{bmatrix} P & Q^T \\
Q & R
\end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \geq 0 \quad \forall u. \tag{\ast}
\]

When \(R \succ 0\), the left hand side inf can be easily computed and turns to be

\[
u^T (P - Q^T R^{-1} Q) u.
\]

Thus, (\ast) is valid if and only if

\[
u^T (P - Q^T R^{-1} Q) u \geq 0 \quad \forall u,
\]

i.e., iff

\[
P - Q^T R^{-1} Q \succeq 0.
\]
More examples of SD-representable functions/sets

- **The largest eigenvalue** $\lambda_{\text{max}}(X)$ regarded as a function of $m \times m$ symmetric matrix $X$ is SDr:

  $$\lambda_{\text{max}}(X) \leq t \iff tI_m - X \succeq 0,$$

  $I_k$ being the unit $k \times k$ matrix.

- **The largest eigenvalue of a matrix pencil.** Let $M, A \in S^m$ be such that $M \succ 0$.
  The eigenvalues of the pencil $[M, A]$ are reals $\lambda$ such that the matrix $\lambda M - A$ is singular, or, equivalently, such that

  $$\exists e \neq 0 : \quad A e = \lambda M e.$$

  The eigenvalues of the pencil $[M, A]$ are the usual eigenvalues of the symmetric matrix $D^{-1} A D^{-T}$, where $D$ is such that $M = D D^T$.

  The largest eigenvalue $\lambda_{\text{max}}(X : M)$ of a pencil $[M, X]$ with $M \succ 0$, regarded as a function of $X$, is SDr:

  $$\lambda_{\text{max}}(X : M) \leq t \iff t M - X \succeq 0.$$
• **Sum of \( k \) largest eigenvalues.** For a symmetric \( m \times m \) matrix \( X \), let \( \lambda(X) \) be the vector of eigenvalues of \( X \) taken with their multiplicities in the non-ascending order:

\[
\lambda_1(X) \geq \lambda_2(X) \geq ... \geq \lambda_m(X),
\]

and let \( S_k(X) \) be the sum of \( k \) largest eigenvalues of \( X \):

\[
S_k(X) = \sum_{i=1}^{k} \lambda_i(X) \quad [1 \leq k \leq m]
\]

\[
[S_1(X) = \lambda_{\text{max}}(X); \ S_m(X) = \text{Tr}(X)]
\]

The functions \( S_k(X) \) are SDr:

\[
S_k(X) \leq t \iff \exists s, Z : \left\{ \begin{array}{l}
(a) \quad ks + \text{Tr}(Z) \leq t \\
(b) \quad Z \succeq 0 \\
(c) \quad X \preceq Z + sI_m
\end{array} \right.
\]

**Proof.** We should prove that

(i) If a pair \( X, t \) can be extended, by properly chosen \( s, Z \), to a solution of \( (a) \) \( - \) \( (c) \), then \( S_k(X) \leq t \);

(ii) If \( S_k(X) \leq t \), then the pair \( X, t \) can be extended by properly chosen \( s, Z \), to a solution of \( (a) \) \( - \) \( (c) \).
\[ S_k(X) \leq t \iff \exists s, Z : \begin{cases} 
(a) & ks + \text{Tr}(Z) \leq t \\
(b) & Z \succeq 0 \\
(c) & X \preceq Z + sI_m \end{cases} \]

“(i) If a pair \( X, t \) can be extended, by properly chosen \( s, Z \), to a solution of \( (a) - (c) \), then \( S_k(X) \leq t \)”

(i): We use the following

**Basic Fact:** The vector \( \lambda(X) \) is a \( \succeq \)-monotone function of \( X \in S^m \): \( X \succeq X' \Rightarrow \lambda(X) \geq \lambda(X') \).

Let \((X, t, s, Z)\) solve \( (a) - (c) \). Then

\[ X \preceq Z + sI_m \]  
\[ \Rightarrow \lambda(X) \leq \lambda(Z + sI_m) = \lambda(Z) + s \left[ \begin{array}{c} 1 \\ \vdots \\ 1 \end{array} \right] \]  
\[ \text{[by (c)]} \]

\[ \Rightarrow \quad S_k(X) \leq S_k(Z) + sk \]
\[ \Rightarrow \quad S_k(X) \leq \text{Tr}(Z) + sk \]  
\[ \text{[since } S_k(Z) \leq \text{Tr}(Z) \text{ due to } (b) \text{]} \]
\[ \Rightarrow \quad S_k(X) \leq t \]  
\[ \text{[by (a)]} \]

3.17
(ii): Let $S_k(X) \leq t$, and let $X = U \text{Diag}\{\lambda\} U^T$, $\lambda = \lambda(X)$, be the eigenvalue decomposition of $X$.

$$s = \lambda_k, \quad Z = U \begin{bmatrix} \lambda_1 - \lambda_k & \cdots & \lambda_{k-1} - \lambda_k \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \text{Diag}\{\lambda(Z)\} U^T,$$

we have

$$Z \succeq 0,$$

$$\text{Diag}\{\lambda(X)\} \leq \text{Diag}\left\{ \lambda(Z) + s \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right\} \Rightarrow X \preceq Z + sI_m,$$

$$t \geq S_k(X) = ks + \text{Tr}(Z),$$

so that $(t, X, s, Z)$ solves the system of LMIs

$$(a) \quad ks + \text{Tr}(Z) \leq t$$
$$(b) \quad Z \succeq 0$$
$$(c) \quad X \preceq Z + sI_m$$
Basic Fact: The vector $\lambda(X)$ is a $\succeq$-monotone function of $X \in S^m$: $X \succeq X' \Rightarrow \lambda(X) \geq \lambda(X')$.

This is an immediate corollary of the following

Variational Characterization of Eigenvalues: For an $m \times m$ symmetric matrix $A$, one has

$$\lambda_k(A) = \min_{E \in \mathcal{E}_k} \max_{e \in E: e^Te = 1} e^T A e,$$

where $\mathcal{E}_k$ is the collection of all linear subspaces of $\mathbb{R}^m$ of the dimension $m - k + 1$.

In particular,

$$\lambda_1(A) = \max_{e: e^Te = 1} e^T A e$$

$$\lambda_m(A) = \min_{e: e^Te = 1} e^T A e$$
VCE has a lot of important consequences, e.g., the following one:

**Eigenvalue Interlacement Theorem:** Let $A$ be a symmetric $m \times m$ matrix, and $\hat{A}$ be a $(m-k) \times (m-k)$ principal submatrix of $A$. Then

$$\lambda_i(A) \geq \lambda_i(\hat{A}) \geq \lambda_{i+k}(A).$$

**Proof of VCE.** Let $\lambda_k = \lambda_k(A)$, and let

$$\mu_k = \min_{E: \dim E = m-k+1} \max_{e \in E: e^r e = 1} e^T A e;$$

we should prove that $\mu_k = \lambda_k(A)$.

Both $\mu_k$ and $\lambda_k$ remain invariant when $A$ is replaced with $UAU^T$ with orthogonal $U$.

$\Rightarrow$ It suffices to consider the case of $A = \text{Diag}\{\lambda(A)\}$.

**$\lambda_k \geq \mu_k$:** Let $E = \{ x : x_1 = \ldots = x_{k-1} = 0 \}$. Then

$$\dim E = m-k+1 \Rightarrow \mu_k \leq \max_{e \in E: e^r e = 1} e^T A e = \max_{e_{k-1}^e, \ldots, e_m^e} \sum_{i=k}^m \lambda_i e_i^2 = \lambda_k.$$ 

**$\lambda_k \leq \mu_k$:** Let $F = \{ x : x_{k+1} = \ldots = x_m = 0 \}$, so that $\dim F = k$. For every subspace $E$ with $\dim E = m-k+1$, we have $\dim E + \dim F > m$, so that there exists a unit vector $f \in F \cap E$. We have

$$\max_{e \in E: e^r e = 1} e^T A e \geq f^T A f = \sum_{i=1}^k \lambda_i f_i^2 \geq \lambda_k \sum_{i=1}^k f_i^2 = \lambda_k.$$ 

Thus, $\mu_k \equiv \min_{E: \dim E = m-k+1} \max_{e \in E: e^r e = 1} e^T A e \geq \lambda_k$. 

3.20
To proceed, we need the following

**Birkhoff Theorem:** Let $P_m$ be the set of double-stochastic $m \times m$ matrices, that is, matrices $[p_{ij}]_{i,j=1}^m$ such that

$$p_{ij} \geq 0; \quad \sum_i p_{ij} = 1 \forall j; \quad \sum_j p_{ij} = 1 \forall i.$$

The vertices of the polytope $P_m$ are exactly the permutation matrices, so that every double stochastic matrix is a convex combination of permutation matrices.

**Sketch of the proof:** The only nontrivial claim is that an extreme point $p$ of $P_m$ is a Boolean (≡ with entries 0/1) matrix.

$P_m$ is cut off $\mathbb{R}^{m^2}$ by $m^2$ inequalities $p_{ij} \geq 0$ and $2m - 1$ linearly independent linear equalities ("if all row sums and all but one column sums in a square matrix are equal to 1, than all row and column sums are equal to 1").

⇒ extreme point $p$ should make $m^2 - (2m - 1)$ of the bounds $p_{ij} \geq 0$ active
⇒ there is a column in $p$ with at most one nonzero
⇒ $p$ has an entry equal to 1, and all remaining entries in the row and the column of this entry are zeros.
Eliminating from $p$ the row and the column of an entry equal to 1, we get a (clearly extreme) point of $P_{m-1}$
⇒ The claim can be proved by induction in $m$. 3.21
Corollary. Let \( f(x) \) be a symmetric w.r.t. permutation of coordinates convex function on \( \mathbb{R}^m \), and let \( \pi \) be a double-stochastic \( m \times m \) matrix. Then

\[
f(\pi x) \leq f(x) \quad \forall x \in \mathbb{R}^m.
\]

Proof. By Birkhoff Theorem, \( \pi x \) is a convex combination of permutations \( x^i \) of \( x \). Therefore, by Jensen’s Inequality, \( f(\pi \lambda) \) is not greater than \( \max_i f(x^i) \), and this is exactly \( f(x) \) due to the symmetry of \( f \).
**Corollary of Corollary:** Let $f(x)$ be a symmetric convex function on $\mathbb{R}^m$. Then the function

$$F(X) = f(\lambda(X))$$

is convex on $S^m$, and, moreover,

$$F(X) = \max_{U:U^TU=I} f(Dg(UXU^T)). \tag{*}$$

**Proof:** It suffices to verify (*); indeed, given (*), $F(\cdot)$ is convex as the upper bound, w.r.t. orthogonal $U$, of the family of (clearly convex) functions $f_U(\cdot)$. For properly chosen orthogonal $U$ we have

$$UXU^T = \text{Diag}\{\lambda(X)\} \Rightarrow \max_{U:U^TU=I} f(Dg(UXU^T)) \geq f(\lambda(X)).$$

To prove the opposite inequality, observe that every matrix of the form $UXU^T$ with orthogonal $U$ is of the form $V\text{Diag}\{\lambda(X)\}V^T$ with orthogonal $V$ as well. Now,

$$[Dg(UXU^T)]_i = [V\text{Diag}\{\lambda(X)\}V^T]_{ii} = \sum_j V_{ij}^2 \lambda_j(X),$$

that is, $Dg(UXU^T) = \pi \lambda(X)$ for the double stochastic matrix $\pi = [V_{ij}^2]_{i,j}$. Therefore

$$f(Dg(UXU^T)) = f(\pi \lambda(X)) \leq f(\lambda(X)).$$
Corollary of Corollary of Corollary: Let $f$ be a convex symmetric function on $\mathbb{R}^m$. Then

$$f(Dg(X)) \leq f(\lambda(X))$$

for every symmetric matrix $X$.

For example, for every symmetric matrix $X$ with the vector of eigenvalues $\lambda$ one has

- The sum of $k$ largest diagonal entries of $X$ does not exceed $S_k(X) = \lambda_1 + \ldots + \lambda_k$.

  $$ [f(x) = \max_{i_1 < i_2 < \ldots < i_k} [f(x_{i_1}) + \ldots + f(x_{i_k})] ]$$ is the sum of $k$ largest entries in $x$.

- The sum of $k$ smallest diagonal entries in $X$ is at least the sum of $k$ smallest of $\lambda_i$'s.

- If $X \succ 0$, then the product of the $k$ smallest diagonal entries in $X$ is at least the product of the $k$ smallest of $\lambda_i$'s. In particular, the product of all diagonal entries in $X$ is $\geq \text{Det}(X)$.

  $$ [g(x) = \min_{i_1 < i_2 < \ldots < i_k} [\ln x_{i_1} + \ldots + \ln x_{i_k}] ]$$ is the sum of logs of $k$ smallest entries in $x > 0$, $f(x) = -g(x)$.
For $z \in \mathbb{R}^m$, let $s_k(z)$ be the sum of $k$ largest entries in $z$.

- **Majorization Principle:** Let $x \in \mathbb{R}^m$. A point $y$ can be represented as $\pi x$ with a double stochastic matrix $\pi$ if and only if

$$s_k(y) \leq s_k(x), \; k < m, \; \text{and} \; s_m(y) = s_m(x)$$

**Corollary:** Let $f(x)$ be a SDr symmetric function on $\mathbb{R}^m$. Then the function

$$F(X) = f(\lambda(X)) : S^m \to \mathbb{R} \cup \{+\infty\}$$

is SDr. In particular, the following functions are SDr with explicit SDR’s:

- $-\text{Det}^\pi(X), \; X \in S^m_+ \; (\pi \in (0, \frac{1}{m}] \; \text{is rational});$
- $\text{Det}^{-\pi}(X), \; X \succ 0 \; (\pi > 0 \; \text{is rational});$
- $|X|_\pi = \|\lambda(X)\|_\pi, \; X \in S^m \; (\pi \in [1, \infty) \; \text{is rational or} \; \pi = \infty).$
Proof. Let \( t \geq f(x) \Leftrightarrow \exists u : A(t, x, u) \succeq 0 \). Then

\[
\begin{align*}
\text{Let } t \geq f(x) & \iff \exists u : A(t, x, u) \succeq 0. \\
& \iff \exists u : A(t, x, u) \succeq 0.
\end{align*}
\]

\[
\begin{align*}
t \geq F(X) & \iff \exists (y \in \mathbb{R}^m, \pi \in P_m) : \\
& \begin{cases}
y_1 \geq y_2 \geq \ldots \geq y_m \\
f(y) \leq t \\
\lambda(X) = \pi y
\end{cases}
\end{align*}
\]

\[
\Rightarrow t \geq F(X) \iff \exists y \in \mathbb{R}^m : \\
\begin{cases}
y_1 \geq y_2 \geq \ldots \geq y_m, f(y) \leq t \\
s_k(\lambda(X)) \leq y_1 + \ldots + y_k, k < m \\
s_m(\lambda(X)) = y_1 + \ldots + y_m
\end{cases}
\]

\[
\Rightarrow t \geq F(X) \iff \exists (y \in \mathbb{R}^m, u) : \\
\begin{cases}
y_1 \geq y_2 \geq \ldots \geq y_m, A(y, t, u) \succeq 0 \\
S_k(X) \leq y_1 + \ldots + y_k, k < m \\
SD\text{-representable!} \\
Tr(X) = y_1 + \ldots + y_m
\end{cases}
\]
Majorization Principle: Let \( x \in \mathbb{R}^n \). A point \( y \) can be represented as \( \pi x \) with a double stochastic matrix \( \pi \) if and only if

\[
s_k(y) \leq s_k(x), \quad k < m, \text{ and } s_m(y) = s_m(x) \quad (\ast)
\]

Proof, “only if” part: If \( y = \pi x \) with double stochastic \( \pi \), then \( s_k(y) \leq s_k(x) \) by Corollary of the Birkhoff Theorem (\( s_k(\cdot) \) are convex symmetric functions!), and of course \( s_m(y) = s_m(x) \).
Proof, “if” part: Let \( x \) and \( y \) satisfy (\( \ast \)); we should prove that \( y = \pi x \) for a double stochastic matrix \( \pi \). By “permutational symmetry” of the statement, we may assume that
\[
x_1 \geq x_2 \geq \ldots \geq x_m, \quad y_1 \geq y_2 \geq \ldots \geq y_m.
\]

Let \( X \) be the set of all permutations of \( x \); by Birkhoff Theorem, \( y = \pi x \) for certain double stochastic \( \pi \) iff \( y \in \text{Conv}(X) \), thus all we should prove is that \( y \in \text{Conv}(X) \). Assume that \( y \notin \text{Conv}(X) \). Then there exists \( e \) such that
\[
e^T y > \max_{x' \in X} e^T x'. \tag{**}
\]

Permuting the entries in \( e \), we do not vary the right hand side in (\( ** \)). If \( e_i < e_j \) for a pair \( i, j \) with \( i > j \), then, swapping \( e_i \) and \( e_j \), we do not decrease \( e^T y \) (since \( y_1 \geq y_2 \geq \ldots \geq y_m \)). Thus, we may assume that \( e \) in (\( \ast \)) satisfies \( e_1 \geq e_2 \geq \ldots \geq e_m \). Then
\[
e^T y = e_1 y_1 + e_2 y_2 + \ldots + e_m y_m
\]
\[
= e_m (y_1 + \ldots + y_m) + (e_{m-1} - e_m) (y_1 + \ldots + y_{m-1})
\]
\[
+ (e_{m-2} - e_{m-1}) (y_1 + \ldots + y_{m-2}) + \ldots + (e_1 - e_2) y_1
\]
\[
= e_m s_m(y) + \underbrace{(e_{m-1} - e_m)}_{\geq 0} s_{m-1}(y)
\]
\[
+ \underbrace{(e_{m-2} - e_{m-1})}_{\geq 0} s_{m-2}(y) + \ldots + \underbrace{(e_1 - e_2)}_{\geq 0} s_1(y)
\]
\[
\leq e_m s_m(x) + (e_{m-1} - e_m) s_{m-1}(x)
\]
\[
+ (e_{m-2} - e_{m-1}) s_{m-2}(x) + \ldots + (e_1 - e_2) s_1(x) \quad [\text{by } (\ast)]
\]
\[
= e^T x - \text{contradicts } (**)!
\]
- **Norm of rectangular matrix.** Let $X$ be a $m \times n$ matrix. Its spectral norm

\[
\|X\| = \max_{\|\xi\|_2 \leq 1} \|X\xi\|_2
\]

is SDr:

\[
t \geq \|X\| \iff \begin{bmatrix} tI_n & X^T \\ X & tI_m \end{bmatrix} \succeq 0.
\]

More generally, let

\[
\sigma_i(X) = \sqrt{\lambda_i(X^TX)}
\]

be the singular values of a rectangular matrix $X$. Then

- **The sum of $k$ largest singular values** $\Sigma_k(X) = \sum_{i=1}^k \sigma_i(X)$ is a SDr function of $X \in \mathbb{R}^{m \times n}$. 

3.29
The sum of $k$ largest singular values $\Sigma_k(X) = \sum_{i=1}^{k} \sigma_i(X)$ is a SDr function of $X \in \mathbb{R}^{m \times n}$. Indeed, it is easily seen that the eigenvalues of linearly depending on $X$ symmetric matrix

$$A(X) = \begin{bmatrix} X \\ X^T \end{bmatrix}$$

are singular values of $X$, minus singular values of $X$, and perhaps a number of zeros. As a result,

$$\Sigma_k(X) = S_{\bar{k}}(A(X))$$

with properly selected $\bar{k}$.
• **SDr of symmetric monotone functions of singular values.** Let $f(\lambda) : \mathbb{R}_+^n \to \mathbb{R} \cup \{\infty\}$ be a symmetric w.r.t. permutations of coordinates and $\geq$-nondecreasing SDr function. Then the function

$$F(X) = f(\sigma(X)) : \mathbb{R}^{m \times n} \to \mathbb{R} \cup \{\infty\}$$

is SDr.

In particular, the functions

$$|X|_{\pi} = \|\sigma(X)\|_{\pi}$$

with rational $\pi \in [1, \infty)$ are SDr with explicit SDR’s.
“$\succeq$-convex quadratic matrix function”

\[
F(X) = (AXB)(AXB)^T + CXD + (CXD)^T + E
\]

\[
F : \mathbb{R}^{p \times q} \rightarrow \mathbb{S}^m
\]

($A, B, C, D, E = E^T$ are constant matrices such that $F(\cdot)$ makes sense and takes its values in $\mathbb{S}^m$) is SDr in the sense that its “$\succeq$graph”

\[
\text{Epi}\{F\} = \{(X, Y) \in \mathbb{R}^{p \times q} \times \mathbb{S}^m : F(X) \preceq Y\}
\]

is an SDr set:

\[
Y \succeq F(X)
\]

$\updownarrow$ [LSC]

\[
\begin{bmatrix}
Y - E - CXD - (CXD)^T & AXB \\
(AXB)^T & I_r
\end{bmatrix} \succeq 0 \quad [B : q \times r]
\]

(by the Schur Complement Lemma).
“\(\succeq\)-convex fractional-quadratic function”. Let \(X\) be a rectangular \(p \times q\) matrix, and \(V\) be a positive definite symmetric \(q \times q\) matrix. Consider the matrix-valued function

\[
F(X, V) = XV^{-1}X^T : \mathbb{R}^{p \times q} \times \text{int} \mathbb{S}^q_+ \rightarrow \mathbb{S}^p
\]

The closure of the “\(\succeq\)graph” of \(F(X, V)\) – the set

\[
\mathcal{G} \equiv \text{cl}\left\{ (X, V, Y) \in \mathbb{R}^{p \times q} \times \text{int} \mathbb{S}^q_+ \times \mathbb{S}^p : F(X, V) \preceq Y \right\}
\]

is SDr:

\[
\mathcal{G} = \left\{ (X, V, Y) \in \mathbb{R}^{p \times q} \times \mathbb{S}^q \times \mathbb{S}^p \mid \begin{bmatrix} Y & X \\ X^T & V \end{bmatrix} \succeq 0 \right\}.
\]

(by the Schur Complement Lemma).
• “⪯-hypograph of the matrix square root. The sets

\[(X, Y) \in S^+_m \times S^+_m : X^2 \preceq Y \} = \{(X, Y) : X \succeq 0, \begin{bmatrix} Y & X \\ X & I \end{bmatrix} \succeq 0\}\]

and

\[(X, Y) \in S^+_m \times S^+_m : X \preceq Y^{1/2} \} = \{(X, Y) : \exists Z : 0 \preceq X \preceq Z, \begin{bmatrix} Y & Z \\ Z & I \end{bmatrix} \succeq 0\}\]

both are SDr. These sets are different:

\[0 \preceq X, X^2 \preceq Y \Rightarrow X \preceq Y^{1/2}, \text{ but } 0 \preceq X \preceq Y^{1/2} \nRightarrow X^2 \preceq Y \]

\[\begin{bmatrix} 6 & 0 \\ 0 & 1 \end{bmatrix} \preceq \begin{bmatrix} 12 & 8 \\ 8 & 12 \end{bmatrix}, \text{ but } \text{Det} \left( \begin{bmatrix} 172 & 192 \\ 192 & 207 \end{bmatrix} \right) = -1260 < 0! \]
Sums-of-Squares

**Situation:** We are given real-valued functions $\phi_0(x) \equiv 1, \phi_1(x), ..., \phi_d(x)$ on some set $X$. These data specify the linear space $\Phi$ of functions $\phi(\cdot)$ which can be represented as linear combinations of $\phi_i(\cdot)$ and their pairwise products, or, which is the same due to $\phi_0(\cdot) \equiv 1$, as linear combinations of their pairwise products:

$$\Phi = \{ f(\cdot) = \sum_{i,j=0}^{d} c_{ij} \phi_i(\cdot) \phi_j(\cdot) \}$$

W.l.o.g. we can assume that $c_{ij} = c_{ji}$. Note that $\Phi$ is the image of $S^{d+1}$ under the linear mapping

$$S^{d+1} \ni C = [c_{ij}]_{0 \leq i,j \leq d} \mapsto A(C)(\cdot) = \sum_{i,j} c_{ij} \phi_i(\cdot) \phi_j(\cdot)$$
\[ S^{d+1} \ni C = [c_{ij}]_{0 \leq i, j \leq d} \mapsto A(C)(\cdot) = \sum_{i,j} c_{ij} \phi_i(\cdot) \phi_j(\cdot) \] & \Phi = A(S^{d+1})

Observation: Sums of squares of linear combinations of functions \( \phi_0, ..., \phi_d \) are exactly the elements of the image of the positive semidefinite cone \( S^{d+1}_+ \) under the mapping \( A \).

Indeed, \([\sum_i \lambda_i \phi_i(\cdot)]^2 = A(\lambda \lambda^T)\), and the matrices from \( S^{d+1}_+ \) are nothing but sums of dyadic matrices.

Corollary: The set of (arrays of coefficients of) polynomials which are sums of squares of linear combinations of given polynomials \( \phi_0, ..., \phi_d \) on \( \mathbb{R}^n \) is SDR.

Indeed, this set is the image of \( S^{d+1}_+ \) under linear mapping \( A(\cdot) \).

Conclusion: A sufficient condition for a function \( f \in \Phi \) to be nonnegative is the possibility to find a \( C \in S^{d+1}_+ \) such that

\[ A[C] = f \] & \( C \succeq 0. \quad (!) \)

When \( X = \mathbb{R}^n \) and all \( \phi_i \) are polynomials, \(!\) is a semidefinite feasibility problem.
Nonnegative polynomials

For every positive integer $k$, the following sets are SDr:

— The set $P_{2k}^{+}(\mathbb{R})$ of coefficients of algebraic polynomials of degree $\leq 2k$ which are nonnegative on the entire axis:

$$P_{2k}^{+} = \left\{ p = (p_0, \ldots, p_{2k})^T : \exists Q = [Q_{ij}]_{i,j=0}^{k} \in S_{+}^{k+1} : p_\ell = \sum_{i+j=\ell} Q_{ij}, \ell = 0, 1, \ldots, 2k \right\}$$

Equivalently: A polynomial $p(t)$ of degree $\leq 2k$ is nonnegative on $\mathbb{R}$ iff it can be obtained from $Q \in S_{+}^{k+1}$ according to

$$p(t) = [1; t; t^2; \ldots; t^{k}]^T Q[1; t; t^2; \ldots; t^{k}]$$

— The set $P_k^{+}(\mathbb{R}_+)$ of coefficients of algebraic polynomials of degree $\leq k$ which are nonnegative on the nonnegative ray $\mathbb{R}_+$

— The set $P_k^{+}([0, 1])$ of coefficients of algebraic polynomials of degree $\leq k$ which are nonnegative on the segment $[0, 1]$

— The set $T_k^{+}(\Delta)$ of coefficients of trigonometric polynomials of degree $\leq k$ which are nonnegative on a given segment $\Delta \in [-\pi, \pi]$. 

3.37
As a corollary, for every segment $\Delta \subset \mathbb{R}$ and every positive integer $k$, the function

$$f(p) = \max_{t \in \Delta} p(t)$$

of the vector $p$ of coefficients of an algebraic (or a trigonometric) polynomial $p(\cdot)$ of degree $\leq k$ is SDr.

Indeed, $\tau \geq f(p)$ is and only if the polynomial $q_{p,\tau}(t) = \tau - p(t)$ of $t$ is non-negative on $\Delta$, and the coefficients of $q$ are affine in $\tau$ and the coefficients of $p$. 
• **SDR of the cone** $P_{2k}^+(\mathbb{R})$: Consider the linear mapping $\Pi$ from the space $S_k^{k+1}$ to the space of polynomials of degree $\leq 2k$:

$$
\Pi([a_{ij}]_{i,j=0}^k) = \sum_{i,j=0}^k a_{ij} t^{i+j}.
$$

**Observation:** The images of dyadic matrices $aa^T$ under the mapping $\Pi$ are exactly squares of polynomials of degree $\leq k$:

$$
\Pi(aa^T) = \sum_{i,j=0}^k a_i a_j t^{i+j} = \left(\sum_{i=0}^k a_i t^i\right)^2.
$$

• The positive semidefinite cone is exactly the set of sums of dyadic matrices. Therefore, by Observation, the image of positive semidefinite cone under the mapping $\Pi$ is exactly the set of polynomials of degree $\leq 2k$ which are *sums of squares*. It remains to note that A univariate polynomial is nonnegative on the entire axis *iff* it is sum of squares, whence

$$
P_{2k}^+(\mathbb{R}) = \Pi(S_k^{k+1}),
$$

and thus $P_{2k}^+$ is SDr.
• SDR of $P_{2k}^+(\mathbb{R})$ induces all other SDRs we need, namely

— SDR of $P_k^+(\mathbb{R}_+)$ due to

$$p(t) \in P_k^+(\mathbb{R}_+) \Leftrightarrow \pi[p](t) \equiv p(t^2) \in P_{2k}^+(\mathbb{R}),$$

— SDR of $P_k^+([0, 1])$ due to

$$p(t) \in P_k^+([0, 1]) \Leftrightarrow \psi[p](t) \equiv (1 + t^2)^k p \left(\frac{t^2}{1 + t^2}\right) \in P_{2k}^+(\mathbb{R})$$

— SDR of $T_k(\Delta)$ due to

$$p(\phi) \in T_k(\Delta) \Leftrightarrow \theta[p](t) \equiv (1 + t^2)^k p(2 \text{atan}(t)) \in P_{2k}^+(\tilde{\Delta})$$

and the coefficients of $\pi[p]$, $\psi[p]$, $\theta[p]$ are affine in the coefficients of $p$. 

3.39
• Why a nonnegative on the axis polynomial is a sum of squares?
Assume a polynomial

\[ p(t) = a(t - s_1)...(t - s_n) \]

of certain degree \( n \) is nonnegative on the entire axis. Then
• the degree is even,
• the leading coefficient \( a \) is positive,
• all real roots, if any, are of even multiplicities.
If \( z, z^* \) is a conjugate pair of complex roots, then the corresponding factor
\((t - z)(t - z^*)\) in \( p \) is a sum of squares of a linear function and a real.
Thus, \( p \) is the product of sums of squares of polynomials, and such a product again is a sum of squares of polynomials.
• In fact, our reasoning says that \( p \) is a product of factors which are sums of at most two squares each. As a result, \( p \) itself is a sum of just two squares, due to the identity

\[ (a^2 + b^2)(c^2 + d^2) = (ac - bd)^2 + (ad + bc)^2. \]
A. Dynamic Stability in Mechanics. The “free” (when no external forces are applied) motions of linearly elastic mechanical systems (buildings, bridges, masts, etc.) are governed by the Newton Law in the form:

\[ M \frac{d^2}{dt^2} x(t) = -Ax(t) \]  

(NL)

where

- \( x(t) \) is the state of the system at time \( t \);
- \( M \succ 0 \) is the mass matrix;
- \( A \succeq 0 \) is the stiffness matrix; \( \frac{1}{2}x^T A x \) is the potential energy of the system at state \( x \).
- It is easily seen that every solution to (NL) is linear combination of basic harmonic oscillations (“modes”)

\[ \cos(\omega_\ell t) \vec{f}_\ell, \sin(\omega_\ell t) \vec{f}_\ell \]

where the eigenfrequencies \( \omega_\ell \) are square roots of the eigenvalues \( \lambda(A : M) \) of the matrix pencil \([M, A]\), and \( f_\ell \) are eigenvectors of the pencil.
\[ \omega = 1.274 \quad \omega = 0.957 \quad \omega = 0.699 \]

“Nontrivial” modes of a spring triangle (3 unit masses linked by springs)
There are 3 modes more with \( \omega = 0 \) (coming from shifts and rotation)

- A typical Dynamic Stability specification is a lower bound on the eigen-frequencies:

\[ \lambda_{\min}(A : M) \geq \lambda_*, \]

which is the matrix inequality

\[ A \succeq \lambda_* M. \]  \hspace{1cm} (S)

- When \( A \) and \( M \) are affine in the design variables, (S) is an LMI!
B. Structural Design. Consider a linearly elastic mechanical system $S$ with stiffness matrix $A \succ 0$ loaded by an external load $f$. Under the load, the system deforms until the tensions caused by the deformation compensate the external forces. The corresponding equilibrium displacement $x_f$ solves the equilibrium equation

$$Ax = f [\Rightarrow x_f = A^{-1}f]$$

The compliance of $S$ w.r.t. load $f$ is the potential energy

$$\text{Compl}_f = \frac{1}{2} x_f^T Ax_f = \frac{1}{2} f^T A^{-1} f$$

stored in the system in the corresponding equilibrium. The compliance quantifies the “rigidity” of $S$ w.r.t. $f$: the less is the compliance, the better $S$ withstands the load.
In a typical Structural Design problem, we are given
- a stiffness matrix $A = A(t)$ affinely depending on a vector $t$ of design parameters,
- a collection $f_1, ..., f_k$ of “loading scenarios”,
- a set $\mathcal{T}$ of allowed values of $t$

and are seeking for the design $t \in \mathcal{T}$ which results in the smallest possible worst-case, w.r.t. the scenarios, compliance, thus arriving at the optimization problem

$$
\min_{t \in \mathcal{T}} \max_{\ell=1,\ldots,k} \frac{1}{2} f_\ell^T A^{-1}(t) f_\ell.
$$
\[
\min_{t \in T} \max_{\ell=1,\ldots,k} \frac{1}{2} f_\ell^T A^{-1}(t) f_\ell.
\]  

(\text{SD})

- When $T$ is SDr, problem (SD) becomes the semidefinite program

\[
\min_{t,\tau} \left\{ \tau : \begin{bmatrix} 2\tau & f_\ell^T \\ f_\ell & A(t) \end{bmatrix} \succeq 0, \quad \ell = 1,\ldots,k, \quad t \in T \right\}
\]

Data for Bridge Design problem [12 nodes, 51 tentative bars, 4-force load]

Optimal bridge (29 bars)  
Equilibrium displacement
C. Boyd's Time Constant of an RC circuit. Consider a circuit comprised of (a) resistors, (b) capacitors, and (c) resistors in serial connection with outer voltages:

A simple circuit

Element OA: outer supply of voltage $V_{OA}$ and resistor with conductance $\sigma_{OA}$
Element AO: capacitor with capacitance $C_{AO}$
Element AB: resistor with conductance $\sigma_{AB}$
Element BO: capacitor with capacitance $C_{BO}$

A chip is a complicated RC circuit where the outer voltages are switching, at certain frequency, between several constant values. In order for chip to work reliably, the time of transition to the steady-state corresponding to given outer voltages should be much less than the time between switches of the voltages. How to model this crucial requirement?
In an RC circuit, the transition period is governed by the Kirchhoff laws which result in the equation

\[ C \dot{w} = -Rw \]  

(H)

where

- \( w \) is the difference between the current state of the circuit and its steady state;
- \( C \succ 0 \) is given by circuit's topology and the capacitances of the capacitors and is affine in the capacitances;
- \( R \succ 0 \) is given by circuit's topology and the conductances of the resistors and is affine in the conductances.

The space of solutions to (H) is spanned by functions

\[ w_\ell(t) = \exp\{-\lambda_\ell t\} f_\ell, \]

where \( \lambda_\ell \) are the eigenvalues of the matrix pencil \([C, R]\).

- \( \lambda_{\min}(R : C) \) can be viewed as the "decay rate" for (H): the "duration" of the transition period is of order of \( \lambda_{\min}^{-1}(R : C) \).

S. Boyd has proposed to use \( \lambda_{\min}^{-1}(R : C) \) as a "time constant" for an RC circuit and to model a lower bound on the speed of the circuit (\( \equiv \) an upper bound on the duration of the transition period) as a lower bound on \( \lambda_{\min}(R : C) \), i.e., as the matrix inequality

\[ R \succeq \lambda_* C. \]  

(B)
When $R$ and $C$ are affine in the design variables, (B) becomes an LMI, which allows to pose numerous circuit design problems with bounds on the speed as SDPs.

\[ R \geq \lambda^* C. \hspace{1cm} (B) \]
Lyapunov Stability Analysis. Consider an uncertain time varying linear dynamical system

\[ \dot{x}(t) = A(t)x(t) \quad \text{(ULS)} \]

where

- \( x(t) \in \mathbb{R}^n \) is the state vector at time \( t \)
- \( A(t) \) takes values in a given uncertainty set \( \mathcal{U} \subset \mathbb{R}^{n \times n} \)
- (ULS) is called stable, if all trajectories of the system converge to 0 as \( t \to \infty \):

\[ A(t) \in \mathcal{U} \forall t \geq 0, \quad \dot{x}(t) = A(t)x(t) \Rightarrow \lim_{t \to \infty} x(t) = 0. \]

How to certify stability?

- Standard sufficient stability condition is the existence of Lyapunov Stability Certificate – a matrix \( X \succ 0 \) such that the function \( L(x) = x^T X x \) decreases exponentially along the trajectories:

\[ \exists \alpha > 0 : \frac{d}{dt} L(x(t)) \leq -\alpha L(x(t)) \text{ for all trajectories} \]

\[ \Rightarrow L(x(t)) \leq \exp\{-\alpha t\} L(x(0)) \Rightarrow x(t) \to 0, \quad t \to \infty \]

For a time-invariant system, this condition is necessary and sufficient for stability.

3.49
Question: When $\alpha > 0$ is such that
\[
\frac{d}{dt} L(x(t)) \leq -\alpha L(x(t))
\] for all trajectories $\dot{x}(t) = A(t)x(t)$, $A(t) \in U$?

Answer:
\[
\frac{d}{dt} \left( x^T(t) X x(t) \right) = (\dot{x}(t))^T X x(t) + x^T(t) X \dot{x}(t)
\]
\[
= x^T(t) A^T(t) X x(t) + x^T(t) X A x(t)
\]
\[
= x^T(t) \left[ A^T(t) X + X A(t) \right] x(t)
\]

Thus,
\[
\frac{d}{dt} L(x(t)) \leq -\alpha L(x(t))
\] for all trajectories
\[
\Leftrightarrow x^T(t) \left[ A^T(t) X + X A(t) \right] x(t) \leq -\alpha x^T(t) X x(t)
\] for all trajectories
\[
\Leftrightarrow A^T X + X A \preceq -\alpha X \quad \forall A \in U
\]

Thus,
\[
\exists (\alpha > 0, X \succ 0) : \frac{d}{dt} \left( x^T(t) X x(t) \right) \leq -\alpha \left( x^T(t) X x(t) \right)
\] for all trajectories
\[
\Leftrightarrow \exists (\alpha > 0, X \succ 0) : A^T X + X A \preceq -\alpha X \quad \forall A \in U
\]
\[
\Leftrightarrow \exists X : X \succeq I, A^T X + X A \preceq -I \quad \forall A \in U
\]
• The existence of a Lyapunov Stability Certificate is equivalent to solvability of the *semi-infinite* system of LMIs in matrix variable $X$:

$$X \succeq I; \quad A^T X + X A \preceq -I \quad \forall (A \in U)$$

(L)

• Every solution to (L) is a Lyapunov Stability Certificate for the uncertain dynamical system

$$\dot{x}(t) = A(t)x(t) \quad [A(t) \in U \forall t]$$

• In some cases, the semi-infinite system of LMIs is equivalent to a usual system of LMIs, so that search for a Lyapunov Stability Certificate reduces to solving an SDP.

**Example 1: Polytopic uncertainty**

$$U = \text{Conv}\{A_1, ..., A_L\}.$$  

In this case (L) clearly is equivalent to the finite system of LMIs

$$X \succeq I; \quad A_\ell^T X + X A_\ell \preceq -I, \ \ell = 1, ..., L.$$
• **Example 2: Norm-bounded uncertainty**

\[ U = \left\{ A = A_0 + P \Delta Q : \Delta \in \mathbb{R}^{p \times q}, \|\Delta\| \leq 1 \right\} \]  

(NB)

• **Example:** Consider a controlled linear time-invariant dynamical system

\[
\begin{align*}
\dot{x}(t) &= Ax(t) + Bu(t) \\
y(t) &= Cx(t)
\end{align*}
\]

• \( x \): state • \( u \): control • \( y \): observed output

“closed” by a feedback

\[ u(t) = Ky(t). \]

Open loop (left) and closed loop (right) systems
\[ \mathcal{U} = \left\{ A = A_0 + P\Delta Q : \Delta \in \mathbb{R}^{p \times q}, \|\Delta\| \leq 1 \right\} \]  

(NB)

The resulting closed loop system is given by

\[ \dot{x}(t) = \tilde{A}x(t), \quad \tilde{A} = A + BKC \]  

(1)

Assuming that \( A, B, C \) are certain, and feedback matrix \( K \) is drifting around nominal feedback \( K^* \):

\[ K = K^* + \Delta, \]

where \( \|\Delta\| \) does not exceed a given level, \( \tilde{A} \) runs through uncertainty set of the form of (NB).
\[ U = \left\{ A = A_0 + P \Delta Q : \Delta \in \mathbb{R}^{p \times q}, \| \Delta \| \leq 1 \right\} \]  \hspace{1cm} \text{(NB)}

**Proposition.** With the uncertainty set (NB), the Lyapunov Stability Certificate semi-infinite system of LMIs

\[
X \succeq I; \quad A^T X + X A \preceq -I \quad \forall (A \in U) \quad \text{(L)}
\]

is equivalent to the LMIs

\[
X \succeq I, \quad \begin{bmatrix}
-I - A_0^T X - X A_0 - \lambda Q^T Q & -X P \\
-P^T X & \lambda I
\end{bmatrix} \succeq 0
\]

in variables \( X, \lambda \).
An instrumental role in the proof of Proposition is played by the following statement which is extremely useful by its own right:

**S-Lemma:** Consider a homogeneous quadratic inequality

\[ x^T Ax \geq 0 \]  \hspace{1cm} (A)

which is strictly feasible: \( \bar{x}^T A \bar{x} > 0 \) for certain \( \bar{x} \).

A homogeneous quadratic inequality

\[ x^T Bx \geq 0 \]  \hspace{1cm} (B)

is a consequence of (A) iff it is a “linear” consequence of (A), i.e., iff (B) can be obtained by summing up a nonnegative multiple of (A) and identically true homogeneous quadratic inequality, or, which is the same, iff

\[ \exists (\lambda \geq 0) : \quad B \geq \lambda A. \]
Proof of Proposition is given by the following fact:

(!) Assume that $E \neq 0$. Then

$$C + D^T \Delta E + E^T \Delta^T D \succeq 0 \quad \forall (\Delta, \|\Delta\| \leq 1)$$

$$\Leftrightarrow \exists \lambda : \begin{bmatrix} C - \lambda E^T E & D^T \\ D & \lambda I \end{bmatrix} \succeq 0$$

In particular, when $Q \neq 0$, one has

$$=-[I + A_0^T X - X A_0] + [-P^T X]^T \Delta Q + Q^T \Delta [-P^T X]
\underbrace{-I - [A_0 + P \Delta Q]^T X - X [A_0 + P \Delta Q]}_{-I - [A_0 + P \Delta Q]^T X - X [A_0 + P \Delta Q]}
\succeq 0 \forall (\Delta, \|\Delta\| \leq 1)$$

$$\Leftrightarrow \exists \lambda : \begin{bmatrix} -I - A_0^T X - X A_0 - \lambda Q^T Q & -XP \\ -P^T X & \lambda I \end{bmatrix} \succeq 0$$

Proof of (!):

$$C + D^T \Delta E + E^T \Delta^T D \succeq 0 \quad \forall (\Delta, \|\Delta\| \leq 1)$$

$$\Leftrightarrow \xi^T C \xi + 2 \xi^T D^T \underbrace{[\Delta E \xi]}_{\eta} \succeq 0 \quad \forall \xi \forall (\Delta, \|\Delta\| \leq 1)$$

$$\Leftrightarrow \xi^T C \xi + 2 \xi^T D^T \eta \succeq 0 \quad \forall \xi \forall (\eta, \|\eta\| \leq \|E \xi\|)$$

$$\Leftrightarrow \xi^T C \xi + 2 \xi^T D^T \eta \succeq 0 \quad \forall (\xi, \eta : \xi^T E^T E \xi - \eta^T \eta \succeq 0)$$

$$\Leftrightarrow \exists \lambda \geq 0 : \begin{bmatrix} C & D \\ D^T & -I \end{bmatrix} \succeq \lambda \begin{bmatrix} E^T E \\ -I \end{bmatrix}$$

[S-Lemma]
SDP approximations of computationally intractable problems

A. SDP relaxations in Combinatorics. In a typical combinatorial problem, we are interested to minimize a “simple” function over a discrete set, e.g.

- **Shortest Path**: Given a graph with arcs assigned nonnegative integer lengths and two nodes $a, b$, find the shortest path from $a$ to $b$ or detect that no path exists.

- **Integer Linear Programming**:
  \[
  \min_x \{ c^T x : Ax \leq b, x \in \mathbb{Z}^n \}
  \]
  \[\text{[} \mathbb{Z}^n : \text{n-dimensional integral vectors}\text{]}\]
  (all entries in $A, b, c$ are integral)

- **Boolean Programming**:
  \[
  \min_x \{ c^T x : Ax \leq b, x \in \mathbb{B}^n \}
  \]
  \[\text{[} \mathbb{B}^n : \text{n-dimensional 0-1 vectors}\text{]}\]
  (all entries in $A, b, c$ are integral)

3.57
• **Knapsack problem:**

\[
\max_x \left\{ \sum_{i=1}^{n} c_i x_i : \sum_{i=1}^{n} a_i x_i \leq b, \ x_i \in \{0; 1\} \right\}
\]

\((c_i, a_i, b \text{ are positive integers})\)

• **“Stones”:** Given \(n\) stones of positive integer weights \(a_1, ..., a_n\), check whether you can partition them into two groups of equal weight, i.e., check whether the linear equation

\[
\sum_{i=1}^{n} a_i x_i = 0
\]

has a solution with \(x_i = \pm 1\).
As far as solution methods are concerned, the majority of generic combinatorial problems — are reducible to each other and are therefore of basically the same complexity — are **NP-complete** — “as difficult as a problem can be”.

- In the above list the only “easy” — known to be efficiently solvable — problem is Shortest Path, while all other problems are of basically the same “maximal possible” complexity.
Most of solution methods for difficult combinatorial problems heavily use *bounding*. Bounding techniques are aimed at building “efficiently computable” lower bounds for the optimal value in combinatorial problem

\[
\min_x \{ f(x) : x \in \mathcal{X} \}. \tag{Ini}
\]

A typical way to find such a bound is given by *relaxation*: we replace \( \mathcal{X} \) with a larger set \( \mathcal{X}^+ \) such that the problem

\[
\min_x \{ f(x) : x \in \mathcal{X}^+ \} \tag{Rel}
\]

is efficiently solvable, and use the optimal value of (Rel) as a lower bound on the optimal value of (Ini):

\[
\mathcal{X} \subset \mathcal{X}^+ \Rightarrow \text{Opt}(\text{Rel}) \leq \text{Opt}(\text{Ini}).
\]
**Generic example:** Let (Ini) be quadratic quadratically constrained problem:

\[
\text{Opt} = \min_x \left\{ x^T Q_0 x + 2 b_0^T x + c_0 : \begin{array}{l}
  f_i(x) = x^T Q_i x + 2 b_i^T x + c_i \leq 0, i = 1, ..., m \\
  h_\ell(x) = x^T R_\ell x + 2 d_\ell^T x + e_\ell = 0, \ell = 1, ..., k
\end{array} \right\} \quad (\text{Ini})
\]

Setting

\[
X(x) = \begin{bmatrix} x \\ 1 \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix}^T = \begin{bmatrix} x x^T \\ x^T \\ 1 \end{bmatrix},
\]

\[
A_i = \begin{bmatrix} Q_i & b_i \\ b_i^T & c_i \end{bmatrix}, \quad i = 0, ..., m,
B_\ell = \begin{bmatrix} R_\ell & d_\ell \\ d_\ell^T & e_\ell \end{bmatrix}, \quad \ell = 1, ..., k,
\]

we can write down (Ini) equivalently as

\[
\min_{X} \left\{ \begin{array}{l}
  \text{Tr}(A_0 X) \\
  \text{Tr}(A_i X) \leq 0, i = 1, ..., m, \\
  \text{Tr}(B_\ell X) = 0, \ell = 1, ..., k,
\end{array} \right\}, \quad X \in \mathcal{X} = \{X = X(x) : x \in \mathbb{R}^n\}. \quad (\text{Med})
\]

Matrices \(X \in \mathcal{X} \in \mathbb{S}^{n+1}\) clearly are \(\succeq 0\), so that

\[
\mathcal{X} \subset \mathcal{X}^+ = \{X \in \mathbb{S}^{n+1} : X \succeq 0, X_{n+1,n+1} = 1\}.
\]

Consequently, the semidefinite program

\[
\text{Opt}_{\text{Rel}} = \min_{X} \left\{ \begin{array}{l}
  \text{Tr}(A_0 X) \\
  \text{Tr}(A_i X) \leq 0, i = 1, ..., m, \\
  \text{Tr}(B_\ell X) = 0, \ell = 1, ..., k,
\end{array} \right\}, \quad X \succeq 0, X_{n+1,n+1} = 1
\]

is a relaxation of (Ini).

3.61
Another way to get the same relaxation is given by

**Weak Lagrange Duality:** Consider an optimization program

\[
\text{Opt} = \min_x \left\{ f_0(x) : \begin{array}{l}
  f_i(x) \leq 0, \ i = 1, \ldots, m; \\
  h_\ell(x) = 0, \ \ell = 1, \ldots, k.
\end{array} \right\} \quad \text{(Ini)}
\]

Let

\[
L(x; \lambda, \mu) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{\ell=1}^k \mu_\ell h_\ell(x) \quad [\lambda_i \geq 0]
\]

be the Lagrange function of (Ini). We clearly have

\[
\lambda \geq 0, x \text{ feasible for (Ini)} \Rightarrow L(x; \lambda, \mu) \leq f_0(x)
\]

and therefore

\[
\lambda \geq 0 \Rightarrow F(\lambda, \mu) \equiv \inf_{x \in \mathbb{R}^n} L(x; \lambda, \mu) \leq \text{Opt}.
\]

It follows that

\[
\text{Opt}_{\text{Lag}} \equiv \sup_{\lambda \geq 0, \mu} F(\lambda, \mu) \leq \text{Opt}.
\]
(Ini): $\text{Opt} = \min_x \left\{ f_0(x): f_i(x) \leq 0, \ i = 1, \ldots, m; \ h_\ell(x) = 0, \ \ell = 1, \ldots, k. \right\}$
$\Rightarrow L(x; \lambda, \mu) = f_0(x) + \sum_i \lambda_i f_i(x) + \sum_\ell \mu_\ell h_\ell(x)$
$\Rightarrow F(\lambda, \mu) = \inf_{x \in \mathbb{R}^n} L(x; \lambda, \mu)$
$\Rightarrow \text{Opt}_* \equiv \sup_{\lambda \geq 0, \mu} F(\lambda, \mu) \leq \text{Opt}$

**Shor’s bounding scheme:** Assume that all functions $f_0, \ldots, f_m, h_0, \ldots, h_k$ are quadratic:

$$f_i(x) = x^T Q_i x + 2b_i^T x + c_i, \quad h_\ell = x^T R_\ell x + 2d_\ell^T x + e_\ell$$

and let us apply the Weak Duality:

$$L(x; \lambda, \mu) = f_0(x) + \sum_i \lambda_i f_i(x) + \sum_\ell \mu_\ell h_\ell(x)$$
$$= x^T [Q(\lambda, \mu)] x + 2[q(\lambda, \mu)]^T x + r(\lambda, \mu)$$

$Q(\lambda, \mu) = Q_0 + \sum_{i \geq 1} \lambda_i Q_i + \sum_\ell \mu_\ell R_\ell,$
$q(\lambda, \mu) = b_0 + \sum_{i \geq 1} \lambda_i b_i + \sum_\ell \mu_\ell d_\ell,$
$r(\lambda, \mu) = c_0 + \sum_{i \geq 1} \lambda_i c_i + \sum_\ell \mu_\ell e_\ell$

What is $\inf_x L(x; \lambda, \mu)$?
\[
L(x; \lambda, \mu) = f_0(x) + \sum_i \lambda_i f_i(x) + \sum_\ell \mu_\ell h_\ell(x) = x^T [Q(\lambda, \mu)] x + 2 [q(\lambda, \mu)]^T x + r(\lambda, \mu)
\]
\[
Q(\lambda, \mu) = Q_0 + \sum_{i \geq 1} \lambda_i Q_i + \sum_\ell \mu_\ell R_\ell, q(\lambda, \mu) = b_0 + \sum_{i \geq 1} \lambda_i b_i + \sum_\ell \mu_\ell d_\ell, r(\lambda, \mu) = c_0 + \sum_{i \geq 1} \lambda_i c_i + \sum_\ell \mu_\ell e_\ell
\]

**Lemma:** A quadratic form \( x^T Q x + 2 q^T x + r \) is \( \geq s \) for all \( x \) iff
\[
\begin{bmatrix}
Q & q \\
q^T & r - s
\end{bmatrix} \succeq 0.
\]

By Lemma,
\[
\inf_x L(x; \lambda, \mu) = \sup \left\{ s : \begin{bmatrix}
\frac{Q(\lambda, \mu)}{q^T(\lambda, \mu)} & \frac{q(\lambda, \mu)}{r(\lambda, \mu) - s}
\end{bmatrix} \succeq 0 \right\}
\]
whence
\[
\text{Opt}_{\text{Lag}} = \max_{\lambda, \mu, s} \left\{ s : \begin{bmatrix}
\frac{Q(\lambda, \mu)}{q^T(\lambda, \mu)} & \frac{q(\lambda, \mu)}{r(\lambda, \mu) - s}
\end{bmatrix} \succeq 0, \lambda \geq 0 \right\} \quad \text{(Lag)}
\]
and this optimal value is a lower bound for
\[
\text{Opt} = \min_x \left\{ f_0(x) : \begin{array}{l}
f_i(x) \leq 0, \ i = 1, \ldots, m; \\
h_\ell(x) = 0, \ \ell = 1, \ldots, k.
\end{array} \right\}
\]
\[
\begin{bmatrix}
f_i(x) = x^T Q_i x + 2 b_i^T x + c_i, h_\ell = x^T R_\ell x + 2 d_\ell^T x + e_\ell
\end{bmatrix}
\]
\[ \text{Opt} = \min_x \left\{ f_0(x) : \begin{array}{l} f_i(x) \leq 0, \ i = 1, \ldots, m; \\ h_\ell(x) = 0, \ \ell = 1, \ldots, k. \end{array} \right\} \]  
\[ [f_i(x) = x^T Q_i x + 2b_i^T x + c_i, \ h_\ell = x^T R_\ell x + 2d_\ell^T x + e_\ell] \]  

The Semidefinite Relaxation and Shor’s Bounding yield, respectively, the lower bounds

\[ \text{OptRel} = \min_X \left\{ \text{Tr}(A_i X) : \begin{array}{l} \text{Tr}(A_i X) \leq 0, \ i = 1, \ldots, m \\ \text{Tr}(A_0 X), \ X \succeq 0, X_{n+1,n+1} = 1 \end{array} \right\} \]  
\[ A_i = \begin{bmatrix} Q_i & b_i^T \\ b_i & c_i \end{bmatrix}, \ i = 1, \ldots, m, B_\ell = \begin{bmatrix} R_\ell & d_\ell^T \\ d_\ell & e_\ell \end{bmatrix}, \ \ell = 1, \ldots, k \]  

and

\[ \text{OptLag} = \max_{\lambda, \mu, s} \left\{ s : \begin{array}{l} Q(\lambda, \mu) q(\lambda, \mu) \geq 0, \ \lambda \geq 0 \ \text{and} \\ Q(\lambda, \mu) = Q_0 + \sum_{i \geq 1} \lambda_i Q_i + \sum_{\ell} \mu_\ell R_\ell, \\ q(\lambda, \mu) = b_0 + \sum_{i \geq 1} \lambda_i b_i + \sum_{\ell} \mu_\ell d_\ell, \\ r(\lambda, \mu) = c_0 + \sum_{i \geq 1} \lambda_i c_i + \sum_{\ell} \mu_\ell e_\ell \end{array} \right\}, \]  

on \text{Opt}.

\bullet \text{It is immediately seen that (Rel) is (equivalent to) the dual of (Lag), so that both bounds are the same (provided that one of the relaxations is strictly feasible)!}
Example: Lovasz $\vartheta$-function

- **A graph** is a finite set of *nodes* linked by *arcs*. A subset $S$ of the nodal set is called *independent*, if no pair of nodes from $S$ are linked by an arc. The **stability number** $\alpha(\Gamma)$ of a graph $\Gamma$ is the maximum cardinality of independent sets of nodes. E.g., the stability number of graph $C_5$

![Graph C_5](image_url)

is 2.

- To compute $\alpha(\Gamma)$ is an NP-complete combinatorial problem.
Shannon capacity $\Theta(\Gamma)$ of a graph $\Gamma$ is defined as follows. Imagine that the nodes are letters of an alphabet. We can sent these letters through a communication channel. When passing through the channel, a letter may be corrupted by noise; as a result, two distinct letters on input to the channel may become the same on the output. We link every pair of letters with this property by an arc, thus getting a graph.

- Assume we are sending $k$-letter words, one letter per unit time, and want to avoid “misunderstandings” — the addressee should be capable to recognize what word was sent, without risk that “no!” will be read as “yes”.

To avoid misunderstandings, we should restrict the “dictionary” of $n$-letter words we actually use to be “independent” in the sense that no two distinct words from the dictionary, as sent through the channel, can produce the same output. If we agree with addressee what is the independent dictionary we use, no misunderstandings will occur.
• In order to fully utilize the capacity of the channel, it makes sense to use a maximum cardinality independent dictionary of $k$-letter words, let this cardinality be $f(k)$. It is clear that

$$f(k + l) \geq f(k)f(l)$$

and that $f^{1/k}(k)$ is above bounded (e.g., by the number of letters). From these properties it follows that

$$\sup_{k \geq 1} f^{1/k}(k) = \lim_{k \to \infty} f^{1/k}(k) \equiv \sigma(\Gamma);$$

$\sigma(\Gamma)$ is called \textit{Shannon capacity} of graph $\Gamma$.

• Since the maximum cardinality of independent single-letter dictionaries is the stability number of the graph, we have

$$\alpha(\Gamma) = f(1) \leq \sigma(\Gamma).$$
\[ \alpha(\Gamma) \leq \sigma(\Gamma). \]  

(*)

- Inequality (*) may be strict. E.g., \( \alpha(C_5) = 2 \):

Graph \( C_5 \)
At the same time, for $C_5$ there exists independent dictionaries with 5 two-letter words, e.g., \{AA, BC, CE, DB, ED\}

Thus, $\sigma(C_5) \geq \sqrt{f(2)} = \sqrt{5}$. 

The question whether this inequality is equality remained open for about 20 years!
• In early 70’s, L. Lovasz found a computable upper bound $\vartheta(\Gamma)$ for $\alpha(\Gamma)$ and proved that

$$\alpha(\Gamma) \leq \sigma(\Gamma) \leq \vartheta(\Gamma)$$

(In particular, $\sqrt{5} \leq \sigma(C_5) \leq \vartheta(C_5) = \sqrt{5}$, whence $\sigma(C_5) = \sqrt{5}$).

• By definition, $\vartheta(\Gamma)$ is the optimal value in the following semidefinite program:

$$\min_{X \in \mathcal{L}} \lambda_{\max}(X) \equiv \min_{X \in \mathcal{L}, \mu} \{ \mu : \mu I \succeq X \} \quad \text{(Lov)}$$

where $\mathcal{L}$ is the set of all symmetric $n \times n$ matrices $X$ ($n$ is the number of nodes in the graph) such that $X_{ij} = 1$ when the nodes $i, j$ are not adjacent.

3.71
Example: For graph $C_5$, the set $\mathcal{L}$ is comprised of all matrices of the form

$$\begin{bmatrix}
1 & x_{AB} & 1 & 1 & x_{EA} \\
x_{AB} & 1 & x_{BC} & 1 & 1 \\
1 & x_{BC} & 1 & x_{CD} & 1 \\
1 & 1 & x_{CD} & 1 & x_{DE} \\
x_{EA} & 1 & 1 & x_{DE} & 1 \\
\end{bmatrix}.$$
The Lovasz upper bound on $\alpha(\Gamma)$ can be obtained from Shor’s Bounding scheme.

Let the nodes of $\Gamma$ be $1,\ldots,n$.

Observe that $\alpha(\Gamma)$ is the optimal value in the Boolean quadratic program:

$$(a) \quad \max_x \sum_{i=1}^{n} x_i \quad (b) \quad 2x_ix_j = 0 \quad \forall \text{ adjacent } i, j \quad (c) \quad x_i^2 - x_i = 0 \quad \iff \quad x_i \in \{0; 1\}$$

(c) associates with $x$ the set of nodes $\{i : x_i = 1\}$;

(b) says that the set $\{i : x_i = 1\}$ is independent;

(a) counts the cardinality of $\{i : x_i = 1\}$.

Applying Shor’s scheme, we come to the “bounding program”

$$\min_{\mu, \nu, Y} \left\{ \mu : \begin{bmatrix} Y + \text{Diag}\{\nu\} & -\frac{1}{2}[\nu + 1] \\ -\frac{1}{2}[\nu + 1]^T & \mu \end{bmatrix} \succeq 0 \right\}, \quad 1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad \text{(Lag)}$$

$\text{[Opt(Lag) } \geq \alpha(\Gamma)]$
\[
\min_{\mu, \nu, Y} \left\{ \mu : \begin{bmatrix}
Y + \text{Diag}\{\nu\} & -\frac{1}{2}[\nu + 1] \\
-\frac{1}{2}[\nu + 1]^T & \mu
\end{bmatrix} \succeq 0 \right\}, \quad 1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \tag{Lag}
\]

- Applying Lemma on Schur Complement, we convert (Lag) to

\[
\min_{\mu \geq 0, \nu, Y} \left\{ \mu : \mu(Y + \text{Diag}\{\nu\}) \succeq \frac{1}{4}(\nu + 1)(\nu + 1)^T \right\}
\]

- Specifying \( \nu \)-variables as ones, we can only increase the optimal value. The resulting problem is

\[
\text{SDP} = \min_{\mu, X} \left\{ \mu : \mu I \succeq -\mu Y + 1 \cdot 1^T \right\}
\]

\[
\text{SDP} \succeq \alpha(\Gamma)
\]

- When \( Y \) runs through the set of symmetric matrices such that \( Y_{ij} = 0 \) for non-adjacent \( i, j \), \( X \) runs through the entire set of symmetric matrices with \( X_{ij} = 1 \) for non-adjacent \( i, j \), so that

\[
\text{SDP} = \min_{\mu, X} \left\{ \mu : \mu I \succeq X \right\}
\]

\[
\text{SDP} \succeq \alpha(\Gamma)
\]
How close is $\vartheta(\Gamma)$ to $\alpha(\Gamma)$?

There exists an important class of perfect graphs for which $\vartheta(\Gamma) = \alpha(\Gamma)$.

However, for general-type graphs it may happen that $\vartheta(\Gamma) \gg \alpha(\Gamma)$.

Lovasz have proved that if $\Gamma$ is an $n$-node graph and $\hat{\Gamma}$ is its complement (two distinct nodes are linked by arc in $\hat{\Gamma}$ iff they are not linked by arc in $\Gamma$), then

$$\vartheta(\Gamma) \vartheta(\hat{\Gamma}) \geq n \Rightarrow \max [\vartheta(\Gamma), \vartheta(\hat{\Gamma})] \geq \sqrt{n}.$$ 

On the other hand, for a random $n$-node graph $\Gamma$ (probability for a pair $i < j$ to be linked by an arc is $\frac{1}{2}$) it holds

$$\max [\alpha(\Gamma), \alpha(\hat{\Gamma})] \leq O(\ln n)$$

with probability approaching 1 as $n \to \infty$.

Thus, for “typical” random graphs

$$\frac{\vartheta(\Gamma)}{\alpha(\Gamma)} \geq O \left( \frac{\sqrt{n}}{\ln n} \right).$$
B. Theorem of Goemans and Williamson. There exist hard combinatorial problems where bounds coming from semidefinite relaxations coincide with the actual optimal value within $absolute$ constant factor. The most famous example is given by the MAXCUT problem which is as follows:

*Given a graph $\Gamma$ with arcs assigned nonnegative weights $a_{ij}$, find a cut of maximal weight*

[A cut in a graph is partitioning $(S, S')$ of the set of nodes into two non-overlapping subsets. The weight of a cut is the sum of weights of all arcs linking a node from $S$ with a node from $S'$.]
♠ MAXCUT is an NP-complete combinatorial problem which can be posed as quadratic program with variables $\pm 1$:

- We lose nothing by assuming that graph is complete (set $a_{ij} = 0$ for pairs $i, j$ of nodes which in fact are not adjacent). Thus, assume that $a_{ij}$ form a symmetric $n \times n$ matrix $A$ with nonnegative entries and zero diagonal.
- A cut $(S, S')$ can be represented by vector $x \in \mathbb{R}^n$ with $x_i = -1$ for $i \in S$ and $x_i = 1$ for $i \in S'$. With this representation, the weight of the cut is
  \[
  \frac{1}{4} \sum_{i,j} a_{ij} (1 - x_i x_j)
  \]  
  (*)
- Thus, MAXCUT is the program
  \[
  OPT = \max_x \left\{ \frac{1}{4} \sum_{i,j} a_{ij} (1 - x_i x_j) : x_i = \pm 1 \right\}.
  \]  
  (MAXCUT)
- Applying the Semidefinite Relaxation scheme, we get an SDP relaxation of MAXCUT as follows:
  \[
  SDP = \max_X \left\{ \frac{1}{4} \sum_{i,j} a_{ij} (1 - X_{ij}) : X = [X_{ij}] \succeq 0, \text{Dg}(X) = 1 \right\}.
  \]  
  (SDP)

3.77
\[ \text{OPT} = \max_x \left\{ \frac{1}{4} \sum_{i,j} a_{ij} (1 - x_i x_j) : x_i = \pm 1 \right\} \quad \text{(MAXCUT)} \]
\[ \text{SDP} = \max_X \left\{ \frac{1}{4} \sum_{i,j} a_{ij} (1 - X_{ij}) : X = [X_{ij}] \succeq 0, \text{Dg}(X) = 1 \right\} \quad \text{(SDP)} \]

**Theorem** [Goemans & Williamson, 1995]

\[ \text{OPT} \leq \text{SDP} \leq \alpha \cdot \text{OPT}, \, \alpha = 1.138... \]

**Proof.** The left inequality is evident. Let \( X^* \) be optimal for (SDP), let \( \xi \sim \mathcal{N}(0, X^* \) and let \( \zeta = \text{sign}[\xi] \). Then

\[
[\text{OPT} \geq] \quad \mathbb{E} \left\{ \frac{1}{4} \sum_{i,j} a_{ij} (1 - \zeta_i \zeta_j) \right\} = \frac{1}{4} \sum_{i,j} a_{ij} (1 - \frac{2}{\pi} \text{asin}(X^*_{ij})) \quad \text{[computation]}
\geq \frac{1}{4} \alpha^{-1} \sum_{i,j} a_{ij} (1 - X^*_{ij})
\quad \text{[due to} \ a_{ij} \geq 0 \ \text{and} \ (1 - \frac{2}{\pi} \text{asin}(t)) \geq \alpha^{-1} (1 - t), \ -1 \leq t \leq 1 \]
\]

Thus,

\[ \text{SDP} \leq \alpha \cdot \text{OPT} \]

3.78
C. Nesterov’s $\frac{\pi}{2}$ Theorem. The GW Theorem states that with $Q$ given by

$$Q_{ij} = \begin{cases} \sum_{p=1}^{n} a_{ip}, & i = j \\ -a_{ij}, & i \neq j \end{cases} \quad (*)$$

where $a_{ij} \geq 0$, the semidefinite upper bound

$$SDP = \max_X \{ \text{Tr}(QX) : X \succeq 0, X_{ii} = 1, i = 1, ..., n \}$$

(SDP)

on the combinatorial quantity

$$OPT = \max_x \{ \text{Tr}(Qxx^T) : x_i = \pm 1, i = 1, ..., n \}$$

(QP)

is tight within the factor $1.138....$

- $Q$ as given by $(*)$ (where $a_{ij} \geq 0$) is a very specific positive semidefinite matrix. What is the relation between $SDP$ and $OPT$ for an arbitrary $Q \succeq 0$?

Nesterov’s $\frac{\pi}{2}$ Theorem: When $Q \succeq 0$, one has

$$OPT \leq SDP \leq \frac{\pi}{2} \cdot OPT.$$
\[
SDP = \max_X \{ \text{Tr}(QX) : X \succeq 0, X_{ii} = 1, i = 1, \ldots, n \} \quad \text{(SDP)}
\]
\[
OPT = \max_x \{ \text{Tr}(Qxx^T) : x_i = \pm 1, i = 1, \ldots, n \} \quad \text{(QP)}
\]

Claim: \( OPT \leq \frac{\pi}{2} SDP \)

Proof. Let \( X^* \) be an optimal solution to (SDP), let \( \xi \sim \mathcal{N}(0, X^*) \) and let \( \zeta = \text{sign}[\xi] \). Then

\[
[OPT \geq] \quad \mathbb{E} \left\{ \zeta^T Q \zeta \right\} = \text{Tr}(Q \frac{2}{\pi} [\text{asin}(X^*_{ij})]_{i,j}) \quad (1)
\]

Lemma: Let \( X \succeq 0 \) and \( |X_{ij}| \leq 1 \). Then \( \text{asin}[X] \succeq X \).

Proof: Denoting \( [X]^k = [X^k_{ij}]_{i,j} \) and taking into account that \( X \succeq 0 \Rightarrow [X]^k \succeq 0, k = 1, 2, \ldots, \) one has

\[
\text{asin}[X] - X = \sum_{k=1}^{\infty} \frac{1 \times 3 \times 5 \times \ldots \times (2k - 1)}{2^k k!(2k + 1)} [X]^{2k+1} \geq 0.
\]

By Lemma and since \( Q \succeq 0 \), the right hand side in (1) is \( \geq \frac{2}{\pi} \text{Tr}(QX^*) = \frac{2}{\pi} SDP \), whence \( SDP \leq \frac{\pi}{2} OPT \).

\[\square\]
We have used the following

**Fact:** If $X = [x_{ij}]_{i,j \leq n}$, $Y = [y_{ij}]_{i,j \leq n}$ are positive semidefinite matrices of the same order, then the entrywise product of $X$ and $Y$ – the matrix

$$X \cdot Y = [x_{ij}y_{ij}]_{i,j \leq n}$$

is positive semidefinite as well.

Indeed, symmetric matrix $Q$ is $\succeq 0$ iff $Q = F^TF$ for some rectangular matrix $F$, or, which is the same, iff $Q$ is a Gram matrix:

$$x_{ij} = f_i^Tf_j$$

for some $f_i \in \mathbb{R}^N$ (treat $f_i$ as the columns of $F$). And entrywise product of Gram matrices again is a Gram matrix:

$$x_{ij} = f_i^Tf_j, y_{ij} = g_i^Tg_j \Rightarrow x_{ij}y_{ij} = \text{Vec}^T(f_ig_i^T)\text{Vec}(f_jg_j^T)$$

\[ \Box \]
The $\frac{\pi}{2}$ Theorem admits important corollaries:

**Corollary 1** [Nesterov '97] Let $T \subset \mathbb{R}_+^n$ be a nonempty $SDr$ compact set, and let $Q$ be an $n \times n$ symmetric matrix. Then the quantities

$$m_*(Q) = \min_x \left\{ x^T Q x : (x_1^2, \ldots, x_n^2)^T \in T \right\},$$

$$m^*(Q) = \max_x \left\{ x^T Q x : (x_1^2, \ldots, x_n^2)^T \in T \right\}$$

admit efficiently computable, via SDP, bounds

$$s_*(Q) \equiv \min_X \left\{ \text{Tr}(QX) : X \succeq 0, (X_{11}, \ldots, X_{nn})^T \in T \right\},$$

$$s^*(Q) \equiv \max_X \left\{ \text{Tr}(QX) : X \succeq 0, (X_{11}, \ldots, X_{nn})^T \in T \right\}$$

such that

$$s_*(Q) \leq m_*(Q) \leq m^*(Q) \leq s^*(Q)$$

and

$$m^*(Q) - m_*(Q) \leq s^*(Q) - s_*(Q) \leq \frac{\pi}{4 - \pi} (m^*(Q) - m_*(Q))$$

Thus, one can bound from above the variation $m^*(Q) - m_*(Q)$ by the efficiently computable quantity $s^*(Q) - s_*(Q)$, and this bound is tight within the absolute constant factor $\frac{\pi}{4 - \pi}$.
Corollary 2 [Nesterov '97] Let $p \in [2, \infty]$, $r \in [1, 2]$, and let $A$ be an $m \times n$ matrix. Consider the problem of computing the operator norm $\|A\|_{p,r}$ of the linear mapping $x \mapsto Ax$, considered as the mapping from the space $\mathbb{R}^n$ equipped with the norm $\| \cdot \|_p$ to the space $\mathbb{R}^m$ equipped with the norm $\| \cdot \|_r$:

$$\|A\|_{p,r} = \max \{ \|Ax\|_r : \|x\|_p \leq 1 \};$$

(it is NP-hard to compute this norm, except for the case of $p = r = 2$). The “computationally intractable” quantity $\|A\|_{p,r}$ admits an efficiently computable upper bound

$$\omega_{p,r}(A) = \min_{\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^n} \left\{ \frac{1}{2} \left[ \|\mu\|_p^{p-2} + \|\lambda\|_r^{2-r} \right] : \begin{bmatrix} \text{Diag}\{\mu\} & A^T \\ A & \text{Diag}\{\lambda\} \end{bmatrix} \succeq 0 \right\}.$$

This bound is exact for a nonnegative matrix $A$, and for an arbitrary $A$ the bound is tight within the factor $\frac{\pi}{2\sqrt{3} - 2\pi/3} = 2.293...$:

$$\|A\|_{p,r} \leq \omega_{p,r}(A) \leq \frac{\pi}{2\sqrt{3} - 2\pi/3} \|A\|_{p,r}.$$

Moreover, if $p \in [1, \infty]$ and $r \in [1, 2]$ are rational, the bound $\omega_{p,r}(A)$ is an SDr function of $A$. 3.83
D. Semidefinite Relaxation on Ellitopes

♠ A basic ellitope is a set $\mathcal{X} \subset \mathbb{R}^n$ represented as

$$\mathcal{X} = \{ x : \exists t \in \mathcal{T} : x^T S_k x \leq t_k, 1 \leq k \leq K \}$$

- $S_k \succeq 0, k \leq K, \sum_k S_k \succ 0$
- $\mathcal{T}$: convex compact set in $\mathbb{R}_+^k$ containing a positive vector and monotone: $0 \leq t' \leq t \in \mathcal{T} \Rightarrow t' \in \mathcal{T}$

♠ An ellitope $\mathcal{Y}$ is a set represented as a linear image of basic ellitope:

$$\mathcal{Y} = \{ y : \exists (t \in \mathcal{T}, x) : y = Px. x^T S_k x \leq t_k, k \leq K \}.$$

Examples: A. Bounded intersection $\mathcal{X}$ of $K$ centered at the origin ellipsoids/elliptic cylinders \{ $x : x^T S_k x \leq 1$ \} $[S_k \succeq 0]$ is a basic ellitope:

$$\mathcal{X} = \{ x : \exists t \in \mathcal{T} := [0, 1]^K : x^T S_k x \leq t_k, k \leq K \}$$

B. $\| \cdot \|_p$-ball in $\mathbb{R}^n$ with $p \in [2, \infty]$ is a basic ellitope:

$$\{ x \in \mathbb{R}^n : \| x \|_p \leq 1 \} = \{ x : \exists t \in \mathcal{T} = \{ t \in \mathbb{R}_+^n, \| t \|_{p/2} \leq 1 \} : x_k^2 \leq t_k, k \leq K \}.$$
Fact: Ellitopes admit fully algorithmic ”calculus:” this family is closed with respect to basic operations preserving convexity and symmetry w.r.t. the origin, like taking finite intersections, linear images, inverse images under linear embeddings, direct products, arithmetic summation.

- What is missing, is taking convex hulls of finite unions.
Fact: When maximizing a quadratic form $y^T Cy$ over an ellitope $\mathcal{Y} = P \mathcal{X}$, $\mathcal{X} = \{x : \exists t \in \mathcal{T} : x^T S_k x \leq t_k, k \leq K\}$ semidefinite relaxation works reasonably well. This is how it works:

- Passing from the quadratic form $y^T Cy$ to the “lifted” form $x^T [P^T C P] x$, we reduce the situation to maximizing quadratic form $x^T D x$ over the basic ellitope $\mathcal{X}$.
- For $\lambda \in \mathbb{R}^K$, let $\phi_{\mathcal{T}}(\lambda) = \max_{t \in \mathcal{T}} t^T \lambda$ be the support function of $\mathcal{T}$. When $\lambda \geq 0$ is such

$$D \preceq \sum_k \lambda_k S_k,$$

and $x \in \mathcal{X}$, there exists $t \in \mathcal{T}$ such that $x^T S_k x \leq t_k$, $k \leq K$,

$$x^T D x \leq x^T [\sum_k \lambda_k S_k] x \leq \sum_k \lambda_k t_k \leq \phi_{\mathcal{T}}(\lambda)$$

$$\Rightarrow \max_{x \in \mathcal{X}} x^T D x \leq \text{Opt} := \min \left\{ \phi_{\mathcal{T}}(\lambda) : \lambda \geq 0, D \preceq \sum_k \lambda_k S_k \right\}$$
\[ X = \{ x \in \mathbb{R}^n : \exists t \in T : x^T S_k x \leq t_k, \ k \leq K \} \quad [S_k \geq 0, \sum_k S_k > 0] \]
\[ \Rightarrow \max_{x \in X} x^T D x \leq \text{Opt} := \min \left\{ \phi_T(\lambda) : \lambda \geq 0, D \preceq \sum_k \lambda_k S_k \right\} \]

**Theorem** [Proposition 4.3.3 in https://www2.isye.gatech.edu/~nemirovs/StatOpt_LN_NS.pdf] One has

\[ \max_{x \in \mathcal{X}} x^T D x \leq \text{Opt} \leq 4 \ln(5K) \max_{x \in \mathcal{X}} x^T D x \]
Application: Near-Optimal Linear Estimation

Consider the following basic statistical problem: Given noisy observation

\[ \omega = Ax + \xi \]

[\[ \xi : \text{standard (zero mean, unit covariance) Gaussian noise} \]

of unknown signal \( x \) known to belong to a given “signal set” \( \mathcal{X} \), recover the linear image \( Bx \) of \( x \).

We quantify the performance of a candidate estimate \( \hat{x}(\cdot) \) by its risk

\[ \text{Risk}[\hat{x}|\mathcal{X}] = \left[ \sup_{x \in \mathcal{X}} \mathbb{E}_\xi \left\{ \|\hat{x}(Ax + \xi) - Bx\|^2 \right\} \right]^{1/2} \]

The simplest estimates are linear ones: \( \hat{x}(\omega) = \hat{x}_H(\omega) := H^T \omega \).

The squared risk of a linear estimate is given by

\[ \text{Risk}^2[\hat{x}|\mathcal{X}] = \max_{x \in \mathcal{X}} \| [B - H^T A]x \|^2 + \text{Tr}(H H^T) \]

The minimum risk linear estimate is given by an optimal solution to the convex optimization problem

\[ \min_H \left\{ \Phi(H) + \text{Tr}(H H^T) \right\}, \Phi(H) := \max_{x \in \mathcal{X}} x^T \left[ [B - H^T A]^T [B - H^T A] \right] x \]
\[
\text{Opt}_* = \min_H \{ \Phi(H) + \text{Tr}(HH^T) \}, \quad \Phi(H) := \max_{x \in \mathcal{X}} x^T \left[ [B - H^T A]^T [B - H^T A] \right] x
\]

**Difficulty:** \( \Phi \), while convex, is, in general, difficult to compute. The only generic “easy cases” here are those of an ellipsoid \( \mathcal{X} \), or \( \mathcal{X} \) given as a convex hull of finite set.

**Partial remedy** when \( \mathcal{X} \) is an ellitope: use semidefinite relaxation.

\[\blacklozenge\text{ Assuming that the ellitope } \mathcal{X} \text{ is basic (this is w.l.o.g.):} \]

\[\mathcal{X} = \{x : \exists t \in \mathcal{T} : x^T S_k x \leq t_k, \ k \leq K\} \]

Semidefinite relaxation combined with the Schur Complement Lemma results in the tractable relaxation

\[\text{Opt} = \min_{H, \lambda} \left\{ \phi_T(\lambda) + \text{Tr}(HH^T) : \lambda \geq 0, \left[ \frac{\sum_k \lambda_k S_k}{[B - H^T A]} \right] \left[ [B - H^T A]^T \right] \succeq 0 \right\} \]

of the problem of interest. We have \( \text{Opt} \leq 4 \ln(5K) \text{Opt}_* \), implying that the efficiently computable optimal solution to the relaxed problem results in linear estimate with optimal, within the factor \( 2 \sqrt{\ln(5K)} \), risk achievable with linear estimates.

**Fact:** The resulting sub-optimal linear estimate is “near-optimal” (optimal within an explicit logarithmic factor) among all estimates, linear and nonlinear alike.
E. The Matrix Cube Theorem. Consider the following problem: MATRCUBE: Given symmetric $m \times m$ matrices $B_0 \succeq 0, B_1, \ldots, B_L$, solve the optimization problem

$$\rho^* = \max \left\{ \rho : A[\rho] \equiv \left\{ B_0 + \sum_{\ell=1}^{L} u_\ell B_\ell : \|u\|_\infty \leq \rho \right\} \subset S^m_+ \right\}$$

i.e., find the largest $\rho$ such that the “matrix box” $A[\rho]$ is contained in the semidefinite cone.

This problem is easy when all “edge matrices” $B_\ell, \ell \geq 1$, are of rank 1, and can be NP-hard already when the “edge matrices” are of rank 2.
**Matrix Cube Theorem** [Ben-Tal & Nemirovski, ’00] Given $\rho \geq 0$, consider the system of LMI’s

\[
X^\ell \succeq \pm B^\ell, \; \ell = 1, \ldots, L, \\
\rho \sum_{\ell=1}^{L} X^\ell \preceq B_0
\]

in matrix variables $X^1, \ldots, X^L$.

(i) If $\left( S[\rho] \right)$ is solvable, then $A[\rho]$ is contained in $S^m_+$

(ii) If $\left( S[\rho] \right)$ is unsolvable, then $A[\vartheta(\mu)\rho]$ is not contained in $S^m_+$. Here

$$\mu = \max_{1 \leq \ell \leq L} \text{Rank}(B^\ell)$$

(note $\ell \geq 1$ in the max!) and $\vartheta(\mu)$ is a universal function such that

$$\vartheta(1) = 1, \quad \vartheta(2) = \frac{\pi}{2}, \quad \vartheta(k) \leq \frac{\pi \sqrt{k}}{2}.$$ 

In particular, the efficiently computable quantity

$$\hat{\rho} = \max \{ \rho : (S[\rho]) \text{ is solvable} \}$$

is a lower bound on $\rho^*$, and this bound is tight within the factor $\vartheta(\mu)$: $\hat{\rho} \leq \rho^* \leq \vartheta(\mu)\hat{\rho}$.
Lyapunov Stability Analysis revisited. Recall that Lyapunov Stability Certificates, if any, for uncertain dynamical system

\[ \dot{x} = A(t)x, \quad [A(t) \in U] \]

are exactly the solutions \( X \) to the semi-infinite system of LMIs

\[ X \succeq I, \quad A^T X + X A \preceq -I \quad \forall (A \in U) \quad (L[U]) \]

Consider the case of “interval uncertainty”:

\[ U = U_\rho \equiv \{ A : |A_{ij} - A^*_{ij}| \leq \rho D_{ij}, \ i, j = 1, \ldots, n \}, \]

where \( A^* \) is the (stable) “nominal matrix”, \( \rho \) is the level of perturbations, and \( D_{ij} \geq 0 \) are “perturbation scales”.

How to compute the Lyapunov Stability Radius

\[ LSR[A^*, D] = \sup \{ \rho : (L[U_\rho]) \text{ is solvable} \} \]

3.92
- The interval uncertainty is a polytopic one, so that the semi-infinite system of LMIs \((L[U][\rho])\) is equivalent to the finite system of LMIs

\[
X \succeq I, \quad A_j^T X + X A_j \preceq -I \quad \forall j = 1, \ldots, J,
\]

\((*)\)

where \(A_1, \ldots, A_J\) are the vertices of the matrix box \(U_\rho\). However, \(J\) can blow up exponentially with the size \(n\) of the underlying dynamical system, so that \((*)\) is not computationally tractable, except for the case when “nearly all” entries in \(A(t)\) are certain.

- In fact, the problem of computing \(LSR\) for a general-type interval uncertainty is \(NP\)-hard.
Observe that

$$\text{LSR}[A^*, D] = \sup \left\{ \rho : \exists X \succeq I : A^T X + X A \preceq -I \forall (A : |A_{ij} - A^*_{ij}| \leq \rho D_{ij}) \right\}$$

$$= \sup \left\{ \rho : \exists X \succeq I : \left[ -I - (A^*)^T X - X A \right] + \sum_{i,j} u_{ij} D_{ij} [e_j e_i^T X + X e_i e_j^T] \succeq 0 \right\}$$

$$= \sup_{X \succeq I} \rho(X), \quad \rho(X) = \sup \left\{ \rho : B_0[X] + \sum_{i,j} u_{ij} B_{ij}[X] \succeq 0 \forall (u : \|u\|_\infty \leq \rho) \right\}$$

$\rho(X)$ is the optimal value in a MATRUCUBE problem with rank 2 edge matrices $B_{ij}[X]$. Applying the Matrix Cube Theorem, we conclude that

The efficiently computable quantity

$$\hat{\text{LSR}}[A^*, D] = \sup_{\rho, X, \{X_{ij}\}} \left\{ \rho : X_{ij} \succeq \pm B_{ij}[X], 1 \leq i, j \leq n \right\}$$

is a lower bound, tight within the factor $\frac{\pi}{2}$, on the Lyapunov Stability Radius $\text{LSR}[A^*, D]$. 

3.94
Similarly to Lyapunov Stability Analysis, the Matrix Cube Theorem allows to build tight, within an absolute constant factor, tractable approximations of numerous Control-originating semi-infinite LMIs affected by interval uncertainty.
Matrix Cube Theorem – Sketch of the Proof

Matrix Cube Theorem: Given \( \rho \geq 0 \), consider the system of LMI's

\[
X^\ell \succeq \pm B^\ell, \quad \ell = 1, \ldots, L, \\
\rho \sum_{\ell=1}^{L} X^\ell \preceq B_0
\]

\((S[\rho])\)

in matrix variables \( X^1, \ldots, X^L \).

(i) If \((S[\rho])\) is solvable, then the “matrix box”

\[
\mathcal{A}[\rho] \equiv \left\{ B_0 + \rho \sum_{\ell} u^\ell B^\ell : \|u\|_\infty \leq 1 \right\}
\]

is contained in \( \mathbb{S}_+^m \)

(ii) If \((S[\rho])\) is unsolvable, then the matrix box \( \mathcal{A}[\vartheta(\mu)\rho] \) is not contained in \( \mathbb{S}_+^m \). Here

\[
\mu = \max_{1 \leq \ell \leq L} \text{Rank}(B^\ell)
\]

\(\text{(note } \ell \geq 1 \text{ in the max!)}\) and \( \vartheta(\mu) \) is a universal function such that

\[
\vartheta(1) = 1, \quad \vartheta(2) = \frac{\pi}{2}, \quad \vartheta(k) \leq \frac{\pi \sqrt{k}}{2}.
\]

(i) is evident: whenever \( X^1, \ldots, X^L \) is a solution to \((S[\rho])\), we have

\[
\|u\|_\infty \leq 1 \Rightarrow u^\ell B^\ell \succeq -X^\ell \forall \ell \Rightarrow B_0 + \rho \sum_{\ell} u^\ell B^\ell \succeq B_0 - \rho \sum_{\ell} X^\ell \succeq 0.
\]

(ii): Assume that \((S[\rho])\) is not solvable, and let us prove that \( \mathcal{A}[\vartheta(\mu)\rho] \) is not contained in the positive semidefinite cone, provided that \( \vartheta(\mu) \) is chosen properly. There is nothing to prove when \( B_0 \not\succeq 0 \). Thus, let \( B_0 \succeq 0 \).
✿ **Step 1.** We have assumed that the system

\[
X^\ell \succeq \pm B_\ell, \; \ell = 1, \ldots, L, \\
\rho \sum_{\ell=1}^L X^\ell \preceq B_0
\]

(S[\rho])

has no solutions. Consider the semidefinite program

\[
\text{Opt} = \min_{X^\ell, t} \left\{ t : \begin{array}{c}
X^\ell \succeq \pm B_\ell, \; \ell = 1, \ldots, L, \\
\rho \sum_{\ell=1}^L X^\ell \preceq B_0 + tI
\end{array} \right\}
\]

(P)

The problem clearly is feasible and has compact level sets, and is therefore solvable. Since (S[\rho]) has no solutions, the optimal value in (P) is positive. Since the problem clearly is strictly feasible, the dual problem is solvable with positive optimal value.
\[
\text{Opt} = \min_{X_{\ell,t}} \left\{ t : \begin{array}{l}
X_{\ell} \geq \pm B_{\ell}, \ell = 1, \ldots, L,
\end{array}\right. \right\} \quad (P)
\]

\textbf{Step 2.} Let us build the dual. Let
- \( U_{\ell} \geq 0 \) be the “aggregation weights” for the constraints \( X_{\ell} \geq B_{\ell}, \)
- \( V_{\ell} \geq 0 \) be the aggregation weights for the constraints \( X_{\ell} \geq -B_{\ell}, \)
- \( W \geq 0 \) be the aggregation weight for the last LMI in \((P)\).

Aggregating the LMIs in \((P)\) with the above weights, we get the inequality
\[
\sum_{\ell} \text{Tr}([U_{\ell} + V_{\ell} - \rho W]X_{\ell}) + t \text{Tr}(W) \geq \sum_{\ell} \text{Tr}([U_{\ell} - V_{\ell}]B_{\ell}) - \text{Tr}(WB_0)
\]
Restricting the weights to be such that the left hand side in this inequality, as a function of \( X_{\ell} \) and \( t \), is identically equal to the objective in \((P)\):
\[
U_{\ell} + V_{\ell} = \rho W, \ell = 1, \ldots, L; \quad \text{Tr}(W) = 1 \quad (*)
\]
we obtain the lower bound \( \sum_{\ell} \text{Tr}([U_{\ell} - V_{\ell}]B_{\ell}) - \text{Tr}(WB_0) \) on \( \text{Opt} \). The dual problem is to maximize this bound:
\[
\max_{U_{\ell}, V_{\ell}, W} \left\{ \sum_{\ell} \text{Tr}([U_{\ell} - V_{\ell}]B_{\ell}) - \text{Tr}(WB_0) : \begin{array}{l}
U_{\ell} + V_{\ell} = W, \ell = 1, \ldots, L
\end{array} \right. \right]\}
\]
and we know that the optimal value in the dual is positive.
\[ 0 < \max_{U, V, W} \left\{ \sum_{\ell} \text{Tr}([U_{\ell} - V_{\ell}]B_{\ell}) - \text{Tr}(W B_0) : U_{\ell} + V_{\ell} = W, \ell = 1, \ldots, L, \text{Tr}(W) = 1, U_{\ell}, V_{\ell}, W \succeq 0 \right\} \]  

\[ (D) \]

\[ \bullet \text{ In } (D), \text{ we can carry out maximization in } U_{\ell}, V_{\ell} \text{ analytically. Indeed, this maximization requires solving the problem of the form} \]

\[ m(B, Z) \equiv \max_{U, V} \{ \text{Tr}([U - V]B) : U \succeq 0, V \succeq 0, U + V = Z \}, \]  

\[ (A) \]

with given \( Z \succeq 0 \). Assume for a moment that \( Z \succ 0 \), and let us pass in \((A)\) to new variables

\[ P = Z^{-1/2}UZ^{-1/2}, \quad Q = Z^{-1/2}VZ^{-1/2}. \]

We have

\[ U \succeq 0 \iff P \succeq 0, \quad V \succeq 0 \iff Q \succeq 0, \quad U + V = Z \iff P + Q = I \]

\[ \text{Tr}([U - V]B) = \text{Tr}(Z^{1/2}[P - Q]Z^{1/2}B) = \text{Tr}([P - Q] \underbrace{Z^{1/2}BZ^{1/2}}_{C}) \]

\[ \Rightarrow m(B, Z) = \max_P \{ \text{Tr}([2P - I]C) : 0 \preceq P \preceq I \} \]
\( \Rightarrow \text{representing } C = U \text{Diag}\{\lambda(C)\}U^T \text{ with orthogonal } U, \)

\[
m(B, Z) = \max_P \{ \text{Tr}([2P - I]C) : 0 \preceq P \preceq I \}
= \max_P \{ \text{Tr}(U^T[2P - I]U \text{Diag}\{\lambda(C)\}) : 0 \preceq P \preceq I \}
= \max_P \{ \text{Tr}([2PU^T - I]U \text{Diag}\{\lambda(C)\}) : 0 \preceq P \preceq I \}
= \max_P \{ \text{Tr}(U^T[2P - I]U \text{Diag}\{\lambda(C)\}) : 0 \preceq P \preceq I \}
= \max_R \{ \text{Tr}([2R - I] \text{Diag}\{\lambda(C)\}) : 0 \preceq R \preceq I \}
= \max_R \{ \sum_i \lambda_i(C)(2R_{ii} - 1) : 0 \preceq R \preceq I \}
= \sum_i |\lambda_i(C)|.
\]

By continuity arguments, the resulting equality (proved when \( Z \succ 0 \)) holds true for \( Z \succeq 0 \) as well.
\[ 0 < \max_{U, V, W} \left\{ \sum_{\ell} \text{Tr}([U_{\ell} - V_{\ell}] B_{\ell}) - \text{Tr}(WB_0) : \begin{array}{c} U_{\ell} + V_{\ell} = W, \ell = 1, \ldots, L \\ \text{Tr}(W) = U_{\ell}, V_{\ell}, W \succeq 0 \end{array} \right\} \]  

\[ \max_{U, V} \left\{ \text{Tr}([U - V] B) : \begin{array}{c} U, V \succeq 0 \\ U + V = Z \end{array} \right\} = \| \lambda(Z^{1/2} B Z^{1/2}) \|_1 \]

\[ \text{After optimization in } U_{\ell} \text{ and } V_{\ell}, \ (D) \text{ becomes} \]

\[ 0 < \max_{W \succeq 0} \left\{ \sum_{\ell} \rho \| \lambda(W^{1/2} B_{\ell} W^{1/2}) \|_1 - \text{Tr}(WB_0(: \text{Tr}(W) = 1) \right\}, \]

so that

\[ \rho \sum_{\ell=1}^{L} \| \lambda(W^{1/2} B_{\ell} W^{1/2}) \|_1 > \text{Tr}(W^{1/2} B_0 W^{1/2}) \]

for appropriately chosen \( W \succeq 0. \)
**Situation:** Assuming that \((S[\rho])\) has no solutions, there exists \(W \succeq 0\) such that
\[
\rho \sum_{\ell=1}^{L} \| \lambda(W^{1/2}B_{\ell}W^{1/2}) \|_1 > \text{Tr}(W^{1/2}B_{0}W^{1/2}).
\] (*)

**Step 3: Probabilistic interpretation of (*)**. Let \(\xi\) be the standard (zero mean, unit covariance matrix) Gaussian random vector in \(\mathbb{R}^m\), and \(A\) be a symmetric \(m \times m\) matrix of rank \(k\). What is the expectation of the modulus of the quadratic form \(\xi^T A \xi\)?

Representing \(A = U \text{Diag}\{\lambda\} U^T\) with orthogonal \(U\) and setting \(\eta = U^T \xi\), observe that the distribution of \(\eta\) is exactly the same as the one of \(\xi\); thus, our question becomes what is the expectation of
\[
\zeta = \left| \sum_{i=1}^{k} \lambda_i \eta_i^2 \right|
\]
where \(\eta_i \sim \mathcal{N}(0,1)\) are independent of each other. Common sense says that the expectation of \(\zeta\) is at least \(O(1)\|\lambda\|_2 \geq O(1)k^{-1/2}\|\lambda\|_1\). Specifically, setting \(\vartheta(k) = \frac{1}{\min \left\{ \int \left| \sum_{i=1}^{k} \lambda_i \eta_i^2 \right| (2\pi)^{-k/2} e^{-\eta_1^2-\cdots-\eta_k^2/2} d\eta_1 \cdots d\eta_k : \|\lambda\|_1 = 1 \right\}}\)

one can easily verify that
\[
\vartheta(1) = 1, \quad \vartheta(2) = \frac{\pi}{2}, \quad \vartheta(k) \leq \frac{\pi \sqrt{k}}{2},
\]
while by definition of \(\vartheta(\cdot)\) one has
\[
\vartheta(\text{Rank}(A)) \mathbb{E} \{\|\xi^T A \xi\|\} \geq \|\lambda(A)\|_1
\]
for every symmetric matrix \(A\).

3.102
Situation: Assuming that \((S[\rho])\) has no solutions, there exists \(W \succeq 0\) such that
\[
\rho \sum_{\ell=1}^{L} \| \lambda(W^{1/2}B_{\ell}W^{1/2}) \|_1 > \text{Tr}(W^{1/2}B_{0}W^{1/2}).
\] (\(\ast\))

Besides this, we have seen that with properly chosen function \(\vartheta(\cdot)\) such that
\[
\vartheta(1) = 1, \quad \vartheta(2) = \frac{\pi}{2}, \quad \vartheta(k) \leq \frac{\pi \sqrt{k}}{2},
\]
for standard Gaussian vector \(\xi\) and every symmetric matrix \(A\) one has
\[
\vartheta(\text{Rank}(A)) \mathbb{E} \left\{ |\xi^T A \xi| \right\} \geq \| \lambda(A) \|_1
\] (\(\ast\ast\))

- Let \(\xi \sim \mathcal{N}(0,I_m)\) and let \(\mu = \max_{\ell \geq 1} \text{Rank}(B_{\ell})\). We have
\[
\mathbb{E} \left\{ \rho \sum_{\ell=1}^{k} \vartheta(\mu) |\xi^T W^{1/2}B_{\ell}W^{1/2} \xi| \right\} \geq \rho \sum_{\ell=1}^{L} \| \lambda(W^{1/2}B_{\ell}W^{1/2}) \|_1 \quad \text{[by (\(\ast\ast\))]} \nonumber
\]
\[
> \text{Tr}(W^{1/2}B_{0}W^{1/2}) \quad \text{[by (\(\ast\))]} = \mathbb{E} \left\{ \xi^T W^{1/2}B_{0}W^{1/2} \xi \right\} \quad \text{[evident]}
\]
Thus,
\[
\mathbb{E} \left\{ \xi^T W^{1/2}B_{0}W^{1/2} \xi - \xi \rho \vartheta(\mu) \sum_{\ell=1}^{k} |\xi^T W^{1/2}B_{\ell}W^{1/2} \xi| \right\} < 0.
\]

It follows that there exists \(\eta = W^{1/2} \xi\) such that
\[
\eta^T B_{0} \eta - \rho \vartheta(\mu) \sum_{\ell=1}^{k} |\eta^T B_{\ell} \eta| < 0
\]
Setting \(u_{\ell} = -\rho \vartheta(\mu) \text{sign}(\eta^T B_{\ell} \eta)\), we get
\[
\|u\|_{\infty} = \rho \vartheta(\mu) \quad \text{and} \quad \eta^T \left[ B_{0} + \sum_{\ell} u_{\ell} B_{\ell} \right] \eta < 0,
\]
i.e.,
\[
A[\vartheta(\mu) \rho] \not\in S^m_+.
\]

3.103
F. Robust Conic Quadratic Programming. Consider a c.q.i.

\[ \|Ax + b\|_2 \leq c^T x + d \tag{CQI} \]

and assume that the data \((A, b, c, d)\) of this c.q.i. is not known exactly and run through a given uncertainty set \(U\).

How to process the Robust Counterpart

\[ \|Ax + b\|_2 \leq c^T x + d \quad \forall (A, b, c, d) \in U, \tag{RC} \]

of (CQI)?

Assume that

- the uncertainty is *side-wise*: the left hand side data \((A, b)\) and the right hand side data \((c, d)\) run, independently of each other, through the respective uncertainty sets \(U_{\text{left}}\), \(U_{\text{right}}\);
- the set \(U_{\text{right}}\) is given by a strictly feasible SDR;
- the left hand side in (CQI) is affected by “ellipsoidal” uncertainty:

\[
U_{\text{left}} = U_{\rho}^{\text{left}} = \left\{ [A, b] = [A^*, b^*] + \sum_{\ell} u_{\ell} [A^\ell, b^\ell] : u^T S_j u \leq \rho^2, \quad 1 \leq j \leq J \right\},
\]

where \(S_j \succeq 0\), \(\sum_j S_j > 0\).
- With these assumptions, it still can be NP-hard to check whether a given \(x\) is feasible for (RC). However, it turns out that (RC) admits a tight SDP approximation.

3.104
\[ \|Ax + b\|_2 \leq c^T x + d \quad \forall \left( [A, b] \in U^{\text{left}}_\rho, (c, d) \in U^{\text{right}} \right) \]  
(\text{RC}[\rho])

\[ U^{\text{left}} = \left\{ [A, b] = [A^*, b^*] + \sum_{\ell} u_\ell [A^\ell, b^\ell] : u^T S_j u \leq \rho^2, 1 \leq j \leq J \right\} \]

Theorem [Ben-Tal, Nemirovski, Roos '01] The semi-infinite conic quadratic inequality (RC[\rho]) admits a tractable approximation, which is certain explicit system (S[\rho]) of LMIs in original design variables \( x \) and additional variables \( u \). The size of (S[\rho]) is polynomial in the size of the data of (RC[\rho]), and the relation between (RC[\rho]) and (S[\rho]) are as follows:

(i) If \( x \) can be extended to a feasible solution of (S[\rho]), then \( x \) is feasible for (RC[\rho]).

(ii) If \( x \) cannot be extended to a feasible solution of (S[\rho]), then \( x \) is not feasible for (RC[\Omega[\rho]])

where the “tightness factor” \( \Omega \) is as follows:

- in the case of \( J = 1 \) (“simple ellipsoidal uncertainty”), \( \Omega = 1 \), i.e., (S[\rho]) is equivalent to (RC[\rho]) (easily follows from S-Lemma);
- in the case of box uncertainty:
  \[ u^T S_j u \leq \rho^2, j = 1, \ldots, J \Leftrightarrow u_j^2 \leq \rho^2, 1 \leq j \leq \dim u, \]

\[ \Omega = \frac{\pi}{2} \] (easily follows from Matrix Box Theorem);

- in general, \( \Omega \leq \sqrt{2 \ln \left( 6 \sum_j \text{Rank}(S_j) \right)} \) (easily follows from “Approximate S-Lemma”).

Note that \( \Omega \leq 6 \), provided that the total rank of \( S_j \) is \( \leq 65,000,000 \).
Proof of $S$-Lemma

$S$-Lemma: Let $A, B$ be symmetric $m \times m$ matrices such that $\bar{x}^T A \bar{x} > 0$ for certain $\bar{x}$. Then the implication

$$\forall x : x^T A x \geq 0 \Rightarrow x^T B x \geq 0$$

(\ast)

holds true iff

$$\exists \lambda \geq 0 : B \succeq \lambda A$$

(\ast\ast)

• $(\ast\ast) \Rightarrow (\ast)$: evident.
• $(\ast) \Rightarrow (\ast\ast)$: Consider the following “relaxation” of $(\ast)$:

$$\forall (X \succeq 0) : \quad \text{Tr}(XA) \geq 0 \Rightarrow \text{Tr}(XB) \geq 0$$

(R)

Step 1: Under the premise of $S$-Lemma, (R) is equivalent to $(\ast\ast)$. Indeed, under the premise of $S$-Lemma, the semidefinite program

$$\min_X \{ \text{Tr}(BX) : X \succeq 0, \text{Tr}(AX) \geq 0 \}$$

is strictly feasible, and (R) just says that the optimal value in this problem (which is either 0, or $-\infty$) is 0. Applying Conic Duality Theorem, this is the case iff the dual problem

$$\max_{\lambda, S} \{ 0 : B = \lambda A + S, S \succeq 0, \lambda \geq 0 \}$$

is feasible, i.e., iff $(\ast\ast)$ takes place.

• Thus, to complete the proof of $S$-Lemma, it suffices to verify that

$$(\ast) \Rightarrow (R).$$
∀x : x^T Ax ≥ 0 ⇒ x^T Bx ≥ 0

∀(X ⪰ 0) : Tr(XA) ≥ 0 ⇒ Tr(XB) ≥ 0  \quad (R)

Goal: to prove that (\ast) ⇒ (R).

Proof: Assume that (\ast) takes place and that X ⪰ 0 is such that Tr(AX) ≥ 0; we should prove that then Tr(BX) ≥ 0 as well.

Let us set

\[ \bar{A} \equiv X^{1/2}AX^{1/2} = U\text{Diag}\{\lambda\}U^T, \quad \eta = X^{1/2}U\xi, \]

where \( \xi \) is a random vector with independent coordinates taking values ±1 with probabilities 1/2. We have

\[ \eta^T A \eta = \xi^T U^T X^{1/2}AX^{1/2}U\xi = \xi^T \text{Diag}\{\lambda\} \xi = \text{Tr}(\text{Diag}\{\lambda\}) = \text{Tr}(X^{1/2}AX^{1/2}) = \text{Tr}(AX) \geq 0 \]

\[ \Downarrow \quad (\ast) \]

\[ \eta^T B \eta \geq 0 \]

\[ \Downarrow \]

\[ 0 \leq \mathbb{E}\{\eta^T B \eta\} = \mathbb{E}\{\xi^T U^T X^{1/2}BX^{1/2}U\xi\} = \text{Tr}(U^T X^{1/2}BX^{1/2}U) = \text{Tr}(X^{1/2}BX^{1/2}) = \text{Tr}(BX) \]

Q.E.D.
Approximate $S$-Lemma

Let $Q_1,\ldots,Q_L$ be positive semidefinite matrices with positive definite sum, let $A$ be a symmetric matrix, and let $a$ be a vector. Let

$$\text{Opt}(\rho) = \max_x \left\{ x^T Ax + 2a^T x : x^T Q_\ell x \leq \rho^2, \ell \leq L \right\}$$

In general, computing $\text{Opt}(\rho)$ is NP-hard. However, we can use Semidefinite Relaxation scheme to bound $\text{Opt}(\rho)$ from above:

$$\text{Opt}(\rho) = \max \left\{ x^T Ax + 2a^T x : x^T Q_\ell x \leq \rho^2, \ell \leq L \right\} = \max \left\{ x^T Ax + 2ta^T x : x^T Q_\ell x \leq \rho^2, \ell \leq L, t^2 \leq 1 \right\}$$

$$\leq \max_Y \left\{ \begin{array}{c} \text{Tr} \left( \begin{bmatrix} A & a^T \\ a & R \end{bmatrix} Y \right) \\ \text{Tr} \left( \begin{bmatrix} Q_\ell \\ I \end{bmatrix} Y \right) \leq \rho^2, \ell \leq L \\ \text{Tr} \left( \begin{bmatrix} I \\ R_{a=dd^T} \end{bmatrix} Y \right) \leq 1, Y \succeq 0 \end{array} \right\} \equiv \text{SDP}(\rho)$$

Approximate $S$-Lemma. One has

$$\text{Opt}(\rho) \leq \text{SDP}(\rho) \leq \text{Opt}(\Omega \rho), \quad \Omega = \sqrt{2 \ln \left( 6 \sum_{\ell=1}^L \text{Rank}(Q_\ell) \right)}.$$
\[
\text{Opt}(\rho) = \max \left\{ x^T Ax + 2a^T x : x^T Q_\ell x \leq \rho^2, \ell \leq L \right\} \\
\leq \max \left\{ \text{Tr} \left( \begin{bmatrix} A & a^T \\ R \end{bmatrix} Y \right) : \text{Tr} \left( \begin{bmatrix} Q_\ell \\ R_\ell \end{bmatrix} Y \right) \leq \rho^2, \ell \leq L, \text{Tr} \left( \begin{bmatrix} 1 \\ R_0=dd^T \end{bmatrix} Y \right) \leq 1, Y \succeq 0 \right\} \equiv \text{SDP}(\rho)
\]

Approximate S-Lemma. One has

\[
\text{Opt}(\rho) \leq \text{SDP}(\rho) \leq \text{Opt}(\Omega \rho), \quad \Omega = \sqrt{2 \ln \left( 6 \sum_{\ell=1}^{L} \text{Rank}(Q_\ell) \right)}.
\]

Proof of upper bound: From \( Q_\ell \succeq 0, \sum_\ell Q_\ell > 0 \) it follows that \( R_0 + \sum_{\ell=1}^{L} R_\ell > 0 \), so that the feasible set of the SDP program in (1) is nonempty and bounded. Thus, the SDP program in (1) is solvable. Let \( Y_\ast \) be its optimal solution, and let

\[
Y_\ast^{1/2} R Y_\ast^{1/2} = U \text{Diag}\{\lambda\} U^T.
\]

Let, further, \( \eta = Y_\ast^{1/2} U \xi \), where \( \xi \) is random vector with independent entries taking values \( \pm 1 \) with probabilities 1/2. Then

\[
\begin{align*}
\text{SDP}(\rho) &= \text{Tr}(RY_\ast) \& \text{Tr}(R_0 Y_\ast) \leq 1 \& \text{Tr}(R_\ell Y_\ast) \leq \rho^2, 1 \leq \ell \leq L \& Y_\ast \succeq 0 \\
\text{Opt}(\rho) &= \max \{ z^T R z : z^T R_0 z \leq 1, z^T R_\ell z \leq \rho^2, \ell = 1, ..., L \}
\end{align*}
\]

\[
\begin{align*}
\mathcal{R} &:= Y_\ast^{1/2} R Y_\ast^{1/2} = U \text{Diag}\{\lambda\} U^T \\
\eta &:= Y_\ast^{1/2} U \xi, \xi_1, ..., \xi_m \in \{-1; 1\} \text{ i.i.d., } \text{Prob}\{\xi_i = 1\} = 1/2
\end{align*}
\]
\[
\text{SDP}(\rho) = \text{Tr}(RY_*) \& \text{Tr}(R_0Y_*) \leq 1 \& \text{Tr}(R_\ell Y_*) \leq \rho^2, 1 \leq \ell \leq L \& Y_* \succeq 0
\]

\[
\text{Opt}(\rho) = \max_{z=(x,t)} \{ z^T Rz : z^T R_0z \leq 1, z^T R_\ell z \leq \rho^2, \ell = 1,...,L \}
\]

\[
\bar{R} := Y_*^{1/2} RY_*^{1/2} = U \text{Diag}\{\lambda\} U^T
\]

\[
\eta := Y_*^{1/2} U \xi, \xi_1,...,\xi_m \in \{-1;1\} \text{ i.i.d.}, \text{Prob}\{\xi_i = 1\} = 1/2
\]

We have

\[
\eta^T R\eta = \xi^T U^T Y_*^{1/2} RY_*^{1/2} U \xi = \xi^T \text{Diag}\{\lambda\} \xi = \text{Tr}(\text{Diag}\{\lambda\}) = \text{Tr}(U \text{Diag}\{\lambda\} U^T) = \text{Tr}(Y_*^{1/2} RY_*^{1/2}) = \text{Tr}(R_* Y_*) = \text{SDP}(\rho),
\]

\[
E \{ \eta^T R_\ell \eta \} = E \{ \xi^T U^T Y_*^{1/2} R_\ell Y_*^{1/2} U \xi \} = \text{Tr}(U^T Y_*^{1/2} R_\ell Y_*^{1/2} U) = \text{Tr}(R_\ell Y_*) \leq \begin{cases} 1, & \ell = 0 \\ \rho^2, & \ell = 1,...,L \end{cases}
\]

**Lemma:** One has \( \text{Prob}\{\eta^T R_0 \eta \leq 1\} \geq \frac{1}{3} \).

**Proof:** We have \( R_0 = dd^T \) and therefore

\[
\eta^T R_0 \eta = \xi^T \underbrace{U^T Y_*^{1/2} d}_h h^T \xi = |h^T \xi|^2.
\]

Besides this,

\[
\|h\|_2^2 = E \{ |h^T \xi|^2 \} = E \{ \eta^T R_0 \eta \} \leq 1.
\]

It is easily seen that when \( h \) is a deterministic vector with \( \|h\|_2 \leq 1 \) and \( \xi \) is the above random vector, then

\[
\text{Prob}\{|h^T \xi| \leq 1\} \geq O(1).
\]

A more advanced reasoning shows that one can take \( O(1) = \frac{1}{3} \).
Situation:

\[
\text{Opt}(\rho) = \max_{z=(x,t)} \left\{ z^T R_z : z^T R_0 z \leq 1, z^T R_\ell z \leq \rho^2, 1 \leq \ell \leq L \right\}
\]

\[\eta^T R_\eta \equiv \text{SDP}(\rho)\]

\[
\eta: \text{random solution to (a) such that}
\]

\[
\begin{align*}
\eta^T R_0 \eta \leq 1 & \quad \text{Prob}\left\{ \eta^T R_0 \eta \leq 1 \right\} \geq \frac{1}{3} \\
\eta^T R_\ell \eta = \xi^T U^T Y_{\ell}^{1/2} R_\ell Y_{\ell}^{1/2} U \xi, 1 \leq \ell \leq L & \quad \text{Prob}\left\{ \eta^T R_\ell \eta \leq \rho^2, 1 \leq \ell \leq L \right\} \\
E\left\{ \eta^T R_\ell \eta \right\} \leq \rho^2, 1 \leq \ell \leq L & 
\end{align*}
\]

Representing \( S_\ell = \sum_{j=1}^{\text{Rank}(Q_\ell)} a_{\ell j} a_{\ell j}^T \), we have

\[
\sum_j \|a_{\ell j}\|_2^2 = \text{Tr}(S_\ell) = E\left\{ \xi^T S_\ell \xi \right\} = E\left\{ \eta^T R_\ell \eta \right\} \leq \rho^2,
\]

\[
\Rightarrow \text{Prob}\left\{ \eta^T R_\ell \eta > \theta \rho^2 \right\} = \text{Prob}\left\{ \xi^T S_\ell \xi > \theta \rho^2 \right\} \leq \text{Prob}\left\{ \sum_j (a_{\ell j}^T \xi)^2 > \theta \sum_j \|a_{\ell j}\|_2^2 \right\}
\]

\[
\leq \sum_j \text{Prob}\left\{ (a_{\ell j}^T \xi)^2 > \theta \|a_{\ell j}\|_2^2 \right\} < 2\text{Rank}(Q_\ell) \exp\{-\theta/2\}.
\]

Setting \( K = \sum_{\ell=1}^L \text{Rank}(Q_\ell) \) and \( \theta = 2 \ln(6K) \), we conclude that

\[
\text{Prob}\left\{ \exists \ell \in \{1, 2, ..., L\} : \eta^T R_\ell \eta > \theta \rho^2 \right\} < \frac{1}{3}.
\]

Taking into account that \( \text{Prob}\left\{ \eta^T R_0 \eta \leq 1 \right\} \geq \frac{1}{3} \), we arrive at

\[
\exists \bar{\eta}: \bar{\eta}^T R \bar{\eta} = \text{SDP}(\rho), \bar{\eta}^T R_0 \bar{\eta} \leq 1, \bar{\eta}^T R_\ell \bar{\eta} \leq \theta \rho^2, \ell = 1, ..., L.
\]

We see that \( \bar{\eta} \) is a feasible solution of (a) with \( \rho \) increased to \( \sqrt{\theta} \rho \), whence

\[
\text{SDP}(\rho) = \bar{\eta}^T R \bar{\eta} \leq \text{Opt}(\Omega \rho), \Omega \equiv \sqrt{\theta} = \sqrt{2 \ln \left( 6 \sum_{\ell=1}^L \text{Rank}(Q_\ell) \right)}
\]
Extremal Ellipsoids

A ellipsoid in $\mathbb{R}^n$ is, by definition, the image of the unit Euclidean ball

$$B_n = \{ u \in \mathbb{R}^n : u^T u \leq 1 \}$$

under an affine mapping $u \mapsto Au + a$:

$$E = \{ x = Au + a : u^T u \leq 1 \}.$$  \hfill (\star)

**Note:**
- An ellipsoid is a convex compact set symmetric w.r.t. $a$. Consequently, *The center $a$ of an ellipsoid $E$ is uniquely defined by the set $E$.*
- An ellipsoid $E$ is “full-dimensional”, that is, possesses a nonempty interior, iff $A$ in (\star) is nonsingular.
- Matrix $A$ in (\star) is *not* uniquely defined by $E$; replacing in (\star) $A$ with $AU$, where $U$ is orthogonal, we preserve the right hand side set. In particular, *Among the matrices $A$ participating in representations of a given ellipsoid $E$, there exists a positive semidefinite one, which is uniquely defined by the set $E.*

3.112
\[ E = \{ x = Au + a : u^T u \leq 1 \}. \] (\textup{(*)})

\footnote{Bottom line: If a set \( E \subset \mathbb{E}^n \) is an ellipsoid, that is, admits a representation (\textup{(*)}), then \( E \) admits a representation (\textup{(*)}) with \( A \succeq 0 \). In this image representation of \( E \), both \( A \succeq 0 \) and \( a \) are uniquely defined by the set \( E \).

• An ellipsoid with image representation given by matrix \( A \succeq 0 \) and vector \( a \) will be denoted \( \mathcal{E}(A, a) \):

\[ \mathcal{E}(A, a) = \{ Au + a : u^T u \leq 1 \} \subset \mathbb{R}^n \] \[ [A \in \mathbb{S}^n_+, a \in \mathbb{R}^n] \]
Consider a quadratic form
\[ f(x) = x^T P x - 2 p^T x \quad (f) \]
on \( \mathbb{R}^n \). This form is below bounded if and only if the following two conditions hold:
- The form is convex: \( P \succeq 0 \)
- The Fermat equation
\[ \nabla f(x) = 0 \iff Px = p \quad (F) \]
has a solution \( x_* \).

In particular, if \( f(\cdot) \) is below bounded, then there exists a representation
\[ f(x) = x^T B^2 x - 2 b^T B x, \quad (*) \]
where \( B \succeq 0 \) and \( b \in \text{Im} B \). Indeed, in the case of 1), 2) one can set \( B = P^{1/2}, \; b = P^{1/2} x_* \).

Vise versa, if \( f(\cdot) \) can be represented in the form \((*)\) with \( B \succeq 0 \) and \( b \in \text{Im} B \), then 1), 2) hold true, so that below boundedness of \( f \) is equivalent to the possibility to represent \( f \) by \((*)\) with \( B \succeq 0, \; b \in \text{Im} B \).
A below bounded quadratic form $f(x)$ can be represented as

$$f(x) = x^T B^2 x - 2b^T B x$$

$$[B \succeq 0, b \in \text{Im} B] \quad (\ast)$$

Note that *Form (\ast) attains its minimum, which is equal to $-b^T b$.* Indeed, relation $b \in \text{Im} B$ means that $b = Bx_*$ for certain $x_*$. Then

$$\nabla f(x_*) = 2B^2 x_* - 2Bb = 2B^2 x_* - 2B^2 x_* = 0$$

that is, $x_*$ is a critical point and thus a minimizer of the convex function $f$. We have

$$f(x_*) = \underbrace{(Bx_*)^T (Bx_*) - 2b^T B x_*}_{\text{b}} = -b^T B x_* = -b^T b.$$

Let $f$ be a below bounded quadratic form on $\mathbb{R}^n$, and let $f_*$ be its minimum value. The "nontrivial" levels sets of $f$, that is, level sets of the form

$$C = \{x : f(x) \leq f_* + r^2\} \quad [r > 0] \quad (C)$$

are called "elliptic cylinders".
\[ C = \{x : f(x) \leq f_* + r^2\} \quad [r > 0] \] (C)

♠ In representation (\(*\)), an elliptic cylinder is

\[ C = \{x : \|Bx - b\|_2^2 \leq r^2\} \]

When \(\theta > 0\), the data \((B, b, r)\) and \((\theta B, \theta b, \theta r)\) define the same cylinder, so that by normalization we may assume that \(r = 1\). The representation

\[ C = \{x : \|b - Bx\|_2^2 \leq 1\} \quad [B \succeq 0, b \in \text{Im}B] \]

is called inequality representation of elliptic cylinder. The data \(B, b\) of this representation are uniquely defined by the set \(C\).
\[ C = \{ x : \| b - Bx \|_2^2 \leq 1 \} \quad [B \succeq 0, b \in \text{Im}B] \]

- \( C \) is bounded iff \( B \succ 0 \), and iff \( C \) is a full-dimensional ellipsoid. Indeed,
- We clearly have \( C = C + \text{Ker}B \). Thus, if \( C \) is bounded, then \( \text{Ker}B = \{0\} \), that is, \( B \succ 0 \). Vice versa, if \( B \succ 0 \), then \( C \) clearly is bounded.
- We have
  \[ B \succ 0 \Rightarrow \{ x : \| Bx - b \|_2^2 \leq 1 \} = \{ x = B^{-1}u + B^{-1}b : u^Tu \leq 1 \} \]
  \[ A \succ 0 \Rightarrow \{ x = Au + a : u^Tu \leq 1 \} = \{ x : \| A^{-1}x - A^{-1}a \|_2^2 \leq 1 \} \]
- When \( B \preceq 0 \) is degenerate, the elliptic cylinder \( C \) can be represented as the sum of the set
  \[ C_0 = \{ x \in \text{Im}B : \| b - Bx \|_2^2 \leq 1 \} \]
  (which is a full-dimensional ellipsoid in the subspace \( \text{Im}B = (\text{Ker}B)^\perp \)) and the linear subspace \( \text{Ker}B \).

3.117
**Bottom line:** We have defined

- Ellipsoids in $\mathbb{R}^n$ – sets representable as
  \[ E = \mathcal{E}(A, a) \equiv \{ x = Au + a : u^Tu \leq 1 \}, \quad (E) \]
  where $A \succeq 0$. The data $A, a$ of this *image representation* of $E$ are uniquely defined by the set $E$ itself.

  Ellipsoid $E$ is full-dimensional (that is, $\text{int } E \neq \emptyset$) if and only if $A \succ 0$, otherwise the ellipsoid is “flat” – it is contained in the plane $a + \text{Im } A$, which is a proper affine subspace of $\mathbb{R}^n$.

- Elliptic cylinders in $\mathbb{R}^n$ – sets representable as
  \[ C = \mathcal{C}(B, b) \equiv \{ x : \| Bx - b \|_2^2 \leq 1 \} \quad (C) \]
  where $B \succeq 0$ and $b \in \text{Im } B$. The data $B, b$ of this *inequality representation* of $C$ are uniquely defined by the set $C$ itself.

  Elliptic cylinder $C$ is bounded if and only if $B \succ 0$, and in this case $C$ is just a full-dimensional ellipsoid, otherwise $C$ is the sum of the kernel of $B$ and a full-dimensional ellipsoid in the image space of $B$.  

3.118
• Full-dimensional ellipsoids $E$ admit both image and inequality representations:

$$A \succ 0 \Rightarrow E \equiv \{ x = Au + a : u^T u \leq 1 \} = \{ x : \|Bx - b\|^2 \leq 1 \}$$

with the parameters of the representations linked by the relations

$$B = A^{-1} \iff A = B^{-1}$$
$$b = A^{-1} a \iff a = B^{-1} b$$
Volume of an Ellipsoid

Under affine transformation

\[ x \mapsto Ax + a : \mathbb{R}^n \to \mathbb{R}^n, \]

\( n \)-dimensional volumes of sets are multiplied by \(|\text{Det}(A)|:\)

\[ \text{Vol}(\{y = Ax + a : x \in U\}) = |\text{Det}(A)|\text{Vol}(U). \]

In particular, \( \text{The volume of ellipsoid } \mathcal{E}(A,a) \text{ is } \text{Det}(A) \text{ times the volume of the unit Euclidean ball in } \mathbb{R}^n. \)

In what follows, it is convenient to choose, as the unit of volume in \( \mathbb{R}^n \), the volume of the unit Euclidean ball rather than the volume of the unit cube. With this convention, \( \text{The volume of ellipsoid } \mathcal{E}(A,a) \text{ is } \text{Det}(A), \text{ and the volume of full-dimensional ellipsoid } \mathcal{C}(B,b) \text{ is } \)

\[ \frac{1}{\text{Det}(B)}. \]
Half-Axes of an Ellipsoid

Let $E = \mathcal{E}(A, a)$, let $e_i$ be the orthonormal eigenbasis of $A$, and $\lambda_i$ be the corresponding eigenvalues. Let $\xi_i(x)$ be the coordinates of $x$ in the basis $e_1, \ldots, e_n$. The fact that $x = Au + a$ is equivalent to the relations

$$\xi_i(x) - \xi_i(a) = \lambda_i \xi_i(u),$$

so that the fact that $x \in E$ is equivalent to

$$\sum_i \frac{(\xi_i(x) - \xi_i(a))^2}{\lambda_i^2} \leq 1$$

Geometrically: $\lambda_i$ are the half-axes $\chi_i(E)$ of $E$, and $e_i$ are the directions of the principal axes of $E$.

For a full-dimensional ellipsoid $E = \mathcal{E}(A, a)$, all half-axes $\chi_i(E) \equiv \lambda_i(A)$ are positive. In terms of the inequality representation $E = \mathcal{C}(B, b)$ of the ellipsoid, the half-axes are

$$\chi_i(E) = \lambda_i^{-1}(B).$$
In the case of degenerate $B$, elliptic cylinder $C = C(B, b)$ is the sum of an ellipsoid $C_0$ in the subspace $\text{Im}B$ and the linear subspace $\text{Ker}B$ which is orthogonal to $C_0$. It makes sense to define the first $\text{Rank}(B)$ half-axes of $C$ as $\chi_i(C) = \lambda_i^{-1}(B)$, where $\lambda_i(B), i = 1, ..., \text{Rank}(B)$, are the nonzero eigenvalues of $B$, and the remaining $n - \text{Rank}(B)$ half-axes of $C$ as $+\infty$. 

3.122
The basic problems on extremal ellipsoids are as follows:

**Outer Approximation:** (O): Given a bounded nonempty set $X \subset \mathbb{R}^n$, find the “smallest” ellipsoid containing $X$.

**Inner Approximation:** (I): Given a nonempty set $X \subset \mathbb{R}^n$, find the “largest” ellipsoid contained in $X$.

In these problems, the “size” of an ellipsoid is an appropriate symmetric function of the half-axes, e.g.

- $\chi_1 \chi_2 \ldots \chi_n$ (the volume),
- $\max_i \chi_i$ (the radius of the smallest circumscribed ball),
- $\min_i \chi_i$ (the radius of the largest inscribed ball),
- $\sum_i \chi_i^\alpha$,
- ...
Extremal ellipsoids have numerous applications, including
- “optimal” methods of Nonsmooth Convex Optimization,
- identification and estimation in Control
- accurate integration of ordinary differential equations,
- ...

**Example 1: Inscribed Ellipsoid Method.** Theoretically optimal, in certain precise sense, method for solving to high accuracy a general nonsmooth Convex Programming program

\[
\min_X f(x)
\]

\((X \text{ is a convex polytope given by linear inequalities, } f \text{ is convex and continuous on } X)\) is the **Inscribed Ellipsoid Method.** At every step of this method, one should solve an auxiliary problem of the form **Find the largest volume ellipsoid contained in a polytope given by a list of linear inequalities.**
**Example 2: Estimation in Dynamical System.** Consider a Discrete Time Linear Dynamical System:

\[
\begin{align*}
    z(t+1) &= Az(t) \\
    y(t) &= Cz(t) + \xi_t
\end{align*}
\]

where

- \( z(t) \) is the state at time \( t \),
- \( y(t) \) is the observation at time \( t \),
- \( \xi_t \) is norm-bounded observation error: \( \|\xi_t\|_2 \leq \rho \),
- \( A \) and \( C \) are known matrices.

**Example:** \( z(t) \) is the position \( x(t) \) and the velocity \( v(t) \) of a plane flying at (unknown) constant velocity, and \( y(t) \) are the observations of the position of the plane coming from a radar:

\[
\begin{bmatrix}
    x(t+1) \\
    v(t+1) \\
    y(t)
\end{bmatrix}
\begin{bmatrix}
    I_3 & I_3 \\
    I_3 & I_3
\end{bmatrix}
\begin{bmatrix}
    x(t) \\
    v(t)
\end{bmatrix}
\]

\[
y(t) = x(t) + \xi_t
\]
\[ z(t + 1) = A z(t) \]
\[ y(t) = C z(t) + \xi_t \]

Since the dynamics is known, all we need to identify the motion is the initial state \( z(0) \). Some information on \( z(0) \) is contained in observations \( y(t) \): given \( y(t) \), we know that \( z(0) \) belongs to the elliptic cylinder

\[ C_t = \{ z : \| C A^t z - y(t) \|_2^2 \leq \rho^2 \}, \]

and all we know at time \( T \) is that \( z(0) \) belongs to the set

\[ C^T = \bigcap_{t=0}^T C_t. \]

We may now want to build an estimate of \( z(0) \) as the center of the smallest ball containing the set \( C^T \), which is the Outer Ellipsoidal Approximation problem where you are interested to minimize the maximal half-axis of the circumscribed ellipsoid.
Example 3: Approximating reachable sets. Consider a controlled Discrete Time Linear Dynamical System:

\[
z(t+1) = A_t z(t) + B_t u(t) + f_t, \; z(0) = z_0
\]  

- \( z(t) \): states; 
- \( u(t) \): controls; 
- \( f_t \): known inputs; 
- \( A_t, B_t \): known matrices.

Assume that the control \( u(t) \) is bounded:

\[
\|u(t)\|_2 \leq \rho_t.
\]  

The reachable set \( Z^T \) of system (1) – (2) at time \( T \) is the set of all possible states \( z \) of the system at time \( t \):

\[
Z^T = \{ z : \exists \{ u(t), \|u(t)\|_2 \leq \rho_t \}_{t=0}^{T-1} : z(T) = z \}.
\]
\[ Z^T = \{ z : \exists \{ u(t), \| u(t) \|_2 \leq \rho_t \}_{t=0}^{T-1} : z(T) = z \}. \]

**Note:**

- \( Z^T \) is “computationally tractable”; e.g., to optimize a linear form \( c^T z \) over \( Z^T \) is the same as to solve the conic quadratic problem

\[
\min_{u(0),\ldots,u(T-1)} \left\{ c^T z(T) : \begin{array}{l}
z(t + 1) = A_t z(t) + B_t u(t) + f_t, \; 0 \leq t < T \\
\| u(t) \|_2 \leq \rho_t, \; 0 \leq t < T, \; z(0) = z_0
\end{array} \right\}
\]

- \( Z^T \) is the arithmetic sum of \( T \) ellipsoids:

\[
z(T) = z_0(T) + \sum_{\tau=0}^{T-1} A_T A_{T-1} \cdots A_{\tau+1} B_{T, \tau} u(\tau),
\]

where \( z_0(\cdot) \) is the trajectory of (1) corresponding to \( u(\cdot) \equiv 0. \Rightarrow \)

\[
Z^T = z_0(T) + \sum_{\tau=0}^{T-1} B_{T, \tau} \{ u : u^T u \leq \rho_t^2 \}.
\]

The reachable set \( Z^T \), while computationally tractable, becomes more and more complicated as \( T \) grows. *In many applications it makes sense to look for simple – ellipsoidal – inner and outer approximations of \( Z^T \).*
Observation O.1: Let $X \subset \mathbb{R}^n$ be a nonempty compact set. Then the set $\mathcal{X}$ of parameters $B, b$ of inequality representations of elliptic cylinders containing $X$ is convex.

To prove that $\mathcal{X}$ is convex, let $\lambda \in [0, 1]$, $(B, b), (C, c) \in \mathcal{X}$, so that $B \succ 0$, $C \succ 0$ and

$$\forall x \in X : \begin{cases} \|Bx - b\|_2 \leq 1 & [b \in \text{Im} B] \\ \|Cx - c\|_2 \leq 1 & [c \in \text{Im} C] \end{cases} \quad (\ast)$$

we should prove that $(D, d) = \lambda (B, b) + (1 - \lambda)(C, c) \in \mathcal{X}$. There is nothing to prove when $\lambda = 0$ or $\lambda = 1$, thus let $0 < \lambda < 1$. From $(\ast)$ and Triangle inequality we get

$$\forall x \in X : \|Dx - d\|_2 \leq \lambda \|Bx - b\|_2 + (1 - \lambda)\|Cx - c\|_2 \leq 1;$$

thus, all we need is to verify that $d \in \text{Im} D$. 

3.129
Situation:

\[ \lambda \in (0,1) \& (D,d) = \lambda(B,b) + (1 - \lambda)(C,c) \]

Claim: \( d \in \text{Im}D \).

Mini-Lemma: Let \( A_i \succeq 0 \) and \( \lambda_i > 0 \), \( i = 1,..,K \), and let \( A = \sum_i \lambda_i A_i \).

Then

\[ \text{Ker}A = \bigcap_i \text{Ker}A_i \quad (a); \quad \text{Im}A = \text{Im}A_1 + ... + \text{Im}A_K \quad (b) \]

Proof: For \( C \succeq 0 \), one has \( \text{Ker}C = \{x : x^T C x = 0\} \). Since \( \lambda_i > 0 \) and \( A_i \succeq 0 \), it follows that \( x^T A x = 0 \) iff \( x^T A_i x = 0 \) for all \( i \), which gives \((a)\). \((b)\) is equivalent to \((a)\) by elementary Linear Algebra. \( \square \)

Since \( 0 < \lambda < 1 \), both \( B \succeq 0 \) and \( C \succeq 0 \) enter the expression \( D = \lambda B + (1 - \lambda)C \) with positive weights. By MiniLemma, it follows that \( \text{Im}D = \text{Im}B + \text{Im}C \), whence \( d = \lambda b + (1 - \lambda)c \in \text{Im}D \) due to \( b \in \text{Im}B \), \( c \in \text{Im}C \). \( \square \)
Observation O.2: “Typical sizes” of full-dimensional ellipsoids $E$ are convex (and thus easy-to-minimize) functions of the parameters $B, b$ of the inequality representation of $E$. This is so, e.g., for the sizes
- $\text{Vol}(E) = \prod_i \chi_i(E)$ (volume)
- $\max_i \chi_i(E)$ (radius of circumscribed ball),
- $\sum_i \chi_i^p(E)$, $p > 0$, where $\chi_i(E)$ are the half-axes of $E$.

Indeed, the half-axes of $E$ are the eigenvalues of the “parameter” $A = B^{-1}$ of the image representation of $E$, that is,

$$\chi_i(E) = \lambda_i^{-1}(B)$$

Therefore

\begin{align*}
(a) & \quad \text{Vol}(E) = \lambda_1^{-1}(B) \cdots \lambda_n^{-1}(B) \\
(b) & \quad \max_i \chi_i(E) = \max_i \lambda_i^{-1}(B) \\
(c) & \quad \sum_i \chi_i^p(E) = \sum_i \lambda_i^{-p}(B)
\end{align*}

are convex symmetric functions of the eigenvalues of $B \succ 0$ and thus are convex functions of $B \succ 0$.

Note: From Calculus of SDr Functions/ Sets it follows that the sizes (a), (b) are SDr functions of $B$; the same is true for size (c) provided that $p > 0$ is rational.

3.131
Summary of observations: With the inequality representation of ellipsoids, typical problems of outer ellipsoidal approximation become problems of minimizing convex $SDr$ functions over convex feasible sets.

⇒ If the feasible set of a problem of outer ellipsoidal approximation is “computationally tractable” (in particular, is $SDr$), the problem itself is computationally tractable (in particular, is an $SDP$).

Note: “If the feasible set ... is computationally tractable” is a big "IF" indeed!
Observation I.1: Let $X \subset \mathbb{R}^n$ be a nonempty convex set. Then the set $\mathcal{X}$ of parameters $A, a$ of image representations of ellipsoids contained in $X$ is convex.

To prove that $\mathcal{X}$ is convex, let $\lambda \in [0, 1]$, $(A', a'), (A'', a'') \in \mathcal{X}$, so that $A \succeq 0$, $A' \succeq 0$ and

$$\forall (u : u^T u \leq 1) : \begin{cases} a' + A'u & \in X \\ a'' + A''u & \in X \end{cases}$$

we should prove that $\lambda(A', a') + (1 - \lambda)(A'', a'') \in \mathcal{X}$, that is,

$$\forall (u : u^T u \leq 1) : [\lambda a' + (1 - \lambda) a''] + [\lambda A' + (1 - \lambda) A'']u$$

$$\equiv \lambda[a' + A'u] + (1 - \lambda)[a'' + A''u] \in X.$$

But this is an immediate corollary of (*) and the convexity of $X$. 

3.133
Observation I.2: “Typical sizes” of an ellipsoid $E$ are concave (and thus easy-to-maximize) functions of the parameters $A,a$ of the image representation of $E$. This is the case, e.g., for the sizes

- $\text{Vol}^p(E) = \prod_i \chi_i^p(E), 0 < p \leq \frac{1}{n}$,
- $\min_i \chi_i(E)$ (“minimal width” of $E$)
- $\sum_i \chi_i^p(E), 0 < p \leq 1$, where $\chi_i(E)$ are the half-axes of $E$.

Indeed, the half-axes of $E$ are the eigenvalues of the “parameter” $A$ of the image representation of $E$ $\Rightarrow$

\[(a) \quad \text{Vol}^p(E) = (\lambda_1(A)\ldots\lambda_n(A))^p \]
\[(b) \quad \min_i \chi_i(E) = \min_i \lambda_i(A) \]
\[(c) \quad \sum_i \chi_i^p(E) = \sum_i \lambda_i^p(A) \]

are concave symmetric functions of the eigenvalues of $A \succeq 0$ and thus are concave functions of $A \succeq 0$.

Note: From Calculus of SDr Functions/Sets it follows that for a rational $p$, the sizes $(a) - (c)$ are SDr functions of $A$. 

3.134
Summary of observations: With the inequality representation of ellipsoids, typical problems of inner ellipsoidal approximation become problems of minimizing convex $SDr$ functions over convex feasible sets.

⇒ If the feasible set of a typical problem of inner ellipsoidal approximation is “computationally tractable” (in particular, is $SDr$), the problem itself is computationally tractable (in particular, is an SDP).

Note: “If the feasible set ... is computationally tractable” is a big ”IF” indeed!
We have seen that the typical problems of inner and outer ellipsoidal approximation are problems of minimizing explicit convex (usually even SDr) functions over convex feasible sets. As we shall see in the meantime, problems of this type are “computationally tractable” if the feasible sets are so.

A sufficient condition for “computational tractability” of a convex set $\mathcal{X}$ is the possibility to solve efficiently the Analysis problem “Given $x$, check whether $x \in \mathcal{X}$.”

In our context, the Analysis problem is

- **in Outer ellipsoidal approximation of a set $X$ — problem**
  
  (AO) Given an ellipsoid $E$, check whether $E \supset X$.

- **in Inner ellipsoidal approximation of a set $X$ — problem**
  
  (AI) Given an ellipsoid $E$, check whether $E \subset X$.

Whether these analysis problems are/are not tractable, it depends on the structure of $X$. 

3.136
Given an ellipsoid $E$, check whether $E \supset X$.

- (AO) is easy when $X$ is a polytope given as a convex hull of a finite set $\{x^1, \ldots, x^M\}$. Indeed, $\text{Conv}\{x^1, \ldots, x^M\} \subset E$ iff $x^i \in E$ for all $i$, and it is easy to check whether or not a point belongs to $E$.

- (AO) can be NP-hard when $X$ is a polytope given by a list of linear inequalities. Indeed, to check whether the unit cube $\{x : \|x\|_\infty \leq 1\}$ belongs to the centered at the origin ellipsoid $\{x : x^TQx \leq r^2\}$, where $Q \succ 0$, is the same as to verify whether

$$\max_{x} \{x^TQx : \|x\|_\infty \leq 1\} \leq r^2,$$

and the latter problem is, essentially, the NP-hard problem of maximizing positive definite homogeneous quadratic form over the unit cube.
(AI) Given an ellipsoid \( E \), check whether \( E \subset X \).

- (AI) is easy when \( X \) is a polytope \( P \) given by a list of linear inequalities \( a_i^T x \leq b_i, \ 1 \leq i \leq M \). Indeed, to check whether an ellipsoid \( E \) is contained in \( P \) is the same as to check whether \( \max_{x \in E} a_i^T x \leq b_i \) for all \( i \), and it is easy to maximize a linear form over an ellipsoid.

- (AI) can be NP-hard when \( X \) is a polytope given as \( \text{Conv}\{x^1, ..., x^M\} \).
**Basic fact** [Boyd et al.] Let $E = \mathcal{E}(A, a)$ and $C = \mathcal{C}(B, b)$ be ellipsoid and elliptic cylinder given, respectively, by image and inequality representations. Then

$$E \equiv \mathcal{E}(A, a) \subset C \equiv \mathcal{C}(B, b) \quad (*)$$

$$ \iff \exists \lambda : \begin{bmatrix}
1 - \lambda & a^T B - b^T \\
\lambda I & AB \\
Ba - b & BA \\
\end{bmatrix} \succeq 0 \quad (**)$$

**Note:** For $E$ fixed, (**) is an LMI in variable $\lambda$ and in the parameters $B, b$ of $C$. For $C$ fixed, (**) is an LMI in variable $\lambda$ and in the parameters $A, a$ of $E$.

Thus, both the facts that
— an ellipsoid is contained in a fixed elliptic cylinder
— an elliptic cylinder contains a fixed ellipsoid
are semidefinite representable!
\[
E \equiv \mathcal{E}(A, a) \subset C \equiv \mathcal{C}(B, b) \iff \exists \lambda : \begin{bmatrix}
1 - \lambda & a^T B - b^T \\
\lambda I & AB \\
Ba - b & BA \\
I & I
\end{bmatrix} \succeq 0
\]

Proof of equivalence:

\[
\{Au + a : u^T u \leq 1\} \subset \{x : \|Bx - b\|_2^2 \leq 1\} \iff \forall (u : u^T u \leq 1) : \|BAu + Ba - b\|_2^2 \leq 1
\]

\[
\iff \forall (v, t : v^T v \leq t^2, t \neq 0) : \|t^{-1}BAv + c\|_2^2 \leq 1 \iff \forall (v, t : v^T v \leq t^2, t \neq 0) : \|BAv + tc\|_2^2 \leq t^2
\]

\[
\iff \forall (v, t : t^2 - v^T v \geq 0) : t^2 - \|BAv + tc\|_2^2 \geq 0 \iff \exists \lambda \geq 0 : t^2 - \|BAv + tc\|_2^2 - \lambda [t^2 - v^T v] \geq 0 \quad \forall (v, t)
\]

\[
\iff \exists \lambda \geq 0 : \begin{bmatrix}
1 - \lambda & c^T AB \\
\lambda & c^T
\end{bmatrix} T \succeq 0
\]

\[
\iff \exists \lambda \geq 0 : \begin{bmatrix}
1 - \lambda & a^T B - b^T \\
\lambda I & AB \\
Ba - b & BA \\
I & I
\end{bmatrix} \succeq 0
\]
Conclusions:

Let $X$ be a union of finitely many ellipsoids. The problem of finding the smallest ellipsoid $E$ containing $X$ can be posed as an explicit semidefinite program, provided that the size to be minimized is
— either the volume $\text{Vol}(E)$,
— or the maximal half-axis $\max_i \chi_i(E)$ of $E$,
— or $\sum_i \chi_i^p(E)$ with rational $p > 0$.

“Good” design variables in the problem are the parameters $B, b$ of the inequality representation of $E$.

In particular, the problem of finding the smallest ellipsoid containing a polytope given as a convex hull of a finite set of points can be posed as an explicit semidefinite program.
Let $X$ be an intersection of finitely many elliptical cylinders. The problem of finding the largest ellipsoid $E$ contained in $X$ can be posed as an explicit semidefinite program, provided that the size to be maximized is

— either $p$-th power of the volume Vol($E$), with rational $p \in [0, 1/n]$, 
— or the minimal half-axis $\min_i \chi_i(E)$ of $E$, 
— or $\sum_i \chi_i^p(E)$ with rational $p$, $0 < p \leq 1$.

“Good” design variables in the problem are the parameters $A, a$ of the image representation of $E$.

In particular, the problem of finding the largest ellipsoid contained in a polytope *given by a finite list of linear inequalities* can be posed as explicit semidefinite program.
Important Difficult Open problem: Outer Ellipsoidal approximation of intersection

\[ \hat{E} = \bigcap_{i=1}^{m} E_i \]

of ellipsoids (or elliptic cylinders).

Source of difficulty: Given two ellipsoids, we understand how to check efficiently that one of them is contained in the other one, but we do not know how to check efficiently that a given ellipsoid contains the intersection of a collection of ellipsoids.

- The latter problem reduces to describing strongly convex quadratic inequalities

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

which are consequences of systems

\[ x^r A_i x + 2 b_i^T x + c_i \leq 0, \quad 1 \leq i \leq m \quad [A_i \succ 0 \ \forall i] \]

of strongly convex quadratic inequalities.

This problem is NP-hard, and the SDP Relaxation, based on replacing the set of all consequences with the set of all linear consequences, fails to work properly!

3.143
\[ \hat{E} = \bigcap_{i=1}^{m} E_i, \quad E_i: \text{ellipsoids} \]

There are several interesting “ad hoc” approximations of the smallest in volume Outer Ellipsoidal approximation of \( \hat{E} \). In all schemes, one builds efficiently two similar to each other concentric ellipsoids \( E, \overline{E} \) which “bracket” \( \hat{E} \):

\[ E \subset \hat{E} \subset \overline{E}, \]

and guarantees certain bounds on the similarity ratio \( \theta \) of the “brackets”.
One scheme allows to ensure $\theta \leq n$ and is based on the following nice fact:

**Fritz John Theorem:** For every convex compact set $X \subset \mathbb{R}^n$ with a nonempty interior, there exists a unique smallest volume ellipsoid $E_{\text{out}}$ containing $X$, same as there exists a unique largest volume ellipsoid $E_{\text{in}}$ contained in $X$.

When shrinking $E_{\text{out}}$ to its center with the coefficient $n$, one gets an ellipsoid which is contained in $X$, and when enlarging $E_{\text{in}}$ by factor $n$ (keeping the center fixed), one gets an ellipsoid which contains $X$.

When $X$ has a symmetry center, the shrinkage/enlargement by factor $n$ can be replaced with shrinkage/enlargement by factor $\sqrt{n}$.

We would like to build $E_{\text{out}}$, but we do not know how to do it efficiently. However, we do know how to build efficiently $E_{\text{in}}$. Building $E_{\text{in}}$ and enlarging it by factor $n$, we, by Fritz John Theorem, get an ellipsoid containing $\hat{E}$, the ratio of the linear sizes of the resulting “brackets” being $n$. 

3.145
• Another scheme allows to ensure \( \theta \leq m + 2\sqrt{m} \) (non-optimality in volume by factor \( \leq (m + 2\sqrt{m})^n \)). Without essential loss of generality, we can assume that

\[
E_i = \{ x : \|B_i x - b_i\|_2^2 \leq 1 \}
\]

\( \hat{E} \) is bounded, and \( \text{int} \hat{E} \neq \emptyset \). We form the analytical barrier for \( \hat{E} \) – the explicit convex function

\[
F(x) = -\sum_i \ln(1 - \|B_i x - b_i\|_2^2)
\]

with the domain \( \text{int} \hat{E} \), solve the convex optimization problem

\[
x_* = \arg\min_{x \in \text{int} \hat{E}} F(x)
\]

(this can be done efficiently) and set

\[
E = \{ x : (x - x_*)^T \nabla^2 F(x_*)(x - x_*) \leq 1 \},
\]

\[
\overline{E} = \{ x : (x - x_*)^T \nabla^2 F(x_*)(x - x_*) \leq (m + 2\sqrt{m})^2 \}\]
In Outer Ellipsoidal approximation of intersection of ellipsoids, SDP Relaxation “recovers its power” when all the ellipsoids in the intersection have a common center (w.l.o.g., 0):

$$\hat{E} = \{x : x^T S_i x \leq 1, i = 1, \ldots, m\} \quad [S_i \succeq 0]$$

Assuming that $\hat{E}$ is bounded ($\Leftrightarrow \sum_i S_i \succ 0$), observe that the optimal circumscribed ellipsoid is centered at the origin. Indeed, if

$$C_+ \equiv \{x : \|Bx - b\|_2^2 \leq 1\} \supseteq \hat{E},$$

then, due to symmetry of $\hat{E}$, we have

$$C_- \equiv \{x : \|Bx + b\|_2^2 \leq 1\} \supseteq \hat{E}$$

as well, whence, due to the convexity of the set $\{(P, p) : C(P, p) \supseteq \hat{E}\}$, we have

$$C \equiv \{x : \|Bx\|_2^2 \leq 1\} \supseteq \hat{E},$$

and $C$ has the same sizes as $C_+$ and $C_-$. Thus, the Outer Ellipsoidal approximation problem becomes

$$\min_B \{\text{Size}(B) : B \succ 0 \& \ x^T B^2 x \leq 1 \forall (x : x^T S_i x \leq 1, i = 1, \ldots, m)\} \quad (*)$$

where $\text{Size}(B)$ is the size of ellipsoid we intend to minimize.

3.147
\[
\min_B \left\{ \text{Size}(B) : B \succeq 0 \& \ x^T B^2 x \leq 1 \forall (x : x^T S_i x \leq 1, i = 1, \ldots, m) \right\} \quad (\ast)
\]

\[\blacklozenge\] The constraint

\[x^T B^2 x \leq 1 \quad \forall (x : x^T S_i x \leq 1, i = 1, \ldots, m)\]

is equivalent to

\[
\max_x \left\{ x^T B^2 x : x^T S_i x \leq 1, i = 1, \ldots, m \right\} \leq 1
\]

and thus admits a “conservative approximation”, given by SDP Relaxation:

\[
\text{Opt}(B) \equiv \max_x \left\{ x^T B^2 x : x^T S_i x \leq 1, i = 1, \ldots, m \right\}
\leq \max_X \left\{ \text{Tr}(B^2 X) : X \succeq 0, \text{Tr}(S_i X) \leq 1, i = 1, \ldots, m \right\}
\]

\[= \min_\mu \left\{ \sum_i \mu_i : \mu \geq 0, B^2 \preceq \sum_i \mu_i S_i \right\} \quad \text{[Semidefinite Duality]} \]

\[= \min_\mu \left\{ \sum_i \mu_i : \mu \geq 0, \left[ \frac{\sum_i \mu_i S_i}{B} \right] \succeq 0 \right\} \quad \text{[Schur Complement Lemma]} \]

3.148
\[
\min_B \left\{ \text{Size}(B) : B \succeq 0 \ \& \ x^T B^2 x \leq 1 \forall (x : x^T S_i x \leq 1, i = 1,.,m) \right\} \quad (\ast)
\]

⇒ The optimization program

\[
\min_{B,\mu} \left\{ \text{Size}(B) : B \succeq 0, \mu \geq 0, \sum_i \mu_i \leq 1, \left[ \frac{\sum_i \mu_i S_i}{B} \right] \succeq 0 \right\} \quad (\ast\ast)
\]

is a conservative approximation of \((\ast)\) – both problems have the same objective, and the projection of the feasible set of \((\ast\ast)\) on the \(B\)-plane is contained in the feasible set of \((\ast)\).

♠ A typical size \(\text{Size}(B)\) is a SDr function of \(B\); whenever this is the case, \((\ast\ast)\) can be posed as an explicit semidefinite program, and its optimal solution induces a feasible and suboptimal solution to \((\ast)\).
\[ \text{Opt} = \min_B \left\{ \text{Size}(B) : \begin{array}{l} \begin{aligned} & B \succeq 0 \\ & x^T B^2 x \leq 1 \forall (x : x^T S_i x \leq 1 \forall i) \end{aligned} \end{array} \right\} \]  

\( \Rightarrow \text{SDP} = \min_{B, \mu} \left\{ \text{Size}(B) : \begin{array}{l} \begin{aligned} & B \succeq 0, \mu \geq 0, \sum_i \mu_i \leq 1 \\ & \left[ \begin{array}{c} \sum_i \mu_i S_i \\ B \\ I \end{array} \right] \succeq 0 \end{aligned} \end{array} \right\} \)  

\( \bigstar \) If \((B, \mu)\) is feasible for \((**\))\, then \(B\) is feasible for \((*)\) \(\Rightarrow\) \(\text{Opt} \leq \text{SDP}\)  

\( \bigspadesuit \) From the Approximate \(S\)-Lemma it follows that the “relaxation inequality”

\[ \max _{x} \left\{ x^T B^2 x : x^T S_i x \leq 1 \forall i \right\} \leq \min _{\mu} \left\{ \sum_i \mu_i : \left[ \begin{array}{c} \sum_i \mu_i S_i \\ B \\ I \end{array} \right] \succeq 0 \right\} \]

is tight:

\[ \text{SDP}(B) \leq \Omega^2 \text{Opt}(B), \quad \Omega = \sqrt{2 \ln \left( 6 \sum_i \text{Rank}(S_i) \right)} \]

It follows that if \(B\) is a feasible solution to the problem of interest \((*)\), then \(\Omega^{-1} B\) can be extended to a feasible solution to \((**)\). All sizes \(\text{Size}(B)\) we have considered are homogeneous:

\[ \text{Size}(\theta B) = \theta^{-\chi} \text{Size}(B) \Rightarrow \text{SDP} \leq \Omega^\chi \text{Opt} \]

\( \Rightarrow (**) \) yields optimal in size, up to factor \(\Omega^\chi\), “ellipsoidal cover” of \(\hat{E} = \{x : x^T S_i x \leq 1, i = 1, \ldots, m\}\).
Inner and Outer Ellipsoidal Approximations of Sums of Ellipsoids

Problems of interest: Given $m$ ellipsoids $W_1, \ldots, W_m$ in $\mathbb{R}^n$, find the best in the volume inner (problem (I)) and outer (problem (O)) ellipsoidal approximations of the arithmetic sum

$$W = \{x = w_1 + w_1 + \ldots + w_m : w_i \in W_i, \ i = 1, \ldots, m\}$$

of the ellipsoids $W_1, \ldots, W_m$.

♠ Note: When shifting one of the sets $A, B, \ldots, Z$ by a vector $a$, the arithmetic sum $A + B + \ldots + Z$ of the sets is shifted by the same vector $a$.

⇒ We may assume w.l.o.g. that all the ellipsoids $W_i$ are centered at the origin:

$$W_i = \{x \in \mathbb{R}^n : x^T Z_i x \leq 1\} \quad [Z_i \succ 0].$$

In this case the solutions to (I) and (O) also can be sought among the ellipsoids centered at the origin.
Outer Ellipsoidal Approximation of Sum of Ellipsoids

**Observation:** Ellipsoid

\[ E = \{ x : x^T Z x \leq 1 \} \quad [Z > 0] \]

contains the arithmetic sum of ellipsoids

\[ W_i = \{ x : x^T Z_i x \leq 1 \}, \quad i = 1, \ldots, m \]

iff

\[
\max_{u = (u^1, \ldots, u^m)} \left\{ \frac{(u^1 + \ldots + u^m)^T Z (u^1 + \ldots + u^m)}{u^i \mathcal{M}[Z] u^i} : \frac{(u^i)^T Z_i u^i}{u^i \mathcal{M}_i u^i} \leq 1, \quad i = 1, \ldots, m \right\} \leq 1
\]

\[
\mathcal{M}[Z] = \begin{bmatrix} Z & Z & \cdots \\ Z & \ddots & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}, \quad \mathcal{M}_1 = \begin{bmatrix} Z_1 \\ \vdots \\ Z \end{bmatrix}, \ldots, \mathcal{M}_m = \begin{bmatrix} \vdots \\ Z_m \end{bmatrix}
\]

♠ Applying Semidefinite Relaxation, we arrive at the following conservative approximation of (O):

\[
\min_{Z, \mu} \left\{ \det^{-1/2}(Z) : \begin{array}{c} Z > 0, \quad \mu \geq 0, \sum_i \mu_i \leq 1 \\
\mathcal{M}[Z] \preceq \sum_i \mu_i \mathcal{M}_i \end{array} \right\} \quad (*)
\]

♠ Matrices \( \mathcal{M}_i \) are positive semidefinite and commute with each other. Applying Nesterov's \( \frac{\pi}{2} \) Theorem, it is easily seen that the optimal solution to (\( * \)) yields an optimal, up to factor \( (\frac{\pi}{2})^{n/2} \), solution to (O).
Observation: An ellipsoid
\[ E = \{ x = Au : u^T u \leq 1 \} \]
is contained in the sum of ellipsoids
\[ W_i = \{ x = A_i u : u^T u \leq 1 \}, \ i = 1, ..., m \]
iff for every vector \( u \) one has
\[ \| A^T \xi \|_2 \leq \sum_i \| A_i^T \xi \|_2. \] (**)

Proof. Let \( P, Q \) be closed nonempty convex sets. From Separation Theorem it immediately follows that
\[ P \subset Q \iff \max_{x \in Q} \xi^T x \geq \max_{x \in P} \xi^T x \ \forall x. \]

With \( P = E \), we have
\[ \max_{x \in P} \xi^T x = \max_{u} \{ \xi^T Au : u^T u \leq 1 \} = \| A^T \xi \|_2. \]

With \( Q = W_1 + ... + W_m \), we have
\[ \max_{x \in Q} \xi^T x = \max_{u_{1, \ldots, u}} \{ \xi^T [A_1 u^1 + \ldots + A_m u^m] : \| u^i \|_2 \leq 1 \ \forall i \} = \sum_i \| A_i^T \xi \|_2. \]
Thus, \( E \subset W_1 + ... + W_m \) if and only if (***) takes place.
\[ \|A^T \xi\|_2 \leq \sum_i \|A_i^T \xi\|_2. \] (\star)

**Observation I:** Given matrices \( A_i \), the simplest way to generate matrix \( A \) satisfying (\star) is to set
\[ A = \sum_i A_i X_i, \quad \|X_i\| \leq 1 \] (\star\star)

**Observation II:** Let \( A = S + C \) with symmetric positive definite \( S \) and skew-symmetric \( C \). Then
\[ \text{Det}(A) = |\text{Det}(A)| \geq \text{Det}(S) \]

Indeed, by “scaling”
\[ A = S + C \leftrightarrow \hat{A} = S^{-1/2} AS^{-1/2} = I + S^{-1/2}CS^{-1/2} \]
we reduce the general case to the one where \( S = I \). Here the statement is evident: since the eigenvalues of skew-symmetric real matrix \( C \) are pairs of conjugate purely imaginary complex numbers \( \pm i\nu_\ell \), we have
\[ \text{Det}(A) = \text{Det}(I + C) = \prod_\ell [(1 - i\nu_\ell)(1 + i\nu_\ell)] = \prod_\ell [1 + \nu_\ell^2] \geq 1 = \text{Det}(I). \]
We arrive at the following conservative approximation of (I):

\[
\max_{\{X_i\}} \left\{ \text{Det}^{1/n} \left( \frac{1}{2} \sum_i [X_i^T A_i + A_i X_i] \right) : \begin{cases} \sum_i [X_i^T A_i + A_i X_i] \succeq 0 \\ \begin{bmatrix} I & X_i \\ X_i & I \end{bmatrix} \succeq 0 \forall i \\ \equiv \|X_i\| \leq 1 \end{cases} \right\} \quad (P)
\]

where \( A_i \succeq 0 \) are the matrices from the image representations of the ellipsoids \( W_i \).

Every feasible solution \( \{X_i\} \) of \((P)\) produces ellipsoid

\[ E = \{x = Au : u^T u \leq 1\}, \quad A = \sum_i A_i X_i \]

which is contained in \( W_1 + ... + W_m \), and the volume of this ellipsoid is at least

\[ \text{Det} \left( \frac{1}{2} \sum_i [X_i^T A_i + A_i X_i] \right). \]
Observation: Problems (O) and (I) (same as all problems of “optimal in volume” ellipsoidal approximation) admit certain symmetry. Specifically, let

\[ y = Qx \]

be a nondegenerate linear transformation of \( \mathbb{R}^n \). Such a transformation multiplies the volumes of all sets by the same factor \(|\text{Det}(Q)|\); consequently, problems (I)/(O) involving ellipsoids

\[ W_i = \{x : x^T Z_i x \leq 1\} \quad [Z_i > 0] \]

can be reduced to similar problems involving the images

\[ \hat{W}_i = \{y : (Q^{-1} y)^T Z_i (Q^{-1} y) \leq 1\} = \{y : y^T \underbrace{[Q^{-T} Z_i Q^{-1}]}_{\hat{Z}_i} y \leq 1\} \]

of ellipsoids \( W_i \) under this transformation.
Let us call ellipsoids $W_i$ co-axial, if, with a proper choice of $Q$, the matrices $\hat{Z}_i$ commute with each other.

Co-axiality is equivalent to the existence of a basis (non necessarily orthogonal) where all quadratic forms $x^T Z_i x$ become diagonal:

$$x^T Z_i x = \sum_j \nu_j^i \xi_j^2(x)$$

$$[\xi_j(x) : \text{coordinates of } x \text{ in the basis}]$$

Linear Algebra says that every two (full-dimensional) ellipsoids $W_1, W_2$ are co-axial. Indeed, if $W_i = \{x : x^T Z_i x \leq 0\}$ and $Z_i \succ 0$, $i = 1, 2$, then, setting $Q = Z_1^{1/2}$, we arrive at commuting matrices

$$\hat{Z}_1 = Z_1^{-1/2} Z_1 Z_1^{-1/2} = I, \quad \hat{Z}_2 = Z_1^{-1/2} Z_2 Z_1^{-1/2}.$$
We have seen that in the co-axial case problems (I) and (O) can be reduced to similar problems for the sum of ellipsoids given by diagonal matrices:

\[ W_i = \{ x : \sum_j \nu^j_i x_j^2 \leq 1 \} \quad [\nu^j_i > 0] \]

It turns out that in the latter case the tractable approximations of (O) and (I) we have presented yield exactly optimal solutions to the respective problems. This is a corollary of the following simple and powerful Symmetry Principle.
**Symmetry Principle:** Consider a convex and solvable optimization problem

\[
\min_{x \in X} f(x) \quad (P)
\]

and assume that it admits a finite group \( G \) of symmetries, that is,

- \( G \) is a finite subset of the group \( \mathcal{L}_n \) of nonsingular \( n \times n \) matrices,
- \( G \) is a sub-group of \( \mathcal{L}_n \): \( U \in G \Rightarrow U^{-1} \in G \), \( U, V \in G \Rightarrow UV \in G \) and
- every \( U \in G \) is a symmetry of \((P)\):

\[
U(X) := \{Ux : x \in X\} = X, \quad f(Ux) = f(x) \quad \forall x \in X.
\]

Then \((P)\) admits a “\( G \)-symmetric” optimal solution \( x_\ast \):

\[
Ux_\ast = x_\ast \quad \forall U \in G.
\]

**Proof.** Let \( \bar{x} \) be an optimal solution to \((P)\). Since \((P)\) is \( G \)-symmetric, every point of the form

\[
U\bar{x}, \quad U \in G
\]

is an optimal solution to \((P)\) along with \( \bar{x} \). Since \((P)\) is convex, it follows that the point

\[
x_\ast = \frac{1}{\text{Card}(G)} \sum_{U \in G} U \bar{x}
\]

also is an optimal solution to \((P)\); this solution is clearly \( G \)-symmetric.
Remark: Assuming $X$ closed, the statement remains valid when $G$ is a compact, rather than finite, group of symmetries of $(P)$. The proof remains essentially the same, with averaging ($\ast$) replaced by integration over the invariant probabilistic measure on $G$. 

3.160
From Symmetry Principle to Co-Axial (O)/(I). Let ellipsoids $W_i$ be given by diagonal matrices:

$$W_i = \{ x : \sum_{j} \nu^i_j x_j^2 \leq 1 \} \quad [\nu^i_j > 0]$$

Consider problem (O):

$$\min_{B, b} \left\{ \text{Det}^{-1}(B) \equiv \prod_j \lambda_j^{-1}(B) : C(B, b) \ni \underbrace{W_1 + \ldots + W_m}_{W}, B \succ 0 \right\} \quad (O)$$

The problem is convex and solvable (the latter – by Fritz John Theorem). Let $J$ be a transformation of $\mathbb{R}^n$ of the form

$$x \mapsto (\epsilon_1 x_1, \epsilon_2 x_2, \ldots, \epsilon_n x_n), \quad \epsilon_j = \pm 1.$$ 

Since $W_i$ are given by diagonal matrices, this transformation keeps $W$ invariant and therefore maps an ellipsoid $C(B, b)$ containing $W$ into another ellipsoid also containing $W$; this “other ellipsoid” is $C(JBJ, Jb)$. Thus, the feasible set of convex and solvable problem (O) is invariant under the transformations

$$J : (B, b) \mapsto (JBJ, Jb) \equiv (J^TBJ, Jb)$$

generated by $2^n$ “reflections” $J$. The transformations $J$ clearly form a finite sub-group of the group of orthogonal rotations of the Euclidean space $S^n \times \mathbb{R}^n$ where the feasible set of (O) lives, and that these transformations preserve the objective in (O). Applying Symmetry Principle, we conclude that (O) admits an optimal solution $(B_*, b_*)$ which remains invariant under all transformations of the form

$$(B, b) \mapsto (J^TBJ, Jb), \quad J = \text{Diag}\{\epsilon_1, \ldots, \epsilon_n\}, \quad \epsilon_i = \pm 1,$$

which clearly is possible iff $b_* = 0$ and $B_*$ is diagonal.

3.161
\[ W_i = \{ x : \sum_j \nu_j^i x_j^2 \leq 1 \} \quad [\nu_j^i > 0] \]

\[
\min_{B,b} \left\{ \frac{1}{\det(B)} \equiv \prod_j \lambda_j^{-1}(B) : C(B, b) \supset W_1 + \ldots + W_m, B > 0 \right\} \quad (O)
\]

We have seen that when solving \((O)\), we lose nothing by assuming that \(b = 0\) and \(B\) is diagonal, so that \((O)\) is equivalent to the problem

\[
\min_{\beta} \left\{ \prod_j \beta_j^{-1} : \beta > 0, \sum_j \beta_j x_j^2 \leq 1 \forall (x = x^1 + \ldots + x^m : \sum_j \nu_j^i (x_j^i)^2) \right\}
\]

\[
\Leftrightarrow \min_{\beta > 0} \left\{ \prod_j \beta_j^{-1} : \sum_j \beta_j \left( \sum_j y_j^i \right)^2 \leq 1 \forall \left( \{y_j^i \geq 0\} : \sum_j \nu_j^i y_j^i \leq 1, 1 \leq i \leq m \right) \right\} \quad (O')
\]

We claim that

\((!)\) A vector \(\beta > 0\) is feasible for \((O')\) if and only if there exists \(\mu \geq 0\) such that \(M[\text{Diag}\{\beta\}] \preceq \sum_i \mu_i M_i\).

\((!)\) says that the matrices \(\text{Diag}\{\beta\}\) associated with feasible solutions to \((O')\) are feasible solutions to the tractable approximation of \((O)\) we have built.

\(\Rightarrow\) Optimal solution to our approximation of \((O)\) is optimal solution of \((O)\) as well.

3.162
\[
\min_{\beta > 0} \left\{ \prod_j \beta_j^{-1} : \sum_j \beta_j \left( \sum_j \sqrt{y_j^i} \right)^2 \leq 1 \quad \forall \left\{ y_j^i \geq 0 : \sum_j \nu_j^i y_j^i \leq 1, 1 \leq i \leq m \right\} \right\} \quad (\mathcal{O}')
\]

**Claim:** (!) A vector \( \beta > 0 \) is feasible for \((\mathcal{O}')\) if and only if there exists \( \mu \geq 0 \) such that 
\[
\sum_i \mu_i \leq 1 \quad \text{and} \quad \mathcal{M}[\text{Diag}\{\beta\}] \preceq \sum_i \mu_i \mathcal{M}_i.
\]

**Proof of (!):** The only nontrivial part of (!) is the claim that (!!) *if \( \beta > 0 \) is feasible for \((\mathcal{O}')\), then there exists \( \mu \geq 0 \) such that...*

By Semidefinite Duality, the property "exists \( \mu \geq 0 \) such that..." is exactly equivalent to the validity of the implication 
\[
Y \in S_{+}^{mn}, \quad \text{Tr}(\mathcal{M}_i Y) \leq 1, 1 \leq i \leq m \Rightarrow \text{Tr}(\mathcal{M}[\text{Diag}\{\beta\}] Y) \leq 1 \quad (1)
\]

so that to prove (!!!) is the same as to prove that 

**(!!!) If \( \beta \) is feasible for \((\mathcal{O}')\), then (1) takes place.**

To prove (!!!), let \( \beta \) be feasible for \((\mathcal{O}')\), and let \( Y \) satisfy the premise in (1). Let us split \( Y \) into \( m^2 \) blocks \( Y^{ik} \) of the size \( n \times n \) each.
Situation: \( \beta \) is feasible for

\[
\min_{\beta > 0} \left\{ \prod_j \beta_j^{-1} : \sum_j \beta_j \left( \sum_j \sqrt{y_{jj}^i} \right)^2 \leq 1 \quad \forall \left( \{y_{jj}^i \geq 0\} : \sum_j \nu_j^i y_{jj}^i \leq 1, 1 \leq i \leq m \right) \right\} \quad (O')
\]

\( Y = [Y^{k\ell} \in \mathbb{R}^{n \times n}]_{k,\ell \leq m} \) satisfies the premise in

\[
y \in S_{+}^{mn}, \text{Tr}(\mathcal{M}_i Y) \leq 1, 1 \leq i \leq m \Rightarrow \text{Tr}(\mathcal{M}[\text{Diag}\{\beta\}]Y) \leq 1 \quad (1)
\]

Goal: to justify the validity of the conclusion in (1).

Taking into account that \( Y \succeq 0 \), we have \( |Y_{ik}^{ij}| \leq \sqrt{Y_{ii}^{ij}Y_{kk}^{ij}} \), whence

\[
\text{Tr}(\mathcal{M}[\text{Diag}\{\beta\}]Y) = \sum_{i,k=1}^{m} \sum_{j=1}^{n} \beta_j Y_{jj}^{ik} \leq \sum_{i,k=1}^{m} \sum_{j=1}^{n} \beta_j \sqrt{Y_{ii}^{ij}Y_{kk}^{ij}} = \sum_{j=1}^{n} \beta_j \left( \sum_{i=1}^{m} \sqrt{Y_{ii}^{ij}} \right)^2
\]

Since \( Y \) satisfies the premise in (1), we have

\[
\text{Tr}(\mathcal{M}_i Y) \equiv \sum_j \nu_j^i Y_{jj}^{ii} \leq 1,
\]

whence, since \( \beta \) is feasible for (O'),

\[
\text{Tr}(\mathcal{M}[\text{Diag}\{\beta\}]Y) = \sum_{j=1}^{n} \beta_j \left( \sum_{i=1}^{m} \sqrt{Y_{ii}^{ij}} \right)^2 \leq 1,
\]

as required in the conclusion of (1). \( \square \)
Let ellipsoids $W_i$ be given by diagonal matrices:

$$W_i = \{ x : \sum_j \nu_{ij}^i x_j^2 \leq 1 \} \quad [\nu_{ij}^i > 0]$$

$$\Rightarrow W_i = \{ x = \text{Diag}\{\theta^i\} u : u^T u \leq 1 \} \quad [\theta_j^i = (\nu_{ij}^i)^{-1/2}]$$

**Problem (I).** In the case of diagonal matrices $A_i \succeq 0$, our approximation scheme recovers exactly optimal ellipsoid contained in $W_1 + ... + W_m$. Moreover, this ellipsoid is just

$$W = \{ x = \underbrace{[A_1 + ... + A_m]}_{A} u : u^T u = 1 \}. \quad (!)$$

Indeed, ellipsoid (!) is given by our approximation scheme:

$$A = \frac{1}{2} \sum_i \ [X_i^T A_i + A_i X_i] \succeq 0 \quad [X_i = I, \|X_i\| \leq 1]$$

thus, the ellipsoid is contained in $W_1 + ... + W_m$.

On the other hand, it is clear that the set $W_1 + ... + W_m$ is contained in the box

$$\{ x : |x_j| \leq \theta_1^j + \theta_2^j + ... + \theta_m^j, j = 1, ..., n \},$$

so that the largest volume ellipsoid contained in this box (which is exactly $W!$) can be only larger than the largest volume ellipsoid contained in $W_1 + ... + W_m$. 

3.165
Application: On-line approximation of reachable sets.

\[ z(t + 1) = A_t z(t) + B_t u(t) + f_t, \quad z(0) = z_0 \]  

The set \( Z^T \) of all states \( z(T) \) of (1) reachable with norm-bounded control:

\[ \|u(t)\|_2 \leq \rho_t, \quad t = 0, 1, ..., T - 1 \]

is the sum of \( T \) ellipsoids and thus can be approximated from inside and from outside by ellipsoids via our techniques. We can further “trade quality for simplicity” and look at on-line approximations, where, given ellipsoidal approximations of \( Z^t \):

\[ E_t \subset Z^t \subset E^t \]

and observing that

\[ Z^{t+1} = A_t Z^t + \{B_t u + f_t : u^T u \leq \rho_t^2\}, \]

we conclude that

\[ A_tE_t + \{B_t u + f_t : u^T u \leq \rho_t^2\} \subset Z^{t+1} \subset A_tE_t + \{B_t u + f_t : u^T u \leq \rho_t^2\} \]

Thus, setting

\[ E_{t+1} = \text{largest volume ellipsoid} \subset A_tE_t + \{B_t u + f_t : u^T u \leq \rho_t^2\} \]
\[ E^t_{t+1} = \text{smallest volume ellipsoid} \supset A_tE_t + \{B_t u + f_t : u^T u \leq \rho_t^2\} \]

we get (non-optimal!) “greedy” inner and outer ellipsoidal approximations of \( Z^{t+1} \) by solving recursively simple problems of approximating sums of just two ellipsoids (co-axial case!).

3.166
\[
\frac{d}{dt} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} -0.8147 & -0.4163 \\ 0.8167 & -0.1853 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} u_1(t) \\ 0.7071u_2(t) \end{bmatrix}, \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \|u(t)\|_2 \leq 1
\]

\[
\Rightarrow z(k+1) = \exp\{P\Delta t\} z(k) + \int_0^\Delta t \exp\{A\} \begin{bmatrix} 1 \\ 0 \\ \frac{-0.8147}{0.7071} \end{bmatrix} ds \begin{bmatrix} u(k) \\ z(0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, [\Delta t = 0.01] \end{bmatrix}
\]

Outer and inner on-line approximation of \(Z^t, t = 10\ell, \ell = 1, \ldots, 10\), and 4 sample trajectories
A Linear Dynamical System

\[
\begin{align*}
\dot{z} &= A(t)z + B(t)u(t) + f(t), \quad t \geq 0 \\
z(0) &\in E_0 \equiv \{z : z^T G_{\text{ini}} z \leq 1\} \quad [G_{\text{ini}} \succ 0]
\end{align*}
\]

with norm-bounded control:

\[\|u(t)\|_2 \leq 1 \quad \forall t,\]

can be viewed as a limit of discrete time systems with norm-bounded control. The above discrete time greedy on-line policies for building ellipsoidal approximations yield continuous-time counterparts as follows:

We associate with \((\ast)\) ordinary differential equations for matrix-valued functions \(G_t\) and \(W_t\):

\[
\begin{align*}
\frac{d}{dt} G_t &= -A^T(t)G_t - G_t A(t) - \left( \frac{n}{\text{Tr}(G_t B(t)B^T(t))} \right)^{1/2} G_t B(t)B^T(t)G_t - \left( \frac{\text{Tr}(G_t B(t)B^T(t))}{n} \right)^{1/2} G_t, \quad t \geq 0, \\
G_0 &= G_{\text{ini}}; \\
\frac{d}{dt} W_t &= -A^T(t)W_t - W_t A(t) - 2W_t^{1/2} (W_t^{1/2} B(t)B^T(t)W_t^{1/2})^{1/2} W_t^{1/2}, \quad t \geq 0, \\
W_0 &= G_{\text{ini}}.
\end{align*}
\]

Let also \(z_t\) be the “central trajectory”:

\[\frac{d}{dt} z_t = A(t)z_t + f(t), \quad z_0 = 0.\]

Then \(G_t \succ 0, W_t \succ 0\) for all \(t \geq 0\), and for all \(t\) one has

\[\{z : (z - z_t)^T W_t (z - z_t) \leq 1\} \subset Z^t \subset \{z : (z - z_t)^T G_t (z - z_t) \leq 1\}\]

where \(Z^t\) is the set of all possible states of \((\ast)\) at time \(t\).
\[
\frac{d}{dt} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} \cos(t) & -\sin(t) \\ \sin(t) & \cos(t) \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + u(t) \begin{bmatrix} \cos(t) \\ \sin(t) \end{bmatrix} + \begin{bmatrix} 10 \\ 10 \end{bmatrix}
\]

\( x(0) = 0, \quad |u(\cdot)| \leq 1, \quad 0 \leq t \leq 30 \)
\[
\frac{d}{dt} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} 0 & -\sin(t) \\ \frac{\sin(t)}{\sin(t)} & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + u(t) \begin{bmatrix} \cos(t) \\ \sin(t) \end{bmatrix} + \begin{bmatrix} 10 \\ 10 \end{bmatrix}
\]

\text{“Snake”}

\[x(0) = 0, \quad |u(\cdot)| \leq 1, \quad 0 \leq t \leq 30\]
\[
\frac{d}{dt} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + u(t) \begin{bmatrix} 0 \\ 0.05 \end{bmatrix}
\]

\[
\begin{align*}
\frac{d^2}{dt^2} x_1(t) &= -x_1(t) + 0.05u(t) \\
x_2(t) &= \frac{d}{dt} x_1(t)
\end{align*}
\]

\[x_1(0) = 0, \ x_2(0) = 1, \ |u(\cdot)| \leq 1, \ 0 \leq t \leq 30\]
IV. COMPUTATIONAL TRACTABILITY OF CONVEX PROGRAMMING
A Mathematical Programming problem is

\[
\min_x \left\{ p_0(x) : x \in X(p) \subset \mathbb{R}^{n(p)} \right\} \tag{p}
\]

- \(n(p)\) is the *design dimension* of problem \((p)\);
- \(X(p) \subset \mathbb{R}^n\) is the *feasible domain* of the problem;
- \(p_0(x) : \mathbb{R}^n \to \mathbb{R}\) is the *objective* of \((p)\).

E.g., a conic program

\[
\min_x \left\{ c^T x : Ax - b \in \mathcal{K} \right\}, \tag{CP}
\]

is a Mathematical Programming program given by

\[
X(p) = \{ x : Ax - b \in \mathcal{K} \}, \quad p_0(x) = c^T x.
\]
**Definition:** A Mathematical Programming program

\[
\min_x \{ p_0(x) : x \in X(p) \subseteq \mathbb{R}^{n(p)} \}
\]

is called *convex*, if

- The domain \( X(p) \) of the program is a convex set;
- The objective \( p(x) \) is convex on \( \mathbb{R}^{n(p)} \).
- E.g., a conic program

\[
\min_x \{ p_0(x) \equiv c^T x : x \in X(p) \equiv \{ x : Ax - b \in K \} \}
\]

is convex.
Claim: (!) Convex optimization programs are “computationally tractable”: there exist solution methods which “efficiently solve” every convex optimization program satisfying “very mild” computability restrictions.

(!!) In contrast to this, no efficient universal solution methods for non-convex Mathematical Programming programs are known, and there are strong reasons to expect that no methods of this type exist.

- To make (!) a rigorous statement, one should specify the notions of
  - a solution method
  - efficiency
Intuitively, a (numerical) solution method is a computer code; when solving a particular optimization program, a computer loaded with this code inputs the data of the program, executes the code on these data and outputs the result – a real array representing the solution, or the message “no solution exists”.

The efficiency of such a solution method at a particular program can be measured by the running time of the code as applied to the data of the program – by the number of elementary operations performed by the computer when executing the code; the less is the running time, the higher is the efficiency.

When formalizing these intuitive considerations, we should specify a number of elements:

- **Model of computations**: What our computer can do, in particular, what are its “elementary operations”?
- **Encoding of program instances**: What are programs we intend to solve and what are the “data of particular programs” the computer works with?
- **Quality of solution**: Solution of what kind we expect to get? An exactly optimal or an approximate one? Even for simple convex programs, it would be unrealistic to expect that the data can be converted to an exactly optimal solution in finitely many elementary operations!
Real Arithmetics Complexity Model

Model of computations: idealized computer capable to store arbitrary many reals and to perform exactly the following standard operations with reals:
- four arithmetic operations
- comparisons
- computing elementary functions like log, exp, $\sqrt{\_\_\_}$, sin,...
(idealization comes from the assumption that reals can be stored and processed exactly!)

Generic optimization problem: a family of Mathematical Programming problems of a given “analytical structure”, like Linear, Conic Quadratic and Semidefinite Programming.
Formally: a generic optimization problem $\mathcal{P}$ is a set of “instances” – optimization programs

$$\min_x \left\{ p_0(x) : x \in X(p) \subset \mathbb{R}^{n(p)} \right\} \quad (p)$$

where every instance $(p) \in \mathcal{P}$ is specified by a finite-dimensional data vector $\text{Data}(p)$.

The dimension of the data vector is called the size of an instance:

$$\text{Size}(p) = \text{dim Data}(p).$$
Examples:

- **Linear Programming** $\mathcal{LP}$: collection of all possible LP programs

  \[
  \min_x \left\{ c^T x : Ax \geq b \right\} \quad [A : m \times n],
  \]

  the data vector of an instance being

  \[
  \left( n, m, c^T, \text{Vec}(A), b^T \right)^T
  \]

  where for $A \in \mathbb{M}^{m,n}$

  \[
  \text{Vec}(A) = (A_{11}, ..., A_{1n}, A_{21}, ..., A_{2n}, ..., A_{m1}, ..., A_{mn}).
  \]

- **Conic Quadratic Programming** $\mathcal{CQP}$: collection of all possible conic quadratic programs

  \[
  \min_x \left\{ c^T x : \|D_i x - d_i\|_2 \leq e_i^T x - c_i, \ i = 1, ..., k \right\} \quad [D_i : m_i \times n]
  \]

  the data vector of an instance being

  \[
  \left( n, k, m_1, ..., m_k, c^T, \text{Vec} \left( \left[ \begin{array}{cc} D_{i1} & d_1 \\ e_i & c_1 \end{array} \right] \right), ..., \text{Vec} \left( \left[ \begin{array}{cc} D_{ik} & d_k \\ e_k & c_k \end{array} \right] \right) \right)^T
  \]
• **Semidefinite programming** $SDP$: collection of all possible semidefinite programs

\[
\min_x \left\{ c^T x : \sum_{i=1}^n x_i A_i - B \succeq 0 \right\} \quad [A_i \in S^m]
\]

the data vector of an instance being

\[
(n, m, c, \text{Vec}(A_1), ..., \text{Vec}(A_n), \text{Vec}(B))^T.
\]
**Accuracy of approximate solutions:** Let \( \mathcal{P} \) be a generic convex optimization problem. Assume that it is equipped with *infeasibility measure* 

\[
\text{Infeas}_\mathcal{P}(x, p)
\]

– a real-valued function of \((p) \in \mathcal{P}\) and \(x \in \mathbb{R}^{n(p)}\) which is nonnegative everywhere, is zero if \(x \in X(p)\) and is convex in \(x\).

Given an infeasibility measure, we can define the notion of an \(\epsilon\)-solution to an instance

\[
(p) : \min_x \{ p_0(x) : x \in X(p) \subseteq \mathbb{R}^{n(p)} \}
\]

of \(\mathcal{P}\) as a point \(x \in \mathbb{R}^{n(p)}\) which is both \(\epsilon\)-feasible and \(\epsilon\)-optimal:

\[
\text{Infeas}_\mathcal{P}(x, p) \leq \epsilon \quad \text{and} \quad p_0(x) - \text{Opt}(p) \leq \epsilon,
\]

where

\[
\text{Opt}(p) \equiv \begin{cases} 
\inf_{X(p)} p_0(x), & X(p) \neq \emptyset \\
+\infty, & \text{otherwise}
\end{cases}
\]

is the optimal value of \((p)\).
Example: Natural infeasibility measures for $\mathcal{LP}$, $\mathcal{CQP}$, $\mathcal{SDP}$ are given by the following construction: An instance of the generic problem $\mathcal{P}$ in question is a conic problem of the form

$$\min_x \left\{ c^T(p)x : A(p)x - b(p) \in K(p) \right\}$$

The infeasibility measure is

$$\text{Infeas}_\mathcal{P}(x, p) = \min_t \left\{ t \geq 0 : A(p)x - b(p) + te[K(p)] \in K(p) \right\},$$

where $e[K]$ is the "central point" of cone $K$, specifically,

- vector of 1's of appropriate dimension, when $K$ is a nonnegative orthant;
- the vector $(0, ..., 0, 1)^T$, when $K$ is the Lorentz cone $L^m$;
- the unit matrix of appropriate size, when $K$ is a semidefinite cone;
- the direct sum of the central points of the direct factors, when $K$ is a direct product of the aforementioned standard cones.
Let $\mathcal{P}$ be a generic optimization problem. A solution method $\mathcal{M}$ for $\mathcal{P}$ is a code for the Real Arithmetics computer such that when loaded by $\mathcal{M}$ and getting on input the data vector $\text{Data}(p)$ of an instance $(p) \in \mathcal{P}$ and $\epsilon > 0$, the computer in finitely many operations returns

- either an $n(p)$-dimensional vector $\text{Res}_\mathcal{M}(p, \epsilon)$ which is an $\epsilon$-solution to $(p)$,
- or a correct message "$(p)$ is infeasible",
- or a correct message "$(p)$ is below unbounded".

The complexity of a solution method $\mathcal{M}$ on input $((p), \epsilon)$ is

$$\text{Compl}_\mathcal{M}(p, \epsilon) = \# \text{ of real arithmetic operations carried out on input (Data}(p), \epsilon)$$
• The **complexity** of a solution method $\mathcal{M}$ on input $((p),\epsilon)$ is

\[
\text{Compl}_\mathcal{M}(p,\epsilon) = \# \text{ of real arithmetic operations carried out on input } (\text{Data}(p),\epsilon)
\]

• A solution method is called **polynomial time** ("theoretically efficient") on $\mathcal{P}$, if its complexity is bounded by a polynomial of the size of $(p)$ and the "number of accuracy digits":

\[
\exists \text{ polynomial } \pi : \forall (p) \in \mathcal{P} \forall \epsilon > 0 : \text{Compl}_\mathcal{M}(p,\epsilon) \leq \pi (\text{Size}(p), \text{Digits}(p,\epsilon))
\]

\[
\text{Digits}(p,\epsilon) = \ln \left( \frac{\text{Size}(p) + \|\text{Data}(p)\|_1 + \epsilon^2}{\epsilon} \right)
\]

\[
\left[ \text{Size}(p) = \dim \text{Data}(p), \|u\|_1 = \sum_{i=1}^{\dim} u_i \right]
\]

• A generic optimization problem $\mathcal{P}$ is called **polynomially solvable** ("computationally tractable"), if it admits a polynomial time solution method.
A polynomial time method:

\[ \exists \text{ polynomial } \pi : \forall (p) \in P \ \forall \epsilon > 0 : \text{Compl}_M(p, \epsilon) \leq \pi (\text{Size}(p), \text{Digits}(p, \epsilon)) \]

\[
\text{Digits}(p, \epsilon) = \ln \left( \frac{\text{Size}(p) + \|\text{Data}(p)\|_1 + \epsilon^2}{\epsilon} \right)
\]

\[
\left[ \text{Size}(p) = \dim \text{Data}(p), \|u\|_1 = \sum_{i=1}^{\dim u} |u_i| \right]
\]

For a polynomial time method, increasing by absolute constant factor (say, by 10) computer’s performance, we can increase by (another) absolute constant factor the size of instances which can be processed in a fixed time and the number of accuracy digits to which the instances are processed in this time. In contrast to this,

- for a solution method with exponential in Size(·) complexity like

\[
\text{Compl}_M(p, \epsilon) \approx f(\epsilon) \exp\{\text{Size}(p)\}
\]

10-fold progress in computer power allows to increase the sizes of problems solvable to a fixed accuracy in a fixed time only by additive absolute constant \( \approx 2 \).

- for a solution method with sublinear in \( 1/\epsilon \) complexity like

\[
\text{Compl}_M(p, \epsilon) \approx f(\text{Size}(p)) \frac{1}{\epsilon}
\]

10-fold progress in computer power allows to increase the # of accuracy digits available in a fixed time only by additive absolute constant \( \approx 1 \).
The complexity bound of a typical polynomial time method is just linear in the # of accuracy digits:

\[ \text{Compl}_M(p, \epsilon) \leq O(1)\text{Size}^\alpha(p)\text{Digits}(p, \epsilon). \]

For such a method, polynomially means that the “arithmetic cost” of an extra accuracy digit is independent of the position of the digit (is it the 1-st or the 10,000-th) and is polynomial in the dimension of the data vector.
Polynomial Solvability of Convex Programming

- We are about to prove that under “mild assumptions” a generic convex optimization problem $\mathcal{P}$ is polynomially solvable.

The assumptions are
- Polynomial computability
- Polynomial growth
- Polynomial boundedness of feasible sets.
1. Polynomial computability

• We say that a generic optimization problem

\[ P = \left\{ (p) : \min_x \{ p_0(x) \mid x \in X(p) \in \mathbb{R}^{n(p)} \} \right\} \]

is polynomially computable, if

1.1. There exists a code \( C_{\text{obj}} \) for the Real Arithmetics computer which, given on input the data vector \( \text{Data}(p) \) of an instance \((p) \in P\) and a vector \( x \in \mathbb{R}^{n(p)} \), reports on output the value \( p_0(x) \) and a subgradient \( p_0'(x) \) of the objective of \((p)\) at \( x \), and the \# of operations in course of this computation \( T_{\text{obj}}(x,p) \) is bounded by a polynomial of \( \text{Size}(p) = \text{dim Data}(p) \):

\[
\forall ((p) \in P, x \in \mathbb{R}^{n(p)}) : \quad T_{\text{obj}}(x,p) \leq \chi \text{Size}^\chi(p)
\]

From now on, \( \chi \) stands for positive constants "characteristic for \( P \" and independent of particular choice of \((p) \in P, \epsilon > 0, \text{etc.}\)
1.2. There exists a code $C_{\text{cons}}$ for the Real Arithmetics computer which, given on input the data vector $\text{Data}(p)$ of an instance $(p) \in \mathcal{P}$, a vector $x \in \mathbb{R}^{n(p)}$ and $\epsilon > 0$, reports on output whether $\text{Infeas}_{\mathcal{P}}(x, p) \leq \epsilon$, and if it is not the case, returns vector $e$ which separates $x$ and the set \{ $y : \text{Infeas}_{\mathcal{P}}(y, p) \leq \epsilon$\}:

$$\text{Infeas}_{\mathcal{P}}(y, p) < \epsilon \Rightarrow e^T x > e^T y.$$ and the # of operations in course of this computation $T_{\text{cons}}(x, \epsilon, p)$ is bounded by a polynomial of $\text{Size}(p)$ and $\text{Digits}(p, \epsilon)$:

$$\forall \left(\begin{array}{c} (p) \in \mathcal{P} \\ x \in \mathbb{R}^{n(p)} \\ \epsilon > 0 \end{array}\right) : T_{\text{cons}}(x, \epsilon, p) \leq \chi (\text{Size}(p) + \text{Digits}(p, \epsilon))^\chi.$$
2. Polynomial growth

- We say that a generic optimization problem

\[ \mathcal{P} = \left\{ (p) : \min_x \left\{ p_0(x) : x \in X(p) \in \mathbb{R}^{n(p)} \right\} \right\} \]

is of polynomial growth, if the objectives and the infeasibility measures, as functions of \( x \), grow polynomially with \( \|x\|_1 \), the degree of the polynomial being a power of \( \text{Size}(p) \):

\[ \forall (p) \in \mathcal{P}, x \in \mathbb{R}^{n(p)} : \\
|p_0(x)| + \text{Infeas}_\mathcal{P}(x, p) \leq (\chi [\text{Size}(p) + \|x\|_1 + \|\text{Data}(p)\|_1])(\chi^{\text{Size}(p)}) \]
3. Polynomial boundedness of feasible sets

- We say that a generic optimization problems $\mathcal{P}$ has polynomially bounded feasible sets, if the feasible set $X(p)$ of every instance $p \in \mathcal{P}$ is bounded and is contained in the centered at the origin Euclidean ball of “not too large” radius:

$$\forall p \in \mathcal{P} : X(p) \subset \{ x \in \mathbb{R}^{n(p)} : \|x\|_2 \leq (\chi [\text{Size}(p) + \|\text{Data}(p)\|_1])^{\chi \text{Size}(p)} \}.$$  

♣ It is easily seen that the generic convex programs $\mathcal{LP}$, $\mathcal{CQP}$, $\mathcal{SDP}$ (same as basically all other generic convex programs) satisfy the assumptions of polynomial computability and polynomial growth.

At the same time, $\mathcal{LP}$, $\mathcal{CQP}$, $\mathcal{SDP}$ (and most of other generic convex programs) “as they are” do not satisfy the assumption of polynomial boundedness. We can enforce polynomial boundedness of feasible sets by rejecting to deal with instances where an upper bound on the norm of a feasible solution is not stated explicitly. To this end we pass from a generic problem $\mathcal{P}$ to the problem $\mathcal{P}_b$ with instances $(p^+) = ((p), R)$:

$$(p) : \min_x \{ p_0(x) : x \in X(p) \}$$

$$\Rightarrow (p^+) : \min_x \{ p_0(x) : x \in X_R(p) = \{ x \in X(p) : \|x\|_\infty \leq R \} \}$$

$$[\text{Data}(p^+) = (\text{Data}(p), R)]$$

Note that $\mathcal{LP}_b \subset \mathcal{LP}$; $\mathcal{CQP}_b \subset \mathcal{CQP}$; $\mathcal{SDP}_b \subset \mathcal{SDP}$ and the generic programs $\mathcal{LP}_b$, $\mathcal{CQP}_b$, $\mathcal{SDP}_b$ satisfy the assumption of polynomial boundedness of feasible sets (same as the assumptions of polynomial computability and polynomial growth).
Theorem [Polynomial Solvability of Convex Programming] Let $\mathcal{P}$ be a generic convex optimization problem which is
(a) polynomially computable
(b) of polynomial growth
(c) with polynomially bounded feasible sets.
Then $\mathcal{P}$ is polynomially solvable.
Key Component: Ellipsoid Algorithm

♣ Consider an optimization program
\[ f_* = \min_X f(x) \quad (P) \]

• \( X \subset \mathbb{R}^n \) is a closed and bounded convex set with a nonempty interior;
• \( f \) is a continuous convex function on \( \mathbb{R}^n \).

♠ Assume that our “environment” when solving \((P)\) is as follows:

A. We have access to a \textit{Separation Oracle} \( \text{Sep}(X) \) for \( X \) – a routine which, given on input a point \( x \in \mathbb{R}^n \), reports whether \( x \in X \), and in the case of \( x \notin X \), returns a \textit{separator} – a vector \( e \neq 0 \) such that
\[ e^T x \geq \max_{y \in X} e^T y \]

B. We have access to a \textit{First Order} Oracle which, given on input a point \( x \in X \), returns the value \( f(x) \) and a \textit{subgradient} \( f'(x) \) of \( f \) at \( x \):
\[ \forall y : f(y) \geq f(x) + (y - x)^T f'(x). \]

\textbf{Note:} When \( f \) is differentiable, one can set \( f'(x) = \nabla f(x) \).

C. We are given positive reals \( R, r, V \) such that for some (unknown) \( c \) one has
\[ \{ x : \|x - c\| \leq r \} \subset X \subset \{ x : \|x\|_2 \leq R \} \]

and
\[ \max_{x \in X} f(x) - \min_{x \in X} f(x) \leq V. \]
**Example:** Consider an optimization program

\[
\min_x \left\{ f(x) \equiv \max_{1 \leq \ell \leq L} [p_{\ell} + q_{\ell}^T x] : x \in X = \{ x : a_i^T x \leq b_i, 1 \leq i \leq m \} \right\}
\]

W.l.o.g. we assume that \( a_i \neq 0 \) for all \( i \).

**♠ A Separation Oracle** can be as follows: given \( x \), the oracle checks whether \( a_i^T x \leq b_i \) for all \( i \). If it is the case, the oracle reports that \( x \in X \), otherwise it finds \( i = i_x \) such that \( a_{i_x}^T x > b_{i_x} \), reports that \( x \not\in X \) and returns \( a_{i_x} \) as a separator. This indeed is a separator:

\[
y \in X \Rightarrow a_{i_x}^T y \leq b_{i_x} < a_{i_x}^T x
\]

**♠ A First Order Oracle** can be as follows: given \( x \), the oracle computes the quantities \( p_\ell + q_\ell^T x \) for \( \ell = 1, ..., L \) and identifies the largest of these quantities, which is exactly \( f(x) \), along with the corresponding index \( \ell \), let it be \( \ell_x \): 

\[
f(x) = p_{\ell_x} + q_{\ell_x}^T x.
\]

The oracle returns the computed \( f(x) \) and, as a subgradient \( f'(x) \), the vector \( q_{\ell_x} \). This indeed is a subgradient:

\[
f(y) \geq p_{\ell_x} + q_{\ell_x}^T y = [p_{\ell_x} + q_{\ell_x}^T x] + (y - x)^T q_{\ell_x} = f(x) + (y - x)^T f'(x).
\]
\[ f^*_x = \min_x f(x) \quad \text{(P)} \]

- \( X \subset \mathbb{R}^n \) is a closed and bounded convex set with a nonempty interior;
- \( f \) is a continuous convex function on \( \mathbb{R}^n \).
- We have access to a **Separation Oracle** which, given on input a point \( x \in \mathbb{R}^n \), reports whether \( x \in X \), and in the case of \( x \notin X \), returns a separator \( e \neq 0 \):
  \[ e^T x \geq \max_{y \in X} e^T y \]
- We have access to a **First Order** Oracle which, given on input a point \( x \in X \), returns the value \( f(x) \) and a subgradient \( f'(x) \) of \( f \):
  \[ \forall y : f(y) \geq f(x) + (y - x)^T f'(x) \]
- We are given positive reals \( R, r, V \) such that for some (unknown) \( c \) one has
  \[ \{ x : \|x - c\| \leq r \} \subset X \subset \{ x : \|x\|_2 \leq R \} \]
  and
  \[ \max_{x \in X} f(x) - \min_{x \in X} f(x) \leq V. \]

♠ **How to build a good solution method for (P)?**

To get an idea, let us start with univariate case.

4.22
Univariate Case: Bisection

♣ When solving a problem \( \min_x \{ f(x) : x \in X = [a, b] \subset [-R, R] \} \), by bisection, we recursively update localizers – segments \( \Delta_t = [a_t, b_t] \) containing the optimal set \( X_{\text{opt}} \).

- **Initialization:** Set \( \Delta_1 = [-R, R] \supset X_{\text{opt}} \)
\[
\min_{x} \{f(x) : x \in X = [a, b] \subset [-R, R]\},
\]

- **Step t:** Given \(\Delta_t \supset X_{\text{opt}}\) let \(c_t\) be the midpoint of \(\Delta_t\). Calling Separation and First Order oracles at \(e_t\), we replace \(\Delta_t\) by *twice smaller* localizer \(\Delta_{t+1}\).

\[
\begin{array}{c}
\text{1.a)} \quad \text{1.b)} \\
\text{2.a)} \quad \text{2.b)} \\
\text{2.c)} \\
\end{array}
\]

1) \(\text{Sep}_X\) says that \(c_t \notin X\) and reports, via separator \(e\), on which side of \(c_t\) \(X\) is.
   - 1.a): \(\Delta_{t+1} = [a_t, c_t]\)
   - 1.b): \(\Delta_{t+1} = [c_t, b_t]\)

2) \(\text{Sep}_X\) says that \(c_t \in X\), and \(O_f\) reports, via \(f'(c_t)\), on which side of \(c_t\) \(X_{\text{opt}}\) is.
   - 2.a): \(\Delta_{t+1} = [a_t, c_t]\)
   - 2.b): \(\Delta_{t+1} = [c_t, b_t]\)
   - 2.c): \(c_t \in X_{\text{opt}}\)

\(\blacklozenge\) *Since the localizers rapidly shrink and \(X\) is of positive length, eventually some of search points will become feasible, and the nonoptimality of the best found so far feasible search point will rapidly converge to 0 as process goes on.*

4.24
\[ \text{Opt}(P) = \min_{x \in X \subset \mathbb{R}^n} f(x) \quad (P) \]

.Bisection admits multidimensional extension, called \textit{Generic Cutting Plane Algorithm}, where one builds a sequence of “shrinking” localizers \( G_t \) – closed and bounded convex domains containing the optimal set \( X_{\text{opt}} \) of \( (P) \).

Generic Cutting Plane Algorithm is as follows:

\textbf{Initialization} Select as \( G_1 \) a closed and bounded convex set containing \( X \) and thus being a localizer.
\[ \text{Opt}(P) = \min_{x \in X \subseteq \mathbb{R}^n} f(x) \quad (P) \]

\textbf{♠ Step} \ t = 1, 2, \ldots: Given current localizer \( G_t \),

- Select current \textit{search point} \( c_t \in G_t \) and call Separation and First Order oracles to form a \textit{cut} – to find \( e_t \neq 0 \) s.t. \( X_{\text{opt}} \subset \hat{G}_t := \{ x \in G_t : e^T_t x \leq e^T_t c_t \} \)

To this end
— call Sep\(_X\), \( c_t \) being the input. If Sep\(_X\) says that \( c_t \notin X \) and returns a separator, take it as \( e_t \) (case A on the picture).

\textbf{Note:} \( c_t \notin X \Rightarrow \text{all points from } G_t \setminus \hat{G}_t \text{ are infeasible} \)
— if \( c_t \in X_t \), call \( O_f \) to compute \( f(c_t), f'(c_t) \). If \( f'(c_t) = 0 \), terminate, otherwise set \( e_t = f'(c_t) \) (case B on the picture).

\textbf{Note:} \textit{When} \( f'(c_t) = 0 \), \( c_t \) \textit{is optimal for} \( (P) \), otherwise \( f(x) > f(c_t) \) \textit{at all feasible} \( x \in G_t \setminus \hat{G}_t \)

- By the two “Note” above, \( \hat{G}_t \) is a localizer along with \( G_t \). Select a closed and bounded convex set \( G_{t+1} \supset \hat{G}_t \) (it also will be a localizer) and pass to step \( t + 1 \).
\begin{align*}
\text{Opt}(P) &= \min_{x \in X \subseteq \mathbb{R}^n} f(x) \\
\end{align*}

\textbf{Summary:} Given current localizer \( G_t \), selecting a point \( c_t \in G_t \) and calling the Separation and the First Order oracles, we can

\begin{itemize}
  \item in the \textit{productive case} \( c_t \in X \), find \( e_t \) such that
  \begin{align*}
  e_t^T (x - c_t) > 0 &\Rightarrow f(x) > f(c_t) \\
  \end{align*}
  \item in the \textit{non-productive case} \( c_t \not\in X \), find \( e_t \) such that
  \begin{align*}
  e_t^T (x - c_t) > 0 &\Rightarrow x \not\in X \\
  \end{align*}
\end{itemize}

\( \Rightarrow \) the set \( \hat{G}_t = \{ x \in G_t : e_t^T (x - c_t) \leq 0 \} \) is a localizer

\textbullet We can select as the next localizer \( G_{t+1} \) any set containing \( \hat{G}_t \).

\textbullet We define approximate solution \( x^t \) built in course of \( t = 1, 2, \ldots \) steps as the best – with the smallest value of \( f \) – of the \textit{feasible} search points \( c_1, \ldots, c_t \) built so far.

If in course of the first \( t \) steps no feasible search points were built, \( x^t \) is undefined.
\[ \text{Opt}(P) = \min_{x \in X \subset \mathbb{R}^n} f(x) \quad (P) \]

◆ Analysing Cutting Plane algorithm
- Let \( \text{Vol}(G) \) be the \( n \)-dimensional volume of a closed and bounded convex set \( G \subset \mathbb{R}^n \).

**Note:** For convenience, we use, as the unit of volume, the volume of \( n \)-dimensional unit ball \( \{ x \in \mathbb{R}^n : \|x\|_2 \leq 1 \} \), and not the volume of \( n \)-dimensional unit box.
- Let us call the quantity \( \rho(G) = \left[ \text{Vol}(G) \right]^{1/n} \) the *radius* of \( G \). \( \rho(G) \) is the radius of \( n \)-dimensional ball with the same volume as \( G \), and this quantity can be thought of as the average linear size of \( G \).

**Theorem.** Let convex problem \( (P) \) satisfying our standing assumptions be solved by Generic Cutting Plane Algorithm generating localizers \( G_1, G_2, \ldots \) and ensuring that \( \rho(G_t) \to 0 \) as \( t \to \infty \). Let \( \bar{t} \) be the first step where \( \rho(G_{\bar{t}+1}) < \rho(X) \). Starting with this step, approximate solution \( x^t \) is well defined and obeys the “error bound”

\[
f(x^t) - \text{Opt}(P) \leq \min_{\tau \leq t} \left[ \frac{\rho(G_{\tau+1})}{\rho(X)} \right] \left[ \max_{X} f - \min_{X} f \right]
\]
\[
\text{Opt}(P) = \min_{x \in X \subset \mathbb{R}^n} f(x) \quad (P)
\]

**Explanation:** Since \( \text{int } X \neq \emptyset \), \( \rho(X) \) is positive, and since \( X \) is closed and bounded, \( (P) \) is solvable. Let \( x_* \) be an optimal solution to \( (P) \).
- Let us fix \( \epsilon \in (0, 1) \) and set \( X_\epsilon = x_* + \epsilon(X-x_*) \).
  \( X_\epsilon \) is obtained by similarity transformation which keeps \( x_* \) intact and “shrinks” \( X \) towards \( x_* \) by factor \( \epsilon \). This transformation multiplies volumes by \( \epsilon^n \Rightarrow \rho(X_\epsilon) = \epsilon \rho(X) \).
- Let \( t \) be such that \( \rho(G_{t+1}) < \epsilon \rho(X) = \rho(X_\epsilon) \). Then \( \text{Vol}(G_{t+1}) < \text{Vol}(X_\epsilon) \Rightarrow \text{the set } X_\epsilon \setminus G_{t+1} \text{ is nonempty} \Rightarrow \text{for some } z \in X, \text{ the point } \)
  \[ y = x_* + \epsilon(z-x_*) = (1-\epsilon)x_* + \epsilon z \]
  does not belong to \( G_{t+1} \).
- \( G_1 \) contains \( X \) and thus \( y \), and \( G_{t+1} \) does not contain \( y \), implying that for some \( \tau \leq t \), it holds
  \[ e_\tau^T y > e_\tau^T c_\tau \quad (!) \]
- We definitely have \( c_\tau \in X \) – otherwise \( e_\tau \) separates \( c_\tau \) and \( X \ni y \), and \( (!) \) witnesses otherwise.
  \[ \Rightarrow c_\tau \in X \Rightarrow e_\tau = f'(c_\tau) \Rightarrow f(c_\tau) + e_\tau^T(y-c_\tau) \leq f(y) \]
  \[ \Rightarrow [\text{by } (!)] \]
  \[ f(c_\tau) \leq f(y) = f((1-\epsilon)x_* + \epsilon z) \leq (1-\epsilon)f(x_*) + \epsilon f(z) \]
  \[ \Rightarrow f(c_\tau) - f(x_*) \leq \epsilon[f(z) - f(x_*)] \leq \epsilon \left[ \max_{X} f - \min_{X} f \right]. \]

**Bottom line:** If \( 0 < \epsilon < 1 \) and \( \rho(G_{t+1}) < \epsilon \rho(X) \), then \( x^t \) is well defined (since \( \tau \leq t \) and \( c_\tau \) is feasible) and \( f(x^t) - \text{Opt}(P) \leq \epsilon \left[ \max_{X} f - \min_{X} f \right] \).
\[
\text{Opt}(P) = \min_{x \in X \subset \mathbb{R}^n} f(x) \quad (P)
\]

“Starting with the first step \( \bar{t} \) where \( \rho(G_{t+1}) < \rho(X) \), \( x^t \) is well defined, and

\[
f(x^t) - \text{Opt} \leq \min_{\tau \leq t} \left[ \frac{\rho(G_{\tau+1})}{\rho(X)} \right] \left[ \max_{X} f - \min_{X} f \right]
\]

\[\epsilon_t\]

♣ We are done. Let \( t \geq \bar{t} \), so that \( \epsilon_t < 1 \), and let \( \epsilon \in (\epsilon_t, 1) \). Then for some \( t' \leq t \) we have

\[
\rho(G_{t'+1}) < \epsilon \rho(X)
\]

⇒ [by bottom line] \( x^{t'} \) is well defined and

\[
f(x^{t'}) - \text{Opt}(P) \leq \epsilon V
\]

⇒ [since \( f(x^t) \leq f(x^{t'}) \) due to \( t \geq t' \)] \( x^t \) is well defined and \( f(x^t) - \text{Opt}(P) \leq \epsilon V
\]

⇒ [passing to limit as \( \epsilon \to \epsilon_t + 0 \)] \( x^t \) is well defined and \( f(x^t) - \text{Opt}(P) \leq \epsilon_t V
\]

□
Opt(P) = \min_{x \in X \subset \mathbb{R}^n} f(x) \hspace{1cm} (P)

\begin{itemize}
  \item [\spadesuit] Corollary: Let (P) be solved by cutting Plane Algorithm which ensures, for some \( \vartheta \in (0, 1) \), that \( \rho(G_{t+1}) \leq \vartheta \rho(G_t) \). Then, for every desired accuracy \( \epsilon > 0 \), finding feasible \( \epsilon \)-optimal solution \( x_\epsilon \) to (P) (i.e., a feasible solution \( x_\epsilon \) satisfying \( f(x_\epsilon) - \text{Opt} \leq \epsilon \)) takes at most

  \[ N = \frac{1}{\ln(1/\vartheta)} \ln \left( \mathcal{R} \left[ 1 + \frac{V}{\epsilon} \right] \right) + 1 \]

  steps of the algorithm. Here

  \[ \mathcal{R} = \frac{\rho(G_1)}{\rho(X)} \]

  says how well, in terms of volume, the initial localizer \( G_1 \) approximates \( X \), and

  \[ V = \max_X f - \min_X f \]

  is the variation of \( f \) on \( X \).

  Note: \( \mathcal{R} \) and \( V/\epsilon \) are under log, implying that high accuracy and poor approximation of \( X \) by \( G_1 \) cost “nearly nothing.”

What matters, is the factor at the log which is the larger the closer \( \vartheta < 1 \) is to 1.

4.31
“Academic” Implementation: Centers of Gravity

♠ In high dimensions, to ensure progress in volumes of subsequent localizers in a Cutting Plane algorithm is not an easy task: we do not know how the cut through $c_t$ will pass, and thus should select $c_t$ in $G_t$ in such a way that whatever be the cut, it cuts off the current localizer $G_t$ a “meaningful” part of its volume.

♠ The most natural choice of $c_t$ in $G_t$ is the center of gravity:

$$c_t = \left(\frac{\int_{G_t} xdx}{\int_{G_t} 1dx}\right),$$

the expectation of the random vector uniformly distributed on $G_t$.

**Good news:** The Center of Gravity policy with $G_{t+1} = \hat{G}_t$ results in

$$\vartheta = \left(1 - \left[\frac{n}{n+1}\right]^n\right)^{1/n} \leq [0.632...]^{1/n} \quad (*)$$

This results in the complexity bound (♯ of steps needed to build $\epsilon$-solution)

$$N = 2.2n \ln \left(R \left[1 + \frac{V}{\epsilon}\right]\right) + 1$$

**Note:** It can be proved that within absolute constant factor, like 4, this is the best complexity bound achievable by whatever algorithm for convex minimization which can “learn” the objective via First Order oracle only.
Reason for (*): Brunn-Minkowski Symmeterization Principle:

Let $Y$ be a convex compact set in $\mathbb{R}^n$, $e$ be a unit direction and $Z$ be “equi-cross-sectional” to $X$ body symmetric w.r.t. $e$, so that

- $Z$ is rotationally symmetric w.r.t. the axis $e$
- for every hyperplane $H = \{x : e^T x = \text{const}\}$, one has
  \[
  \text{Vol}_{n-1}(X \cap H) = \text{Vol}_{n-1}(Z \cap H)
  \]

Then $Z$ is a convex compact set.

Equivalently: Let $U, V$ be convex compact nonempty sets in $\mathbb{R}^n$. Then

\[
\text{Vol}^{1/n}(U + V) \geq \text{Vol}^{1/n}(U) + \text{Vol}^{1/n}(V)
\]

In fact, convexity of $U, V$ is redundant!
Disastrously bad news: Centers of Gravity are not implementable, unless the dimension $n$ of the problem is like 2 or 3.

Reason: We have no control on the shape of localizers. When started with a polytope $G_1$ given by $M$ linear inequalities (e.g., a box), $G_t$ for $t \gg n$ can be a more or less arbitrary polytope given by $M + t - 1$ linear inequalities. Computing center of gravity of a general-type high-dimensional polytope is a computationally intractable task — it requires astronomically many computations already in the dimensions like 5 – 10.

Remedy: Maintain the shape of $G_t$ simple and convenient for computing centers of gravity, sacrificing, if necessary, the value of $\vartheta$.

The most natural implementation of this remedy is enforcing $G_t$ to be ellipsoids. As a result,

- $c_t$ becomes computable in $O(n^2)$ operations (nice!)
- $\vartheta = [0.632...]^{1/n} \approx \exp\{-0.367/n\}$ increases to $\vartheta \approx \exp\{-0.5/n^2\}$, spoiling the complexity bound

$$N = 2.2n \ln \left( R \left[ 1 + \frac{V}{\epsilon} \right] \right) + 1$$

to

$$N = 4n^2 \ln \left( R \left[ 1 + \frac{V}{\epsilon} \right] \right) + 1$$

(unpleasant, but survivable...)

4.34
 Ellipsoid in $\mathbb{R}^n$ is the image of the unit $n$-dimensional ball under one-to-one affine mapping:

$$E = E(B, c) = \{x = Bu + c : u^T u \leq 1\}$$

where $B$ is $n \times n$ nonsingular matrix, and $c \in \mathbb{R}^n$.

- $c$ is the center of ellipsoid $E = E(B, c)$: when $c + h \in E$, $c - h \in E$ as well
- When multiplying by $n \times n$ matrix $B$, $n$-dimensional volumes are multiplied by $|\text{Det}(B)|$

$$\Rightarrow \text{Vol}(E(B, c)) = |\text{Det}(B)|, \quad \rho(E(B, c)) = |\text{Det}(B)|^{1/n}. $$
\[ E = E(B, c) = \{ x = Bu + c : u^T u \leq 1 \} \]

**Simple fact:** Let \( E(B, c) \) be ellipsoid in \( \mathbb{R}^n \) and \( e \in \mathbb{R}^n \) be a nonzero vector. The “half-ellipsoid”

\[ \hat{E} = \{ x \in E(B, c) : e^T x \leq e^T c \} \]

is covered by the ellipsoid \( E^+ = E(B^+, c^+) \) given by

\[
c^+ = c - \frac{1}{n+1} Bp, \quad p = B^T e / \sqrt{e^T B B^T e}
\]

\[
B^+ = \frac{n}{\sqrt{n^2-1}} B + \left( \frac{n}{n+1} - \frac{n}{\sqrt{n^2-1}} \right) (Bp)p^T,
\]

- \( E(B^+, c^+) \) is the ellipsoid of the smallest volume containing the half-ellipsoid \( \hat{E} \), and the volume of \( E(B^+, c^+) \) is **strictly smaller** than the one of \( E(B, c) \):

\[
\varphi := \frac{\rho(E(B^+, c^+))}{\rho(E(B, c))} \leq \exp\{-\frac{1}{2n^2}\}.
\]

- Given \( B, c, e \), computing \( B^+, c^+ \) costs \( O(n^2) \) arithmetic operations.
Opt\( (P) = \min_{x \in X \subset \mathbb{R}^n} f(x) \) \hspace{1cm} (P)

Ellipsoid method is the Cutting Plane Algorithm where

- all localizers \( G_t \) are ellipsoids:
  \[ G_t = E(B_t, c_t), \]
- the search point at step \( t \) is \( c_t \), and
- \( G_{t+1} \) is the smallest volume ellipsoid containing the half-ellipsoid
  \[ \widehat{G}_t = \{ x \in G_t : e_t^T x \leq e_t^T c_t \} \]

Computationally, at every step of the algorithm we once call the Separation oracle \( \text{Sep}_X \), (at most) once call the First Order oracle \( \mathcal{O}_f \) and spend \( O(n^2) \) operations to update \((B_t, c_t)\) into \((B_{t+1}, c_{t+1})\) by explicit formulas.

Complexity bound of the Ellipsoid algorithm is
\[
N = 4n^2 \ln \left( \mathcal{R} \left[ 1 + \frac{V}{\epsilon} \right] \right) + 1
\]

\[
\mathcal{R} = \frac{\rho(G_1)}{\rho(X)} \leq \frac{R}{r}, \quad V = \max_{x \in X} f(x) - \min_{x \in X} f(x)
\]

Pay attention:
- \( \mathcal{R}, V, \epsilon \) are under log \( \Rightarrow \text{large magnitudes in data entries and high accuracy are not issues} \)
- the factor at the log depends only on the structural parameter of the problem (its design dimension \( n \)) and is independent of the remaining data.
What is Inside Simple Fact

♠ Messy formulas describing the updating

$$(B_t, c_t) \rightarrow (B_{t+1}, c_{t+1})$$

in fact are easy to get.

- Ellipsoid $E$ is the image of the unit ball $B$ under affine transformation. Affine transformation preserves ratio of volumes
  
  $\Rightarrow$ Finding the smallest volume ellipsoid containing a given half-ellipsoid $\hat{E}$ reduces to finding the smallest volume ellipsoid $B^+$ containing half-ball $\hat{B}$:

- The “ball” problem is highly symmetric, and solving it reduces to a simple exercise in elementary Calculus.
Why Ellipsoids?

When enforcing the localizers to be of “simple and stable” shape, why we make them ellipsoids (i.e., affine images of the unit Euclidean ball), and not something else, say paralleloptopes (affine images of the unit box)?

Answer: In a “simple stable shape” version of Cutting Plane Scheme all localizers are affine images of some fixed $n$-dimensional solid $C$ (closed and bounded convex set in $\mathbb{R}^n$ with a nonempty interior). To allow for reducing step by step volumes of localizers, $C$ cannot be arbitrary. What we need is the following property of $C$:

One can fix a point $c$ in $C$ in such a way that whatever be a cut

$$\hat{C} = \{ x \in C : e^T x \leq e^T c \} \quad [e \neq 0]$$

this cut can be covered by the affine image of $C$ with the volume less than the one of $C$:

$$\exists B, b : \hat{C} \subset BC + b \& |\text{Det}(B)| < 1 \quad (!)$$

Note: The Ellipsoid method corresponds to unit Euclidean ball in the role of $C$ and to $c = 0$, which allows to satisfy (!) with $|\text{Det}(B)| \leq \exp\{-\frac{1}{2n}\}$, finally yielding $\vartheta \leq \exp\{-\frac{1}{2n^2}\}$. 

4.39
• Solids $C$ with the above property are “rare commodity.” For example, $n$-dimensional box does not possess it.

• Another “good” solid is $n$-dimensional simplex (this is not that easy to see!). Here (!) can be satisfied with $|\text{Det}(B)| \leq \exp\{-O(1/n^2)\}$, finally yielding $\vartheta = (1 - O(1/n^3))$.

$\Rightarrow$ From the complexity viewpoint, “simplex” Cutting Plane algorithm is worse than the Ellipsoid method.

The same is true for handful of other known so far (and quite exotic) “good solids.”
Ellipsoid Method: pro’s & con’s

♣ Academically speaking, Ellipsoid method is an indispensable tool underlying basically all results on efficient solvability of generic convex problems, most notably, the famous theorem of L. Khachiyan (1978) on polynomial time solvability of Linear Programming with rational data in Rational Arithmetic Complexity model.

♠ What matters from theoretical perspective, is “universality” of the algorithm (nearly no assumptions on the problem except for convexity) and complexity bound of the form “structural parameter outside of log, all else, including required accuracy, under the log.”

♣ Another theoretical (and to some extent, also practical) advantage of the Ellipsoid algorithm is that as far as the representation of the feasible set $X$ is concerned, all we need is a Separation oracle, and not the list of constraints describing $X$. The number of these constraints can be astronomically large, making impossible to check feasibility by looking at the constraints one by one; however, in many important situations the constraints are “well organized,” allowing to implement Separation oracle efficiently.
Theoretically, the only (and minor!) drawbacks of the algorithm is the necessity for the feasible set $X$ to be bounded, with known “upper bound,” and to possess nonempty interior. As of now, there is not way to cure the first drawback without sacrificing universality. The second “drawback” is artifact: given nonempty set

$$X = \{x : g_i(x) \leq 0, 1 \leq i \leq m\},$$

we can extend it to

$$X^\epsilon = \{x : g_i(x) \leq \epsilon, 1 \leq i \leq m\},$$

thus making the interior nonempty, and minimize the objective within accuracy $\epsilon$ on this larger set, seeking for $\epsilon$-optimal $\epsilon$-feasible solution instead of $\epsilon$-optimal and exactly feasible one.

This is quite natural: to find a feasible solution is, in general, not easier than to find an optimal one. Thus, either ask for exactly feasible and exactly optimal solution (which beyond LO is unrealistic), or allow for controlled violation in both feasibility and optimality!
From practical perspective, theoretical drawbacks of the Ellipsoid method become irrelevant: for all practical purposes, bounds on the magnitude of variables like $10^{100}$ are the same as no bounds at all, and infeasibility like $10^{-10}$ is the same as feasibility. And since the bounds on the variables and the infeasibility are under log in the complexity estimate, $10^{100}$ and $10^{-10}$ are not a disaster.

Practical limitations (rather severe!) of Ellipsoid algorithm stem from method’s sensitivity to problem’s design dimension $n$. Theoretically, with $\epsilon, V, R$ fixed, the number of steps grows with $n$ as $n^2$, and the effort per step is at least $O(n^2)$ a.o.

⇒ Theoretically, computational effort grows with $n$ at least as $O(n^4)$,
⇒ $n$ like 1000 and more is beyond the “practical grasp” of the algorithm.

Note: Nearly all modern applications of Convex Optimization deal with $n$ in the range of tens and hundreds of thousands!
♠ By itself, growth of *theoretical* complexity with $n$ as $n^4$ is not a big deal: for Simplex method, this growth is exponential rather than polynomial, and nobody dies – in reality, Simplex does *not* work according to its disastrous theoretical complexity bound. Ellipsoid algorithm, unfortunately, works more or less according to its complexity bound.

⇒ *Practical scope of Ellipsoid algorithm is restricted to convex problems with few tens of variables.*

**However:** Low-dimensional convex problems from time to time do arise in applications. More importantly, these problems arise “on a permanent basis” as auxiliary problems within some modern algorithms aimed at solving extremely *large-scale* convex problems.

⇒ *The scope of practical applications of Ellipsoid algorithm is nonempty, and within this scope, the algorithm, due to its ability to produce high-accuracy solutions* (and surprising stability to rounding errors) *can be considered as the method of choice.*
How It Works

\[ \text{Opt} = \min_x f(x), \; X = \{ x \in \mathbb{R}^n : a_i^T x - b_i \leq 0, \; 1 \leq i \leq m \} \]

♠ Real-life problem with \( n = 10 \) variables and \( m = 81,963,927 \) “well-organized” linear constraints:

<table>
<thead>
<tr>
<th>CPU, sec</th>
<th>( t )</th>
<th>( f(x^t) )</th>
<th>( f(x^t) - \text{Opt} \leq )</th>
<th>( \rho(G_t) / \rho(G_1) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>1</td>
<td>0.000000</td>
<td>6.7e4</td>
<td>1.0e0</td>
</tr>
<tr>
<td>0.53</td>
<td>63</td>
<td>0.000000</td>
<td>6.7e3</td>
<td>4.2e-1</td>
</tr>
<tr>
<td>0.60</td>
<td>176</td>
<td>0.000000</td>
<td>6.7e2</td>
<td>8.9e-2</td>
</tr>
<tr>
<td>0.61</td>
<td>280</td>
<td>0.000000</td>
<td>6.6e1</td>
<td>1.5e-2</td>
</tr>
<tr>
<td>0.63</td>
<td>436</td>
<td>0.000000</td>
<td>6.6e0</td>
<td>2.5e-3</td>
</tr>
<tr>
<td>1.17</td>
<td>895</td>
<td>-1.615642</td>
<td>6.3e-1</td>
<td>4.2e-5</td>
</tr>
<tr>
<td>1.45</td>
<td>1250</td>
<td>-1.983631</td>
<td>6.1e-2</td>
<td>4.7e-6</td>
</tr>
<tr>
<td>1.68</td>
<td>1628</td>
<td>-2.020759</td>
<td>5.9e-3</td>
<td>4.5e-7</td>
</tr>
<tr>
<td>1.88</td>
<td>1992</td>
<td>-2.024579</td>
<td>5.9e-4</td>
<td>4.5e-8</td>
</tr>
<tr>
<td>2.08</td>
<td>2364</td>
<td>-2.024957</td>
<td>5.9e-5</td>
<td>4.5e-9</td>
</tr>
<tr>
<td>2.42</td>
<td>2755</td>
<td>-2.024996</td>
<td>5.7e-6</td>
<td>4.1e-10</td>
</tr>
<tr>
<td>2.66</td>
<td>3033</td>
<td>-2.024999</td>
<td>9.4e-7</td>
<td>7.6e-11</td>
</tr>
</tbody>
</table>
Similar problem with $n = 30$ variables and $m = 1,462,753,730$ “well-organized” linear constraints:

<table>
<thead>
<tr>
<th>CPU, sec</th>
<th>$t$</th>
<th>$f(x^t)$</th>
<th>$f(x^t) - \text{Opt} \leq$</th>
<th>$\rho(G_t)/\rho(G_1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>1</td>
<td>0.000000</td>
<td>5.9e5</td>
<td>1.0e0</td>
</tr>
<tr>
<td>1.56</td>
<td>649</td>
<td>0.000000</td>
<td>5.9e4</td>
<td>5.0e-1</td>
</tr>
<tr>
<td>1.95</td>
<td>2258</td>
<td>0.000000</td>
<td>5.9e3</td>
<td>8.1e-2</td>
</tr>
<tr>
<td>2.23</td>
<td>4130</td>
<td>0.000000</td>
<td>5.9e2</td>
<td>8.5e-3</td>
</tr>
<tr>
<td>5.28</td>
<td>7080</td>
<td>-19.044887</td>
<td>5.9e1</td>
<td>8.6e-4</td>
</tr>
<tr>
<td>10.13</td>
<td>10100</td>
<td>-46.339639</td>
<td>5.7e0</td>
<td>1.1e-4</td>
</tr>
<tr>
<td>15.42</td>
<td>13308</td>
<td>-49.683777</td>
<td>5.6e-1</td>
<td>1.1e-5</td>
</tr>
<tr>
<td>19.65</td>
<td>16627</td>
<td>-50.034527</td>
<td>5.5e-2</td>
<td>1.0e-6</td>
</tr>
<tr>
<td>25.12</td>
<td>19817</td>
<td>-50.071008</td>
<td>5.4e-3</td>
<td>1.1e-7</td>
</tr>
<tr>
<td>31.03</td>
<td>23040</td>
<td>-50.074601</td>
<td>5.4e-4</td>
<td>1.1e-8</td>
</tr>
<tr>
<td>37.84</td>
<td>26434</td>
<td>-50.074959</td>
<td>5.4e-5</td>
<td>1.0e-9</td>
</tr>
<tr>
<td>45.61</td>
<td>29447</td>
<td>-50.074996</td>
<td>5.3e-6</td>
<td>1.2e-10</td>
</tr>
<tr>
<td>52.35</td>
<td>31983</td>
<td>-50.074999</td>
<td>1.0e-6</td>
<td>2.0e-11</td>
</tr>
</tbody>
</table>
Consider a generic Convex Programming problem $\mathcal{P}$ which is polynomially computable, of polynomial growth and with polynomially bounded feasible sets. In order to solve an instance

$$\min_{x \in X(\mathcal{P})} p_0(x) \quad (p)$$

within accuracy $\epsilon$, we act as follows:

• We rewrite $(p)$ as

$$\min_{x \in X(p)} p_0(x), \quad X = \{x : \|x\|_2 \leq R, \text{Infeas}(\mathcal{P}, x, p) \leq \epsilon\} \quad (*)$$

where $R$ is the a priori bound on the size of $X(p)$ given by the polynomial boundedness feasible sets assumption.

• The polynomial computability assumption allows to equip $(*)$ with First Order and Separation oracles.

• Assuming $(p)$ feasible, polynomial growth assumption allows to bound from above $\text{Var}_R(p_0)$ and to bound from below the radius $r > 0$ of a ball contained in the feasible set of $(*)$.

We now are in a position to solve $(*)$ by the Ellipsoid method. The complexity bound for the method combines with the bounds on the effort to mimic the First Order and the Separation oracles to yield a polynomial-time bound on the complexity of finding $\epsilon$-solution to $(p)$. 4.47
The theorem on polynomial time solvability of Convex Programming is “constructive” – we can explicitly point out the underlying polynomial time solution algorithm (e.g., the Ellipsoid method). However, from the practical viewpoint this is a kind of “existence theorem” – the resulting complexity bounds, although polynomial, are “too large” for practical large-scale computations.

The intrinsic drawback of the Ellipsoid method (and all other “universal” polynomial time methods in Convex Programming) is that the method utilizes just the convex structure of instances and is unable to facilitate our a priori knowledge of the particular analytic structure of these instances.

- In late 80’s, a new family of polynomial time methods for “well-structured” generic convex programs was found – the Interior Point methods which indeed are able to facilitate our knowledge of the analytic structure of instances.
- \( \mathcal{LP}, \mathcal{CQP}, \mathcal{SDP} \) are especially well-suited for processing by the IP methods, and these methods yield the best known so far theoretical complexity bounds for the indicated generic problems.
As far as practical computations are concerned and high-accuracy solutions are sought, the IP methods

- in the case of Linear Programming, are competitive (to say the least) with the Simplex method

- in the case of Conic Quadratic and Semidefinite Programming, are the best known so far numerical techniques.
V. INTERIOR POINT ALGORITHMS FOR LP/CQP/SDP
Preliminaries: The Newton method and the Interior Penalty Scheme

♠ The classical Newton method for unconstrained minimization of a smooth convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ with an open domain is the linearization scheme for solving the Fermat equation

$$\nabla f(x) = 0.$$  \hspace{1cm} (*)&

Given current iterate $x_t$, we linearize (*) at $x_t$:

$$\nabla f(x) \approx \nabla f(x_t) + \nabla^2 f(x_t)(x - x_t);$$

the next iterate is the solution to the linearized Fermat equation:

$$\nabla f(x_t) + \nabla^2 f(x_t)(x - x_t) = 0$$

$$\Rightarrow \quad x_{t+1} = x_t - [\nabla^2 f(x_t)]^{-1}\nabla f(x_t) \quad \text{(Nwt)}$$

• Assuming that $x_*$ is a nondegenerate minimum of $f$:

$$\nabla f(x_*) = 0, \quad \nabla^2 f(x_*) > 0,$$

the Newton method converges to $x_*$ quadratically, *provided that it is started close enough to $x_*$*

$$\exists (r > 0, C < \infty) : \|x_t - x_*\|_2 \leq r \Rightarrow \|x_{t+1} - x_*\|_2 \leq C\|x_t - x_*\|_2^2 \leq \frac{1}{2}\|x_t - x_*\|_2.$$  \hspace{1cm} (5.1)

• In order to ensure *global* convergence of the method, one incorporates linesearch, thus coming to the *damped Newton scheme*

$$x_{t+1} = x_t - \gamma_t[\nabla^2 f(x_t)]^{-1}\nabla f(x_t).$$
A Convex Programming program

$$\min_x \left\{ c^T x : x \in X \subset \mathbb{R}^n \right\} \tag{C}$$

with closed and bounded feasible domain $X$ ($\text{int } X \neq \emptyset$) can be represented as a “limiting case” of convex unconstrained problems.

Indeed, introducing an interior penalty $F(\cdot) : \text{int } X \to \mathbb{R}$ such that

- $F$ is smooth and $\nabla^2 F(x) \succ 0$ for $x \in \text{int } X$,
- $F(x_i) \to \infty$ along every sequence $\{x_i \in \text{int } X\}$ converging to a point $x \in \partial X$,

one can approximate (C) by a “penalized” problem

$$\min_x \left\{ f_t(x) \equiv c^T x + \frac{1}{t} F(x) \right\} \tag{C_t}$$

- For every $t > 0$, $f_t$ is a smooth convex function with the domain int $X$, and $f_t$ attains its minimum on the domain at a unique point $x_*(t)$;
- As $t \to \infty$, the path $x_*(t)$ converges to the solution set of (C).
- In order to solve (C), one can trace the path $x_*(t)$, iterating the updating
  
  (a) $t_i \mapsto t_{i+1} > t_i$
  (b) $x_i \mapsto x_{i+1}$ “close enough” to $x_*(t_{i+1})$

Usually, (b) is obtained by minimizing $f_{t_{i+1}}(\cdot)$ with the (damped) Newton method started at $x_i$. 

5.2
\[ \min_x \left\{ c^T x : x \in X \right\}; \quad F : \text{int } X \rightarrow \mathbb{R} \]

\[ f_t(x) = c^T x + \frac{1}{t} F(x) \]

\[ x_*(t) = \arg\min_x f_t(x) \]

\[ (a) \quad t_i \mapsto t_{i+1} > t_i \]

\[ (b) \quad x_i \mapsto x_{i+1} - [\nabla^2 f_{t_{i+1}}(x_i)]^{-1} \nabla f_{t_{i+1}}(x_i) \]

In 1985-94, it was discovered that

• With an appropriate choice of the interior penalty \( F \), the Interior Penalty Scheme admits a \textit{polynomial time} implementation;

• LP, CQP and SDP are especially well-suited for the resulting IP (Interior Point) methods.
We are interested in a generic conic problem

\[ \min_x \left\{ c^T x : A x - B \in K \right\} \]  

(CP)

where \( K \) is a **canonical cone** – a direct product of several Semidefinite and Lorentz cones:

\[ K = S_{k_1}^{\times} \times \ldots \times S_{k_p}^{\times} \times L_{k_{p+1}} \times \ldots \times L_{k_m} \subset E = S_{k_1}^{\times} \times \ldots \times S_{k_p}^{\times} \times R_{k_{p+1}}^{\times} \times \ldots \times R_{k_m}^{\times}. \]

We equip the Semidefinite and the Lorentz cones by **canonical barriers**:

- The canonical barrier for \( S_{k}^{\times} \) is
  \[ S_k(X) = -\ln \det(X) : \text{int} S_{k}^{\times} \to \mathbb{R}; \]
  the **parameter** of this barrier is \( \theta(S_k) = k \).

- The canonical barrier for \( L_{k} \) is
  \[ L_k(x) = -\ln(x_k^2 - x_1^2 - \ldots - x_{k-1}^2) = -\ln(x^T J_k x), \]
  \[ J_k = \begin{bmatrix} -1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & -1 \end{bmatrix}; \]
  the **parameter** of this barrier is \( \theta(L_k) = 2 \).
The canonical barrier $K$ for $K$ is the direct sum of the canonical barriers of the factors:

$$K(X) = S_{k_1}(X_1) + ... + S_{k_p}(X_p) + L_{k_{p+1}}(X_{p+1}) + ... + L_{k_m}(X_m),$$

where

$$X_i \in \begin{cases} \text{int } S^{k_i}, & i \leq p \\ \text{int } L^{k_i}, & p < i \leq m \end{cases};$$

the parameter of this barrier is the sum of parameters of the components:

$$\theta(K) = \theta(S_{k_1}) + ... + \theta(S_{k_p}) + \theta(L_{k_{p+1}}) + ... + \theta(L_{k_m}) = \sum_{i=1}^{p} k_i + 2(m - p).$$
\[ K = S_1^{k_1} \times \ldots \times S_p^{k_p} \times L_{p+1}^{k_{p+1}} \times \ldots \times L_m^{k_m} \]

\[ K(X) = -\sum_{i=1}^{p} \ln \det(X_i) - \sum_{i=p+1}^{m} \ln(X_i^T J_i X_i), \quad J_k = \begin{bmatrix} -1 & \cdots & -1 \\ \vdots & \ddots & \vdots \\ -1 & \cdots & 1 \end{bmatrix}; \]

\[ \theta(K) = \sum_{i=1}^{p} k_i + 2(m - p). \]

**Elementary properties of canonical barriers:**

- **[barrier property]** \( K(\cdot) \) is \( C^\infty \) strongly convex: \( \nabla^2 K(\cdot) > 0 \) function on \( \text{int} \ K \), and

\[ X^i \in \text{int} \ K, \lim_{i \to \infty} X^i = X \in \partial K \Rightarrow K(X^i) \to \infty, i \to \infty; \]

- **[logarithmic homogeneity]**

\[ X \in \text{int} \ K, t > 0 \Rightarrow K(tX) = K(X) - \theta(K) \ln t \]
\[ \Rightarrow \nabla K(tX) = t^{-1} \nabla K(X); \quad \langle \nabla K(X), X \rangle_E = -\theta(K) \]

- **[self-duality]** The mapping \( X \mapsto -\nabla K(X) \) is a one-to-one mapping of \( \text{int} \ K \) onto \( \text{int} \ K \), and this mapping is self-inverse:

\[ X \in \text{int} \ K, S = -\nabla K(X) \Leftrightarrow S \in \text{int} \ K, X = -\nabla K(S). \]
Central Path

♦ Consider a primal-dual pair of conic problems associated with a canonical cone $K$:

\[
\begin{align*}
\min_x \{ c^T x : A x - B \in K \} & \quad \text{(CP)} \\
\max_S \{ \langle B, S \rangle_E : A^* S = c, S \in K \} & \quad \text{(CD)}
\end{align*}
\]

\[
\Leftrightarrow \begin{align*}
\min_X \{ \langle C, X \rangle_E : X \in (L - B) \cap K \} & \quad \text{(P)} \\
\max_S \{ \langle B, S \rangle_E : S \in (L^\perp + C) \cap K \} & \quad \text{(D)}
\end{align*}
\]

\[
A^* : \langle X, Ax \rangle_E \equiv x^T A^* X
\]

♣ Assume from now on that $\text{Ker} A = \{0\}$ and (CP), (CD) are strictly feasible.

• The canonical barrier of $K$ induces the barrier $F(x) = K(Ax - B)$ for the feasible set of (CP), and thus defines the path

\[
x_*(t) = \arg\min_x \left[ c^T x + \frac{1}{t} F(x) \right]
\]

which turns out to be well-defined for all $t > 0$.

• The image $X_*(t) = Ax_*(t) - B$ of the path $x_*(t)$ is a path in $\text{int} K$ fully characterized by the following two properties:

◊ $X_*(t)$ is strictly primal feasible

♥ $-t^{-1} \nabla K(X_*(t))$ is strictly dual feasible
\[
x_*(t) = \arg\min_x [c^T x + \frac{1}{t} F(x)] \implies X_*(t) = A x_*(t) - B
\]

**Claim:** \(X_*(t)\) is fully characterized by the following two properties:

- \(\diamondsuit\) \(X_*(t)\) is strictly primal feasible
- \(\heartsuit\) \(-t^{-1} \nabla K(X_*(t))\) is strictly dual feasible

Indeed, \(X_*(t)\) is the minimizer of the function \(\langle C, X \rangle_E + t^{-1} K(X)\) over the set of strictly feasible primal solutions, whence

\[
C + t^{-1} \nabla K(X_*(t)) \in \mathcal{L}^\perp;
\]

besides this, \(-t^{-1} \nabla K(X_*(t))\) \(\in \text{int } K\).
\[
\min_{X} \{ \langle C, X \rangle_E : X \in (\mathcal{L} - B) \cap K \} \quad (P) \quad \max_{S} \{ \langle B, S \rangle_E : S \in (\mathcal{L}^\perp + C) \cap K \} \quad (D)
\]

\[\Rightarrow \text{Primal central path } X_*(t): \begin{cases} 
(a) X_*(t) \text{ is strictly primal feasible} \\
(b) -t^{-1} \nabla K(X_*(t)) \text{ is strictly dual feasible}
\end{cases}\]

- Due to primal-dual symmetry, the dual problem (D) defines the dual central path \( S_*(t) \):

\[\begin{cases} 
(c) S_*(t) \text{ is strictly dual feasible} \\
(d) -t^{-1} \nabla K(S_*(t)) \text{ is strictly primal feasible}
\end{cases}\]

- The paths are closely related:

\[X_*(t) = -t^{-1} \nabla K(S_*(t)); \quad S_*(t) = -t^{-1} \nabla K(X_*(t)).\]

Indeed, setting \( S = -t^{-1} \nabla K(X_*(t)) \), we see that \( S \) is strictly dual feasible by (b), while

\[ -t^{-1} \nabla K(S) = -t^{-1} \nabla K(-t^{-1} \nabla K(X_*(t))) = -\nabla K(-\nabla K(X_*(t))) \quad \text{[by logarithmic homogeneity of } K] \\
= X_*(t) \quad \text{[self-duality of } K]\]

i.e., \(-t^{-1} \nabla K(S)\) is strictly primal feasible. Thus, \( S \) satisfies \((c), (d)\), whence

\[ S \equiv -t^{-1} \nabla K(X_*(t)) = S_*(t). \]
\[ \min_x \{ e^T x : Ax - B \in K \} \quad (\text{CP}) \]
\[ \max_S \{ \langle B, S \rangle_E : A^* S = c, S \in K \} \quad (\text{CD}) \]
\[ \Rightarrow \min_X \{ \langle C, X \rangle_E : X \in (L - B) \cap K \} \quad (\text{P}) \]
\[ \max_S \{ \langle B, S \rangle_E : S \in (L^\perp + C) \cap K \} \quad (\text{D}) \]
\[ \Rightarrow \text{Primal-Dual Central Path } (X_*(t), S_*(t)) : \]
\[ \left\{ \begin{array}{l}
X_*(t) \text{ is strictly primal feasible} \\
S_*(t) \text{ is strictly dual feasible} \\
X_*(t) = -t^{-1} \nabla K(S_*(t)) \Leftrightarrow S_*(t) = -t^{-1} \nabla K(X_*(t)).
\end{array} \right. \]

\[ \blacklozenge \quad \text{The Duality Gap on the primal-dual central path equals to } \frac{\theta(K)}{t}. \quad \text{Thus,} \]
\[ X_*(t) \] is \( \frac{\theta(K)}{t} \)-primal optimal, and \( S_*(t) \) is \( \frac{\theta(K)}{t} \)-dual optimal:
\[ \text{DualityGap}(X_*(t), S_*(t)) = [\langle C, X_*(t) \rangle_E - \text{Opt}(\text{P})] + [\text{Opt}(\text{D}) - \langle B, S_*(t) \rangle_E] \]
\[ = \langle S_*(t), X_*(t) \rangle_E = t^{-1} \langle -\nabla K(X_*(t)), X_*(t) \rangle_E \]
\[ = t^{-1} \theta(K). \]

\[ \spadesuit \quad \text{Consequently, our “ideal goal” could be to move along the primal-dual central path, thus approaching the primal and the dual optimal sets.} \]
\[ \textbf{However:} \quad \text{We do not know how to stay on a ”curved” path, although can move close to the path.} \]
In a neighbourhood of the central path

\[
\min_X \{\langle C, X \rangle_E : X \in (\mathcal{L} - B) \cap \mathcal{K} \} \quad (\text{P}) \quad \max_S \{\langle B, S \rangle_E : S \in (\mathcal{L}^\perp + C) \cap \mathcal{K} \} \quad (\text{D})
\]

⇒ Primal-Dual Central Path \((X_*(t), S_*(t))\):
\[
\begin{cases}
X_*(t) \text{ is strictly primal feasible} \\
S_*(t) \text{ is strictly dual feasible} \\
X_*(t) = -t^{-1}\nabla K(S_*(t)) \iff S_*(t) = -t^{-1}\nabla K(X_*(t)).
\end{cases}
\]

♠ Given a triple \((t, X, S)\), where \(X\) is strictly primal feasible, and \(S\) is strictly dual feasible, a good for our purposes measure of closeness of \((X, S)\) to \((X_*(t), S_*(t))\) turns out to be

\[
dist(t, X, S) = \sqrt{\langle [\nabla^2 K(X)]^{-1} [tS + \nabla K(X)], tS + \nabla K(X) \rangle_E}
\]

\[
= \sqrt{\langle [\nabla^2 K(S)]^{-1} [tX + \nabla K(S)], tX + \nabla K(S) \rangle_E}.
\]

The duality gap in an \(O(1)\)-neighbourhood of the primal-dual central path is basically the same as at the central path:

\[
dist(t, X, S) \leq 1 \Rightarrow \text{DualityGap}(X, S) \leq \frac{2\theta(K)}{t}.
\]

♠ Consequently, our “realistic goal” could be to trace the primal-dual central path as \(t \to \infty\), staying in (or periodically entering) an \(O(1)\)-neighbourhood \(\mathcal{N}_{O(1)}\) of the path.

5.10
How to trace the central path?

♦ The central path is given by

<table>
<thead>
<tr>
<th>Strict primal feasibility:</th>
<th>Strict dual feasibility:</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) $X \in \mathcal{L} - B \equiv \text{Im}A - B$</td>
<td>(c) $S \in \mathcal{L}^\perp + C$</td>
</tr>
<tr>
<td>(b) $X \succ 0$</td>
<td>(d) $S \succ 0$</td>
</tr>
</tbody>
</table>

Augmented complementary slackness:

(e) $S + t^{-1}\nabla K(X) = 0$

$G_t(X, S) = 0$

♦ The most natural way to trace the path is as follows:
Given a current triple $t_i, X_i, S_i$ with strictly primal-dual feasible $X_i, S_i$, we
• increase the penalty parameter $t$: $t_i \mapsto t_{i+1} > t_i$;
• linearize at $t_{i+1}, X_i, S_i$ the system of nonlinear equations (e), thus coming to the system of linear equations for the (approximate) “corrections” $\Delta X = X_*(t_{i+1}) - X_i$, $\Delta S = S_*(t_{i+1}) - S_i$:

$$\Delta X \in \mathcal{L}, \Delta S \in \mathcal{L}^\perp, G_{t_{i+1}}(X_i, S_i) + \frac{\partial G_{t_{i+1}}(X_i, S_i)}{\partial X} \Delta X + \frac{\partial G_{t_{i+1}}(X_i, S_i)}{\partial S} \Delta S = 0 \quad (N)$$

• solve (N), thus getting the corrections (“search directions”) $\Delta X_i$, $\Delta S_i$, and update $X_i, S_i$ according to

$$X_{i+1} = X_i + \alpha_i \Delta X_i, \quad S_{i+1} = S_i + \beta_i \Delta S_i.$$
Note: The Augmented Complementary Slackness (ACS) equation can be written in many equivalent forms:

\[ S + t^{-1} \nabla K(X) = 0, \ X + t^{-1} \nabla K(S) = 0, \ldots \]

Different equivalent formulations of ACS equation result in different linearizations and thus in different path-following schemes.

Example: Primal path-following method. Let us use the ACS equation “as it is”:

\[ S + t^{-1} \nabla K(X) = 0. \]

Then the system for corrections becomes

\[ \Delta X = A \Delta x \ \ [\Leftrightarrow \Delta X \in \mathcal{L} = \text{Im}A] \]
\[ A^* \Delta S = 0 \ \ [\Leftrightarrow \Delta S \in \mathcal{L}^\perp] \]

\[ \Delta S + t_i^{-1} [\nabla^2 K(X_i)] \Delta X = -S_i - t_i^{-1} \nabla K(X_i), \]

which is equivalent to

\[ \Delta X = A \Delta x \]
\[ \Delta S = -t_i^{-1} [\nabla^2 K(X_i)] \Delta X - S_i - t_i^{-1} \nabla K(X_i), \]

\[ t_i^{-1} A^* [\nabla^2 K(X_i)] A \Delta x = -\underbrace{A^* S_i}_{c} - t_i^{-1} A^* \nabla K(X_i). \]
Setting

\[ F(x) = K(Ax - B), \]

the method becomes

\[
\begin{align*}
  t_i & \mapsto t_{i+1} > t_i, \\
  x_{i+1} & = x_i - [\nabla^2 F(x_i)]^{-1}[t_{i+1}c + \nabla F(x_i)], \\
  X_{i+1} & = Ax_{i+1} - B, \\
  S_{i+1} & = \ldots
\end{align*}
\]

which is exactly the classical Interior Penalty Scheme for tracing the path

\[ x_\ast(t) = \text{argmin}_x \left[ c^T x + t^{-1}F(x) \right]. \]
\[
\begin{align*}
\min_x \left\{ c^T x : Ax - B \in K \right\} \\
&\Rightarrow \\
&\begin{array}{l}
t_i \mapsto t_{i+1} > t_i, \\
x_{i+1} = x_i - [\nabla^2 F(x_i)]^{-1}[t_{i+1}c + \nabla F(x_i)], \\
F(x) = K(Ax - B);
\end{array} \\
&\begin{array}{l}
X_{i+1} = Ax_{i+1} - B, \\
S_{i+1} = \ldots
\end{array}
\end{align*}
\]

\textbf{Theorem.} Let the starting point \((t_0, X_0, S_0)\) in the Primal Path-Following method belong to the neighbourhood \(\mathcal{N}_{0.1}\) of the central path, i.e.,

- \(t_0 > 0, X_0\) is strictly primal feasible, \(S_0\) is strictly dual feasible;
- \(\sqrt{\langle [\nabla^2 K(X_0)]^{-1}[t_0S_0 + \nabla K(X_0)], t_0S_0 + \nabla K(X_0) \rangle_E} \leq 0.1\).

With the penalty updating rule

\[
t_{i+1} = \left(1 + \frac{0.1}{\sqrt{\theta(K)}}\right)t_i,
\]

the Primal Path-Following method is well-defined and keeps all iterates in \(\mathcal{N}_{0.1}\). In particular, it takes no more than

\[
O(1)\sqrt{\theta(K)} \ln \left(2 + \frac{\theta(K)}{t_0\epsilon} \right)
\]

steps of (PF) to get a feasible \(\epsilon\)-solution of \((CP)\).
Theorem implies the best known so far polynomial time complexity bounds for LP, CQP and SDP.

Writing the Augmented Complementarity Slackness equation in the "symmetric" form

\[ X + t^{-1} \nabla K(S) = 0, \]

one arrives at the *Dual* Path-Following method with exactly the same theoretical properties as the Primal method.
2D feasible set of a toy SDP ($\mathbf{K} = S^3_+.$)

"Continuous curve" is the primal central path
Dots are iterates $x_i$ of the Primal Path-Following method.

<table>
<thead>
<tr>
<th>Itr#</th>
<th>Objective</th>
<th>DualityGap</th>
<th>Itr#</th>
<th>Objective</th>
<th>DualityGap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.100000</td>
<td>2.96</td>
<td>7</td>
<td>-1.359870</td>
<td>8.4e-4</td>
</tr>
<tr>
<td>2</td>
<td>-0.906963</td>
<td>0.51</td>
<td>8</td>
<td>-1.360259</td>
<td>2.1e-4</td>
</tr>
<tr>
<td>3</td>
<td>-1.212689</td>
<td>0.19</td>
<td>9</td>
<td>-1.360374</td>
<td>5.3e-5</td>
</tr>
<tr>
<td>4</td>
<td>-1.301082</td>
<td>6.9e-2</td>
<td>10</td>
<td>-1.360397</td>
<td>1.4e-5</td>
</tr>
<tr>
<td>5</td>
<td>-1.349584</td>
<td>2.1e-2</td>
<td>11</td>
<td>-1.360404</td>
<td>3.8e-6</td>
</tr>
<tr>
<td>6</td>
<td>-1.356463</td>
<td>4.7e-3</td>
<td>12</td>
<td>-1.360406</td>
<td>9.5e-7</td>
</tr>
</tbody>
</table>
Semidefinite Case

♠ In spite of being “theoretically perfect”, Primal and Dual Path-Following methods in practice are inferior as compared with the methods based on “less straightforward” forms of the ACS equation. Let us look at these “more advanced” methods in the SDP case:

\[ K = S_k^+ \subset E = S^k, \quad K(X) = -\ln \det(X). \]

In this case,

- \[ \nabla K(X) = -X^{-1}, \quad [\nabla^2 K(X)] H = X^{-1} H X^{-1}; \]
- The ACS equation reads

\[ S = t^{-1} X^{-1} \iff SX = t^{-1} I. \] (*)

♠ An important class of equivalent representations of (*) is as follows: given a “scaling matrix” \( Q \succ 0 \), one can rewrite (*) in two equivalent forms:

\[ Q^{-1} SXQ = t^{-1} I, \quad QXSQ^{-1} = t^{-1} I, \]

whence also

\[ QXSQ^{-1} + Q^{-1} SXQ = 2 t^{-1} I; \] (**)

in fact, (*) and (**) regarded as nonlinear equations with positive definite unknowns \( X, S \) are equivalent to each other.

5.17
\[ QXSQ^{-1} + Q^{-1}SXQ = 2t^{-1}I; \]  \hfill (**)

**Explanation:** Let \( Q \in S^k \) be nonsingular. The *\( Q \)-scaling*  
\[ X \mapsto QXQ \]
is a one-to-one linear mapping of \( S^k \) onto itself, the inverse being the mapping  
\[ X \mapsto Q^{-1}XQ^{-1}. \]

*\( Q \)-scaling is a symmetry of the positive semidefinite cone – it maps the cone onto itself.* 

⇒ Given a primal-dual pair of semidefinite programs  
\[
\begin{align*}
\text{Opt}(P) &= \min_X \{ \text{Tr}(CX) : X \in [\mathcal{L} - B] \cap S^k_+ \} \quad (P) \\
\text{Opt}(D) &= \max_S \{ \text{Tr}(BS) : S \in [\mathcal{L}^\perp + C] \cap S^k_+ \} \quad (D)
\end{align*}
\]
and a nonsingular matrix \( Q \in S^k \), one can pass in \( (P) \) from variable \( X \) to variable \( \hat{X} = QXQ \), while passing in \( (D) \) from variable \( S \) to variable \( \tilde{S} = Q^{-1}SQ^{-1} \). The resulting problems are  
\[
\begin{align*}
\text{Opt}(\hat{P}) &= \min_{\hat{X}} \{ \text{Tr}(\hat{C}\hat{X}) : \hat{X} \in [\hat{\mathcal{L}} - \hat{B}] \cap S^k_+ \} \quad (\hat{P}) \\
\text{Opt}(\tilde{D}) &= \max_{\tilde{S}} \{ \text{Tr}(\tilde{B}\tilde{S}) : \tilde{S} \in [\tilde{\mathcal{L}}^\perp + \tilde{C}] \cap S^k_+ \} \quad (\tilde{D})
\end{align*}
\]

\[ \hat{B} = QBQ, \hat{\mathcal{L}} = \{ QXQ : X \in \mathcal{L} \}, \tilde{C} = Q^{-1}CQ^{-1}, \tilde{\mathcal{L}}^\perp = \{ Q^{-1}SQ^{-1} : S \in \mathcal{L}^\perp \} \]

♠ \( \hat{P} \) and \( \tilde{D} \) are dual to each other, the primal-dual central path of this pair is the image of the primal-dual path of \( (P) \), \( (D) \) under the *primal-dual \( Q \)-scaling*  
\[ (X, S) \mapsto (\hat{X} = QXQ, \tilde{S} = Q^{-1}SQ^{-1}) \]

\( Q \) preserves closeness to the path, etc.

5.18
♠ Writing down the ACS equation as

\[ QXSQ^{-1} + Q^{-1}SXQ = 2t^{-1}I \]  

we in fact

- pass from \((\mathcal{P}), (\mathcal{D})\) to the equivalent primal-dual pair of problems \((\hat{\mathcal{P}}), (\hat{\mathcal{D}})\)
- write down the ACS equation for the latter pair in the simplest primal-dual symmetric form
  \[ \hat{X}\hat{S} + \hat{S}\hat{X} = 2t^{-1}I, \]
- “scale back” to the original primal-dual variables \(X, S\), thus arriving at (!).
\[ QXSQ^{-1} + Q^{-1}SXQ = 2t^{-1}I \] (**) 

- With the ACS equation written in the form of (**) , one can use iteration-dependent scaling matrices \( Q_i \). The system defining the search directions at \( i \)-th iteration becomes

\[
\Delta X \in \mathcal{L}, \quad \Delta S \in \mathcal{L}^\perp,
Q_i[\Delta XS_i + X_i\Delta S]Q_i^{-1} + Q_i^{-1}[S_i\Delta X + \Delta SX_i]Q_i = 2t_{i+1}^{-1}I - Q_iX_iS_iQ_i^{-1} - Q_i^{-1}S_iX_iQ_i
\]

- Popular choices of the scaling matrices \( Q_i \) are:
  - \( Q_i = I \) [Alizadeh-Haeberly-Overton method]
  - \( Q_i = S_i^{1/2} \) [the \( XS \)-method]
  - \( Q_i = X_i^{-1/2} \) [the \( SX \)-method]
  - \( Q_i = \left( X_i^{-1/2}(X_i^{1/2}S_iX_i^{1/2})^{-1/2}X_i^{1/2}S_i \right)^{1/2} \) [Nesterov-Todd method]
Note: The XS-, the SX-, and the NT-method are based on commutative scalings, where the matrices

\[ \tilde{X}_i = Q_i X_i Q_i, \quad \tilde{S}_i = Q_i^{-1} S_i Q_i^{-1} \]

commute with each other. Specifically,

- in the XS-method, \( \tilde{S} = I \)
- in the SX-method, \( \tilde{X} = I \)
- in the NT-method, \( \tilde{S} = \tilde{X} \).
Theorem. Let a strictly-feasible primal-dual pair (P), (D) of semidefinite programs be solved by a primal-dual path-following method based on commutative scalings, and let the penalty updating policy in the method be

\[ t_{i+1} = \left(1 + \frac{0.1}{\sqrt{k}}\right) t_i. \]  

(U)

Assume that the starting triple \((t_0, X_0, S_0)\) is such that
- \(X_0\) is strictly primal feasible, \(S_0\) is strictly dual feasible, \(t_0 = k^{-1} \text{Tr}(X_0S_0)\);
- The triple \((t_0, X_0, S_0)\) is close to the central path:

\[ \text{dist}(t_0, X_0, S_0) := \sqrt{\langle [\nabla^2 K(X_0)]^{-1} [t_0 S_0 + \nabla K(X_0)], t_0 S_0 + \nabla K(X_0) \rangle_E} \]
\[ \equiv \sqrt{\text{Tr}([t_0 X_0^{1/2} S_0 X_0^{1/2} - I]^2)} \leq 0.1. \]

Then the method is well-defined and keeps all iterates in \(N_{0.1}\). In particular, it takes no more than

\[ O(1) \sqrt{k} \ln \left( 2 + \frac{k}{t_0 \epsilon} \right) \]

steps of the method to build feasible \(\epsilon\)-solutions of (P), (D).
To improve the practical performance of primal-dual path-following methods, in actual computations
— the penalty parameter is updated in a “more aggressive,” as compared to (U), fashion;
— the primal-dual methods are allowed to travel in “much wider,” as compared to $\mathcal{N}_{0.1}$, neighbourhoods of the central path.

The constructions and the complexity results we have presented are “incomplete” in the sense that they do not take into account the necessity to come close to the central path before starting path-tracing and do not take care of the case when the pair (P), (D) is not strictly feasible. All these “gaps” can be easily closed via the same path-following technique as applied to appropriate augmented versions of the original problem.
Complexity bounds for $\mathcal{LP}_b$

♣ A program from $\mathcal{LP}_b$:

$$(p) : \min_x \{c^T x : Ax \geq b, \|x\|_\infty \leq R\} \quad [A \in \mathbb{M}^{m,n}]$$

can be solved within accuracy $\epsilon$ in

$$N_{\mathcal{LP}} = O(1) \sqrt{m} \ln \left( \frac{\|\text{Data}(p)\|_1 + \epsilon^2}{\epsilon} \right)$$

iterations.

The computational effort per iteration is dominated by the necessity, given a positive definite diagonal matrix $\Delta$ and a vector $r$, to assemble the matrix and to solve the linear system

$$[A; I; -I]^T \Delta [A; I; -I] x = h$$

and to solve the linear system.

• In the case $m = O(n)$, the overall complexity of solving $(p)$ within accuracy $\epsilon$ is cubic in $n$:

$$O(1)mn^2 \ln \left( \frac{\|\text{Data}(p)\|_1 + \epsilon^2}{\epsilon} \right)$$
Complexity bounds for $\mathcal{CQP}_b$

A program from $\mathcal{CQP}_b$:

$$(p) : \{ c^T x : \|D_i x - d_i\|_2 \leq e_i^T x - c_i, \ i = 1, \ldots, k; \|x\|_2 \leq R \}$$

can be solved within accuracy $\epsilon$ in

$$N_{\mathcal{CQP}} = O(1) \sqrt{k} \ln \left( \frac{\|\text{Data}(p)\|_1 + \epsilon^2}{\epsilon} \right)$$

iterations.

The computational effort per iteration is dominated by the necessity, given vectors $\delta_i, \ i = 1, \ldots, k$ and a vector $r$, to assemble the matrices

$$H_i = D_i^T (I - \delta_i \delta_i^T) D_i, \ i = 1, \ldots, k$$

and to solve a $\dim x \times \dim x$ linear system

$$Hu = r$$

with positive definite matrix $H$ “readily given” by $H_1, \ldots, H_k$. 

5.25
Complexity bounds for $SDP_b$

A program from $SDP_b$:

\[
(p): \quad \min_x \left\{ c^T x : A(x) = \sum_{i=1}^{n} x_i A_i - B \succeq 0, \|x\|_2 \leq R \right\}
\]

can be solved within accuracy $\epsilon$ in

\[
N_{SDP} = O(1) \sqrt{\mu} \ln \left( \frac{\|\text{Data}(p)\|_1 + \epsilon^2}{\epsilon} \right)
\]

iterations, where $\mu$ is the row size of matrices $A_1, ..., A_n$.

The computational effort per iteration is dominated by the necessity, given a positive definite matrix $X$ of the same size and block-diagonal structure as those of $A_i$ and a vector $rs$

- to compute $n \times n$ symmetric matrix $\hat{H}$ with entries

\[
\hat{H}_{ij} = \text{Tr}(X^{-1}A_i X^{-1}A_j), \quad i, j = 1, ..., n;
\]

- to solve $n \times n$ linear system

\[
Hu = r
\]

with positive definite matrix $H$ “readily given” by $\hat{H}$.
VI. FIRST ORDER METHODS
Simple methods for extremely large-scale problems

The arithmetic complexity of a step in all known polynomial time methods for Convex Programming grows up *nonlinearly* with the design dimension $n$ of the problem – at least as $O(n^2)$, if not as $O(n^3)$ (the only exception are extremely sparse real-world LPs with favourable sparsity patterns).

What to do when the design dimension is of order of tens and hundreds of thousands, and the problem is not a “very sparse LP”?

Nonlinear convex problems of huge design dimension do arise in numerous applications, e.g., in

- SDP relaxations of large combinatorial problems,
- Structural Design (especially for 3D structures),
- Signal Processing, High-dimensional Statistics, Machine Learning
- 3D Medical imaging problems
Example of Medical Imaging problem: PET Image Reconstruction

PET (Positron Emission Tomography) is a powerful, non-invasive, medical diagnostic imaging technique for measuring the metabolic activity of cells in the human body. It has been in clinical use since the early 1990s. PET imaging is unique in that it shows the chemical functioning of organs and tissues, while other imaging techniques - such as X-ray, computerized tomography (CT) and magnetic resonance imaging (MRI) - show anatomic structures.
Physics of PET. A PET scan uses radioactive tracer – a biologically active fluid with a radio-active component capable of emitting positrons. When administered to a patient, the tracer distributes within the body and, with properly chosen biologically active “carrier”, concentrates in desired locations, e.g., in the areas of high metabolic activity where cancer tumors can be expected.

- The tracer disintegrates, emitting positrons.
- A positron immediately annihilates with a near-by electron, giving rise to two photons flying at the speed of light off the point of annihilation in nearly opposite directions. They are registered outside the patient by cylindrical PET scanner consisting of several rings of detectors.
- When two detectors “simultaneously” (within $\sim 10^{-8}$ sec time window) are hit by photons, this event is registered, indicating that somewhere on the line linking the detectors (LOR – “Line of Response”) a disintegration act took place.
• The measured data is the collection of numbers of LOR’s counted by different pairs of detectors (“bins”), and the problem is to recover from these measurements the 3D density of the tracer.

♣ Mathematically, the PET Image Reconstruction problem, after appropriate discretization, becomes the problem of recovering a vector $\lambda \geq 0$ from a noisy observation $y$ of the vector $P\lambda$:

$$\lambda \mapsto y = P\lambda + \text{noise} \quad \mapsto \quad \text{estimate of } \lambda.$$ 

Specifically,

• entries of $\lambda$ are indexed by voxels – small cubes into which we partition the field of view; $\lambda_j$ is the average density of the tracer in voxel $j$;

• entries of $y$ are indexed by bins (pairs of detectors); $y_i$ is the number of LORs registered by bin $i$;

• $P = [p_{ij}]$ is a given matrix; $p_{ij}$ is the probability for a LOR originating in voxel $j$ to be registered by bin $i$.

the statistical model of PET states that the entries $y_i$ in $y$ are realizations of independent Poisson random variables with the expectations $(P\lambda)_i$. 

6.4
In the PET Reconstruction problem, we are interested, given observations $y$, to find the Maximum Likelihood estimate $\lambda^*$ of tracer's density:

$$\lambda^* = \arg\min_{\lambda \geq 0} \left[ \sum_{j=1}^{n} p_j \lambda_j - \sum_{i=1}^{m} y_i \ln(\sum_{j} p_{ij} \lambda_j) \right] \quad [p_j = \sum_{i} p_{ij}] \quad \text{(PET)}$$

(PET) is a nicely structured constrained convex program; the only difficulty – a true one! – is in huge sizes of (PET): for problems of actual interest,

- the design dimension $n$ varies from 300,000 to 3,000,000
- the number $m$ of log-terms in the objective varies from 6,000,000 to 25,000,000
As far as nonlinear programs are concerned, design dimension $n \sim 10^4 - 10^5 - 10^6$ makes it necessary to use “cheap” algorithms – those with nearly linear in $n$ arithmetic cost of a step (otherwise you never will finish the very first iteration). This requirement rules out all “advanced” polynomial time optimization techniques and leaves us with, essentially, just two options:

I. Traditional tools of smooth unconstrained minimization: gradient descent, conjugate gradients, quasi-Newton methods, etc.

II. Simple subgradient-type techniques for solving convex nonsmooth constrained optimization problems: subgradient descent, restricted memory bundle methods, etc.
• We are interested in extremely large-scale constrained convex problems, and thus intend to focus on cheap subgradient-type techniques. The question of primary importance here is: (?) What are the limits of performance of cheap optimization techniques?

• When answering (?), we shall restrict ourselves with the black-box-represented convex programs. As a matter of fact, this is exactly the “working environment” for cheap optimization algorithms.
Let us fix a family $\mathcal{P}(X)$ of convex programs

$$\min_x \{ f(x) : x \in X \}; \quad \text{(CP)}$$

where $X \subset \mathbb{R}^n$ is a given instance-independent convex compact set, and $f : \mathbb{R}^n \to \mathbb{R}$ is convex.
\[
\min_x \{ f(x) : x \in X \}; \quad \text{(CP)}
\]

♣ A **black-box-oriented** solution method \( \mathcal{B} \) for \( \mathcal{P}(X) \) is as follows:

- When starting to solve (CP), \( \mathcal{B} \) is given an accuracy \( \epsilon > 0 \) and knows that the problem belongs to a given family \( \mathcal{P}(X) \). However, \( \mathcal{B} \) does **not** know in advance what is the particular problem it deals with.

- When solving the problem, \( \mathcal{B} \) has an access to the First Order oracle for \( f \). Given on input \( x \in \mathbb{R}^n \), the oracle returns \( f(x) \) and a subgradient \( f'(x) \) of \( f \) at \( x \). \( \mathcal{B} \) generates a sequence of **search points** \( x_1, x_2, ... \) and calls the First Order oracle to get values and subgradients of \( f \) at these points. The rules for building \( x_t \) can be arbitrary, except for the fact that they should be **non-anticipative**: \( x_t \) can depend only on the information \( f(x_1), f'(x_1), ..., f(x_{t-1}), f'(x_{t-1}) \) on \( f \) accumulated by \( \mathcal{B} \) at the first \( t - 1 \) steps.

- After a number \( T = T_{\mathcal{B}}(f, \epsilon) \) of calls to the oracle, \( \mathcal{B} \) terminates and outputs a result \( z_{\mathcal{B}}(f, \epsilon) \) which should depend solely on the information on \( f \) accumulated by \( \mathcal{B} \) at the \( T \) search steps, and must be an \( \epsilon \)-solution to (CP):

\[
z_{\mathcal{B}}(f, \epsilon) \in X \& \ f(z_{\mathcal{B}}(f, \epsilon)) - \min_X f \leq \epsilon.
\]
The complexity of \( \mathcal{P}(X) \) w.r.t. a solution method \( B \) is

\[
\text{Compl}_B(\epsilon) = \max_{f \in \mathcal{P}(X)} T_B(f, \epsilon)
\]

which is the minimal number of steps sufficient for \( B \) to solve within accuracy \( \epsilon \) every instance of \( \mathcal{P}(X) \).

The Information-based complexity of a family \( \mathcal{P}(X) \) of problems is

\[
\text{Compl}(\epsilon) = \min_B \text{Compl}_B(\epsilon),
\]

the minimum being taken over all solution methods. Relation

\[
\text{Compl}(\epsilon) = N
\]

means that

- there exists a solution method \( B \) capable to solve within accuracy \( \epsilon \) every instance of \( \mathcal{P}(X) \) in no more than \( N \) calls to the First Order oracle;
- for every solution method \( B \), there exists an instance of \( \mathcal{P}(X) \) such that \( B \) solves the instance within the accuracy \( \epsilon \) in at least \( N \) steps.

The information-based complexity \( \text{Compl}(\epsilon) \) of a family \( \mathcal{P}(X) \) is a lower bound on “actual” computational effort, whatever it means, sufficient to find \( \epsilon \)-solution to every instance of the family.
Main results on Information-based complexity of Convex Programming

Let

\[ X \subset \mathbb{R}^n - \text{a convex compact set, } \text{int} \ X \neq \emptyset \]
\[ \mathcal{P}(X) = \left\{ \left\{ \min_{x \in X} f(x) \right\} : f \text{ is convex on } \mathbb{R}^n \text{ and is normalized by } \max_X f - \min_X f \leq 1. \right\} \]

For the family \( \mathcal{P}(X) \),

I. Complexity of finding high-accuracy solutions in fixed dimension is independent of the geometry of \( X \). Specifically,

\[ \forall (\epsilon \leq \epsilon(X)) : O(1) n \ln \left(2 + \frac{1}{\epsilon} \right) \leq \text{Compl}(\epsilon); \]
\[ \forall (\epsilon > 0) : \text{Compl}(\epsilon) \leq O(1) n \ln \left(2 + \frac{1}{\epsilon} \right), \]

where

\( O(1) \) are appropriately chosen positive absolute constants,
\( \epsilon(X) \) depends on the geometry of \( X \), but never is less than \( \frac{1}{n^2} \).
\( X \subset \mathbb{R}^n \) – a convex compact set, \( \text{int} X \neq \emptyset \)

\[ \mathcal{P}(X) = \left\{ \{ \min_{x \in X} f(x) \} : f \text{ is convex on } \mathbb{R}^n \text{ and normalized by } \max_x f - \min_x f \leq 1 \right\} \]

II. Complexity of finding solutions of fixed accuracy in high dimensions does depend on the geometry of \( X \). Here are 3 typical results:

Let \( X = \{ x \in \mathbb{R}^n : \| x \|_{\infty} \leq 1 \} \). Then

\[ \epsilon \leq \frac{1}{2} \Rightarrow O(1)n \ln(\frac{1}{\epsilon}) \leq \text{Compl}(\epsilon) \leq O(1)n \ln(\frac{1}{\epsilon}). \]  

\((\| \cdot \|_{\infty}-\text{Ball})\)

Let \( X = \{ x \in \mathbb{R}^n : \| x \|_2 \leq 1 \} \). Then

\[ n \geq \frac{1}{\epsilon^2} \Rightarrow \frac{O(1)}{\epsilon^2} \leq \text{Compl}(\epsilon) \leq \frac{O(1)}{\epsilon^2}. \]

\((\| \cdot \|_2-\text{Ball})\)

Let \( X = \{ x \in \mathbb{R}^n : \| x \|_1 \leq 1 \} \). Then

\[ n \geq \frac{1}{\epsilon^2} \Rightarrow \frac{O(1)}{\epsilon^2} \leq \text{Compl}(\epsilon) \leq \frac{O(\ln n)}{\epsilon^2}. \]

\((\| \cdot \|_1-\text{Ball})\)

\((O(1) \text{ in the lower bound can be replaced with } O(\ln n), \text{ provided that } n \gg \frac{1}{\epsilon^2}).\)

6.12
Consequences for large-scale convex minimization:

**Bad news:** I says that we have no hope to guarantee high-accuracy solutions (like $\epsilon = 10^{-6}$) when solving large-scale problems with black-box-oriented methods: it would require at least $O(n)$ calls to the first order oracle with at least $O(n)$ a.o. per call, i.e., totally at least $O(n^2)$ a.o. (with known methods – even $O(n^4)$ a.o.), which is too much for large $n$...

**Good news:** II says that there exist cases when medium accuracy solutions can be found in (nearly) dimension-independent number of oracle calls...
♣ Good news: There exist cases when medium accuracy solutions of convex programs

\[
\min_{x \in X} f(x), \quad \max_X f - \min_X f \leq 1
\]  

(*)
can be found in (nearly) dimension-independent number of oracle calls, e.g., the cases of

\[
X = B_2^n \equiv \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\} \quad (\| \cdot \|_2\text{-Ball})
\]
or

\[
X = B_1^n \equiv \{x \in \mathbb{R}^n : \|x\|_1 \leq 1\} \quad (\| \cdot \|_1\text{-Ball})
\]

(but, unfortunately, not the case when \(X\) is a box).
Problems of minimizing over a \( \| \cdot \|_p \)-ball, \( p = 1, 2 \), are not that typical. Fortunately, the corresponding (nearly) dimension-independent complexity bounds remain valid when \( X \) in (*) is a subset of a “good” set \( B^p_n \), \( p = 1, 2 \), and the normalization condition on \( f \) in (*) is strengthened to

\[
|f(x) - f(y)| \leq \|x - y\|_p \quad \forall x, y \in X.
\]

In particular, \( O\left(\frac{\ln n}{\epsilon^2}\right) \) oracle calls are sufficient to minimize, within accuracy \( \epsilon \), a convex function \( f \) over the standard simplex

\[
\Delta_n = \{x \in \mathbb{R}^n : x \geq 0, \sum_i x_i = 1\},
\]

provided that \( f \) is Lipschitz continuous, with constant 1, w.r.t. \( \| \cdot \|_1 \) (i.e., that the magnitudes of all first order partial derivatives of \( f \) are \( \leq 1 \).)

More good news: The nearly dimension independent complexity bounds for minimization over ball and simplex are given by cheap minimization methods!
Where the lower complexity bounds come from? (cases of ball and box)

Let $2 \leq p \leq \infty$ and $X = \{x : \|x\|_p \leq 1\}$. Consider the families of convex functions

$$F_k = \{f(x) \equiv \max_{1 \leq i \leq k} [\epsilon_i x_i + \delta_i]\}$$

given by all $2^k$ collections $\epsilon_i = \pm 1$ and all collections $\{\delta_i\}_{i=1}^k$ with $0 \leq \delta_i \leq \frac{1}{2^{k^{1/p}}}$.

Observe that when $f \in F_k$, the variation of $f$ on $X$ does not exceed 2, and the $\|\cdot\|_\infty$-Lipschitz constant of $f$ does not exceed 1.

We claim that

(!) For every $k \leq n$, the $\frac{1}{4k^{1/p}}$-complexity of the class of problems $\min_{x \in X} f(x)$ is at least $k - 1$

whence, of course,

(!!) For $0 < \epsilon < \frac{1}{4}$, the $\epsilon$-complexity of the class of optimization problems $\min_X f(x)$ with Lipschitz continuous, with constant 1 w.r.t. $\|\cdot\|_\infty$, objectives $f$ is at least $\min[n, \lfloor\frac{1}{4\epsilon}\rfloor^p] - 1$. 

6.16
We should prove that if $B$ is a method for solving problems

$$\min_{x \in X} f_{\epsilon, \delta}(x) = \max_{1 \leq i \leq k} [\epsilon_i x_i + \delta_i]$$

$$[X = \{x \in \mathbb{R}^n : \|x\|_p \leq 1\}]$$

which, as applied to every problem of this type, terminates after at most $k - 1$ steps, then the accuracy to which the method solves at least one problem from the family is worse than $\epsilon \equiv \frac{1}{2k^{1/p}}$.

We lose nothing when assuming that $B$, as applied to every problem from the family, performs exactly $k$ steps, and the approximate solution is the last – the $k$-th – search point.

Let us associate with $B$ the following construction:

**First step.** Let
- $x^1$ be the first search point generated by $B$ (this point depends solely on $B$),
- $i_1$ be the index of the largest in absolute value coordinate of $x^1$,
- $\epsilon_{i_1}^* = \pm 1$ be such that $\epsilon_{i_1}^* x_{i_1}^1 = |x_{i_1}^1|$,
- $\delta_{i_1}^* = \frac{1}{2k^{1/p}}$

We set

$$\mathcal{F}^1 = \left\{ f(x) = \max_{1 \leq i \leq k} [\epsilon_i x_i + \delta_i] : |\epsilon_i| = 1, \epsilon_{i_1} = \epsilon_{i_1}^*, \delta_{i_1} = \delta_{i_1}^* > \max_{i \neq i_1} \delta_i \geq 0 \right\}$$

**Note:** All functions from $\mathcal{F}^1$ coincide with each other in a neighbourhood of $x^1$, so that the Oracle, being asked at $x^1$ about every one of the objectives from $\mathcal{F}^1$, reports the same.
Step $\ell + 1$, $1 \leq \ell < k$. At the beginning of $\ell$-th step, we have $\ell$ points $x^1, \ldots, x^\ell$ and a set of objectives

$$\mathcal{F}^\ell = \left\{ f(x) = \max_{1 \leq i \leq k} [\epsilon_i x_i + \delta_i] : \begin{array}{ll} |\epsilon_i| = 1, & i = 1, \ldots, k \\
\epsilon_i = \epsilon_i^*, & s = 1, \ldots, \ell \\
\delta_i = \delta_i^*, & s = 1, \ldots, \ell \\
\delta_{i_1}^* > \ldots > \delta_{i_\ell}^* > \max_{i \notin \{i_1, \ldots, i_\ell\}} \delta_i \geq 0 \end{array} \right\}$$

such that

(A$\ell$): $x^1, \ldots, x^\ell$ are the first $\ell$ points of the trajectory of $B$ as applied to every objective $f \in \mathcal{F}^\ell$

(B$\ell$): for every $s \leq \ell$, $\max_{i \notin \{i_1, \ldots, i_\ell\}} |x^s_i| \leq |x^s_{i_\ell}| = \epsilon_i x_i^s$

At step $\ell$, we shrink $\mathcal{F}^\ell$ to $\mathcal{F}^{\ell+1}$ and extend $\{x^1, \ldots, x^\ell\}$ to $\{x^1, \ldots, x^{\ell+1}\}$ as follows:

- By (A$\ell$), $x^1, \ldots, x^\ell$ are the first $\ell$ points of the trajectory of $B$ as applied to every one of the objectives $f \in \mathcal{F}^\ell$, and by (B$\ell$) all these objectives are identically equal to each other in a neighbourhood of $\{x^1, \ldots, x^\ell\} \Rightarrow (\ell + 1)$-st point $x^{\ell+1}$ of the trajectory of $B$ as applied to every one of the objectives $f \in \mathcal{F}^\ell$ is the same.

- Consider the coordinates of $x^{\ell+1}$ with indexes different from $i_1, \ldots, i_\ell$, and let $i_{\ell+1}$ be the index of the largest in magnitude of these coordinates. We choose $\epsilon_{i_{\ell+1}}^* = \pm 1$ in such a way that $\epsilon_{i_{\ell+1}}^* x_{i_{\ell+1}}^{\ell+1} = |x_{i_{\ell+1}}^{\ell+1}|$ thus ensuring (B$\ell+1$), choose $\delta_{i_{\ell+1}}^* \in (0, \delta_{i_{\ell+1}}^*)$ and set

$$\mathcal{F}^{\ell+1} = \left\{ f(x) = \max_{1 \leq i \leq k} [\epsilon_i x_i + \delta_i] : \begin{array}{ll} |\epsilon_i| = 1, & i = 1, \ldots, k \\
\epsilon_i = \epsilon_i^*, & s = 1, \ldots, \ell + 1 \\
\delta_i = \delta_i^*, & s = 1, \ldots, \ell + 1 \\
\delta_{i_1}^* > \ldots > \delta_{i_{\ell+1}}^* > \max_{i \notin \{i_1, \ldots, i_{\ell+1}\}} \delta_i \geq 0 \end{array} \right\}$$

thus ensuring (A$\ell+1$).
After \( k \) steps of the construction, we end up with a single-function family

\[
\mathcal{F}^k = \{f_k(x) = \max_{1 \leq s \leq k} [\epsilon^*_i x_{i_s} + \delta^*_i]\}
\]

such that the trajectory \( x^1, \ldots, x^k \) of \( B \) as applied to \( f_k(\cdot) \) satisfies

\[
\epsilon^*_i x^s_{i_s} \geq 0, \ s = 1, \ldots, k,
\]

whence, in particular, \( f_k(x_k) > 0 \). On the other hand,

\[
\min_{x \in X} f_k(x) \leq -\frac{1}{k^{1/p}} + \max_i \delta^*_i = -\frac{1}{k^{1/p}} + \frac{1}{2k^{1/p}} = \epsilon_k \equiv -\frac{1}{2k^{1/p}}.
\]

Thus, the result \( x_k \) of \( B \) as applied to \( f_k(\cdot) \) is not an \( \epsilon_k \)-solution of \( \min_X f_k \), as claimed.
The simplest of the cheapest – Subgradient Descent  
(N. Shor, 1967)

The Subgradient Descent method (SD) for solving a convex program

\[
\min_{x \in X} f(x) \quad (P)
\]

- \(X\) – convex compact set in \(\mathbb{R}^n\)
- \(f\) – Lipschitz continuous on \(X\) convex function

is the recurrence

\[
x_{t+1} = \Pi_X(x_t - \gamma_t f'(x_t)) \quad [x_1 \in X] \quad (SD)
\]

where

- \(\gamma_t > 0\) are stepsizes
- \(\Pi_X(x) = \arg\min_{y \in X} \|x - y\|_2^2\) is the standard projector on \(X\),
- \(f'(x)\) is a subgradient of \(f\) at \(x\):

\[
f(y) \geq f(x) + (y - x)^T f'(x) \quad \forall y \in X.
\]
Note: We always assume that \( \text{int } X \neq \emptyset \) and that the subgradients \( f'(x) \) reported by the First Order oracle at points \( x \in X \) satisfy the requirement

\[
f'(x) \in \text{cl}\{f'(y) : y \in \text{int } X\}.
\]

With this assumption, for every norm \( \| \cdot \| \) on \( \mathbb{R}^n \) and for every \( x \in X \) one has

\[
\|f'(x)\|_* \equiv \max_{\xi : \|\xi\| \leq 1} \xi^T f'(x) \leq L_{\| \cdot \|}(f) \equiv \sup_{x \neq y, x, y \in X} \frac{|f(x) - f(y)|}{\|x - y\|}.
\]
When, why and how SD converges?

\[ x_{t+1} = \Pi_X(x_t - \gamma_t f'(x_t)) \]  

(\text{SD})

\[ \clubsuit \text{ We start with a simple geometric fact:} \]

(!) \text{ Let } X \subset \mathbb{R}^n \text{ be a closed convex set and } x \in \mathbb{R}^n.

\text{Then the vector } e = x - \Pi_X(x) \text{ forms an acute angle with}
\text{every vector of the form } y - \Pi_X(x), y \in X:
\text{(}x - \Pi_X(x)\text{)}^T(y - \Pi_X(x)) \leq 0 \quad \forall y \in X.

\text{In particular,}
\text{ } y \in X \Rightarrow \|y - \Pi_X(x)\|^2 \leq \|y - x\|^2 - \|x - \Pi_X(x)\|^2

Indeed, when \( y \in X \) and \( 0 \leq t \leq 1 \), one has
\[ \phi(t) = \| [\Pi_X(x) + t(y - \Pi_X(x))] - x \|^2 \geq \|\Pi_X(x) - x\|^2 = \phi(0), \]
whence \( 0 \leq \phi'(0) = 2(\Pi_X(x) - x)^T(y - \Pi_X(x)). \) Consequently,
\[ \|y - x\|^2 = \|y - \Pi_X(x)\|^2 + \|\Pi_X(x) - x\|^2 + 2(y - \Pi_X(x))^T(\Pi_X(x) - x) \geq \|y - \Pi_X(x)\|^2 + \|\Pi_X(x) - x\|^2. \]

\textbf{Corollary:} \text{ For every } u \in X \text{ one has}
\[ \gamma_t(x_t - u)^T f'(x_t) \leq \frac{1}{2} \|x_t - u\|^2 + \frac{1}{2} \|x_{t+1} - u\|^2 + \frac{1}{2} \gamma_t^2 \|f'(x_t)\|^2 \]

Indeed, by (!) we have
\[ d_{t+1} \leq \frac{1}{2} \|x_t - u\| - \gamma_t f'(x_t) \|_2^2 = d_t - \gamma_t (x_t - u)^T f'(x_t) + \frac{1}{2} \gamma_t^2 \|f'(x_t)\|^2. \]

6.22
\( f_* = \min_{x \in X} f(x) \)  \hspace{1cm} (1)
\[
x_{t+1} = \Pi_X (x_t - \gamma_t f'(x_t)) \hspace{1cm} (2)
\]
\[
\gamma_t (x_t - u)^T f'(x_t) \leq \frac{d_t}{2} \left( \|x_t - u\|^2_2 - \|x_{t+1} - u\|^2_2 \right) + \frac{1}{2} \gamma_t \|f'(x_t)\|_2^2 \quad \forall u \in X \hspace{1cm} (3)
\]

Summing up inequalities (3) over \( t = T_0, T_0 + 1, ..., T \), we get
\[
\sum_{t=T_0}^{T} \gamma_t (f(x_t) - f(u)) \leq d_{T_0} - d_{T+1} + \sum_{t=T_0}^{T} \frac{1}{2} \gamma_t \|f'(x_t)\|_2^2 \leq \Theta + \sum_{t=T_0}^{T} \frac{1}{2} \gamma_t \|f'(x_t)\|_2^2
\]

\[
[ \Theta = \max_{x,y \in X} \frac{1}{2} \|x - y\|_2^2 ]
\]

Setting \( u = x_* \equiv \arg\min_X f \), we arrive at the bound
\[
\forall (T, T_0, T \geq T_0 \geq 1) : \epsilon_T \equiv \min_{t \leq T} f(x_t) - f_* \leq \Theta + \frac{1}{2} \sum_{t=T_0}^{T} \gamma_t^2 \|f'(x_t)\|_2^2 \sum_{t=T_0}^{T} \gamma_t
\]
\( \forall (T, T_0, T \geq T_0 \geq 1) : \epsilon_T \equiv \min_{t \leq T} f(x_t) - f_* \leq \frac{\Theta + \frac{1}{2} \sum_{t=T_0}^{T} \gamma_t \|f'(x_t)\|_2^2}{\sum_{t=T_0}^{T} \gamma_t} \)

The resulting relation leads to various convergence results.

**Example 1: “Divergent Series”**. Let \( \gamma_t \to 0 \) as \( t \to \infty \), while \( \sum_t \gamma_t = \infty \). Then

\[
\lim_{T \to \infty} \epsilon_T = 0.
\]

**Proof.** Set \( T_0 = 1 \) and note that

\[
\frac{\sum_{t=1}^{T} \gamma_t^2 \|f'(x_t)\|_2^2}{\sum_{t=1}^{T} \gamma_t} \leq L_{\|\cdot\|_2}(f) \frac{\sum_{t=1}^{T} \gamma_t^2}{\sum_{t=1}^{T} \gamma_t} \to 0, \quad T \to \infty.
\]
\[
\begin{align*}
    f_* &= \min_{x \in X} f(x) \\
    \forall (T, T_0, T \geq T_0 \geq 1): 
    \epsilon_T &\equiv \min_{t \leq T} f(x_t) - f_* 
    \leq \frac{\Theta + \frac{1}{2} \sum_{t=T_0}^{T} \gamma_t \|f'(x_t)\|^2}{\sum_{t=T_0}^{T} \gamma_t} \\
    [\Theta = \frac{1}{2} \max_{x,y \in X} \|x - y\|_2^2]
\end{align*}
\]

Example 2: “Optimal stepsizes”:

\[
\gamma_t = \sqrt{2\Theta} \frac{1}{\|f'(x_t)\|_2 \sqrt{t}} \Rightarrow \epsilon_T \equiv \min_{t \leq T} f(x_t) - f_* \leq O(1) \frac{L \|\cdot\|_2(f) \sqrt{\Theta}}{\sqrt{T}}, \quad T \geq 1
\]

Proof. Setting \( T_0 = \lfloor T/2 \rfloor \), we get

\[
\epsilon_T \leq \frac{\Theta + \Theta \sum_{t=T_0}^{T} \frac{1}{t}}{\sum_{t=T_0}^{T} \frac{1}{\sqrt{t} \|f'(x_t)\|_2}} \leq \frac{\Theta + \Theta \sum_{t=T_0}^{T} \frac{1}{t}}{\sum_{t=T_0}^{T} \frac{1}{\sqrt{t} L \|\cdot\|_2(f)}} \leq L \|\cdot\|_2(f) \sqrt{\Theta} \frac{1+O(1)}{O(1) \sqrt{T}} \leq O(1) \frac{L \|\cdot\|_2(f) \sqrt{\Theta}}{\sqrt{T}}
\]
\[ f_\star = \min_{x \in X} f(x) \]

\[ \Rightarrow x_{t+1} = \Pi_X (x_t - \gamma_t f'(x(t))), \quad \gamma_t = \frac{\max_{x,y \in X} \|x-y\|_2}{\sqrt{\|f'(x_t)\|_2}} \]

\[ \Rightarrow \epsilon_T \equiv \min_{1 \leq t \leq T} f(x_t) - f_\star \leq O(1) \frac{L_{\|\cdot\|_2, X}(f) \max_{x,y \in X} \|x-y\|_2}{\sqrt{T}} \]

**Good news:** We have arrived at efficiency estimate which is dimension-independent, provided that the “\(\|\cdot\|_2\)-variation” of the objective on the feasible domain

\[ \text{Var}_{\|\cdot\|_2, X}(f) = L_{\|\cdot\|_2}(f) \max_{x,y \in X} \|x-y\|_2 \]

is fixed. Moreover, when \(X\) is a Euclidean ball in \(\mathbb{R}^n\), this efficiency estimate “is as good as an efficiency estimate of a black-box-oriented method can be”, provided that the dimension is large:

\[ n \geq \left( \frac{\text{Var}_{\|\cdot\|_2, X}(f)/\epsilon}{\epsilon} \right)^2 \]
$$\epsilon_T \equiv \min_{1 \leq t \leq T} f(x_t) - f^* \leq O(1)\text{Var}_{\|\cdot\|_2, X}(f)/\sqrt{T}$$

**Bad news:** Our “dimension-independent” efficiency estimate

• is pretty slow

• is indeed dimension-independent only for problems with “Euclidean geometry” — those with moderate $\| \cdot \|_2$-variation. As a matter of fact, in applications problems of this type are pretty rare.

SD as applied to $\min_{\|x\|_2 \leq 1} \|Ax - b\|_1$, $A : 50 \times 50$

[red: efficiency estimate; blue: actual error]
An evident drawback of SD is that all information on the objective accumulated so far is “summarized” in the current iterate, and this “summary” is very incomplete. With better usage of past information, one arrives at bundle methods which outperform SD significantly in practice, while preserving the most attractive theoretical property of SD – dimension-independent and optimal, in favourable circumstances, rate of convergence.
Bundle-Level method for solving $f_* = \min_{x \in X} f(x)$

At the beginning of step $t$ of BL, we have at our disposal
- the first-order information $\{f(x_\tau), f'(x_\tau)\}_{1 \leq \tau < t}$ on $f$ along the previous search points $x_\tau \in X$, $\tau < t$;
- current iterate $x_t \in X$.

At step $t$ we
- compute $f(x_t), f'(x_t)$; this information, along with the past first-order information on $f$, provides is with the current model of the objective
  $$f_t(x) = \max_{\tau \leq t} [f(x_\tau) + (x - x_\tau)^T f'(x_\tau)]$$

This model underestimates the objective and is exact at the points $x_1, \ldots, x_t$;
- define the best found so far value $f^t = \min_{\tau \leq t} f(x_\tau)$ of $f$.

define the current lower bound $f_t$ on $f_*$ by solving the auxiliary problem
  $$f_t = \min_{x \in X} f_t(x) \quad \text{(LP}_t)$$

**Note:** current gap $\Delta_t = f^t - f_t$ upper-bounds the inaccuracy of the best found so far solution;
- compute the current level $\ell_t = f_t + \lambda \Delta_t$ ($\lambda \in (0, 1)$ is a parameter);
- build a new search point by solving the auxiliary problem
  $$x_{t+1} = \arg\min_{x} \{ \|x - x_t\|_2^2 : x \in X, f_t(x) \leq \ell_t \} \quad \text{(QP}_t)$$

and loop to step $t + 1$.  

6.29
Why and how BL converges?

Preliminary observations:
♠ The models $f_t(x) = \max_{\tau \leq t} [f(x_\tau) + (x - x_\tau)^T f'(x_\tau)]$ grow with $t$ and underestimate $f$, while the best found so far values of the objective decrease with $t$ and overestimate $f_*$. Thus,

$$f_1 \leq f_2 \leq f_3 \leq \ldots \leq f_*$$
$$f^1 \geq f^2 \geq f^3 \leq \ldots \geq f_*$$
$$\Delta_1 \geq \Delta_2 \geq \ldots \geq 0$$

♠ Let us say that a group of subsequent iterations $J = \{s, s + 1, \ldots, r\}$ form a segment, if $\Delta_r \geq (1 - \lambda) \Delta_s$. We claim that If $J = \{s, s + 1, \ldots, r\}$ is a segment, then

(i) All the sets $L_t = \{x \in X : f_t(x) \leq \ell_t\}$, $t \in J$, have a point in common, specifically, (any) minimizer $u$ of $f_r(\cdot)$ over $X$;

(ii) For $t \in J$, one has $\|x_t - x_{t+1}\|_2 \geq \frac{(1 - \lambda) \Delta_r}{L \|\cdot\|_2(f)}$. 

6.30
We claim that if $J = \{s, s+1, \ldots, r\}$ is a segment, then
(i) All the sets $L_t = \{x \in X : f_t(x) \leq \ell_t\}, t \in J$, have a point in common, specifically, (any) minimizer $u$ of $f_r(\cdot)$ over $X$;
(ii) For $t \in J$, one has $\|x_t - x_{t+1}\|_2 \geq \frac{(1-\lambda)\Delta_r}{L_{\|\cdot\|_2(f)}}$.

Indeed,
(i): for $t \in J$ we have
$$f_t(u) \leq f_r(u) = f_r = f_r - \Delta_r \leq f^t - \Delta_r \leq f^t - (1-\lambda)\Delta_s \leq f^t - (1-\lambda)\Delta_t = \ell_t.$$

(ii): We have $f_t(x_t) = f(x_t) \geq f^t$, and $f_t(x_{t+1}) \leq \ell_t = f^t - (1-\lambda)\Delta_t$. Thus, when passing from $x_t$ to $x_{t+1}$, $t$-th model decreases by at least $(1-\lambda)\Delta_t \geq (1-\lambda)\Delta_r$. It remains to note that $f_t(\cdot)$ is Lipschitz continuous w.r.t. $\|\cdot\|_2$ with constant $L_{\|\cdot\|_2(f)}$. 

6.31
Main observation: The cardinality of a segment $J \equiv \{s, s + 1, \ldots, r\}$ of iterations can be bounded as follows:

$$\text{Card}(J) \leq \frac{\text{Var}^2_{\| \cdot \|_2, X(f)} \lambda}{(1 - \lambda)^2 \Delta_r^2}.$$  

Indeed, when $t \in J$, the sets $L_t = \{x \in X : f_t(x) \leq \ell_t\}$ have a point $u$ in common, and $x_{t+1}$ is the projection of $x_t$ onto $L_t$. It follows that

$$\|x_{t+1} - u\|_2^2 \leq \|x_t - u\|_2^2 - \|x_t - x_{t+1}\|_2^2 \quad \forall t \in J$$

$$\Rightarrow \sum_{t \in J} \|x_t - x_{t+1}\|_2^2 \leq \|x_s - u\|_2^2 \leq \max_{x,y \in X} \|x - y\|_2^2$$

$$\Rightarrow \text{Card}(J) \leq \frac{\max_{x,y \in X} \|x - y\|_2^2}{\min_{t \in J} \|x_t - x_{t+1}\|_2^2} \frac{L^2_{\| \cdot \|_2}(f) \max_{x,y \in X} \|x - y\|_2^2}{(1 - \lambda)^2 \Delta_r^2} \quad [\text{by (ii)}]$$

Corollary: For every $\epsilon$, $0 < \epsilon < \Delta_1$, the number $N$ of steps before a gap $\leq \epsilon$ is obtained (i.e., before an $\epsilon$-solution is found) does not exceed the bound

$$N(\epsilon) = \frac{\text{Var}^2_{\| \cdot \|_2, X(f)} \lambda}{\lambda(1 - \lambda)^2(2 - \lambda)\epsilon^2}.$$
Proof of Corollary. Assume that $N$ is such that $\Delta_N > \epsilon$, and let us bound $N$ from above.

• Let us split the set of iterations $I = \{1, \ldots, N\}$ into segments $J_1, \ldots, J_m$ as follows:  
  - $J_1$ is the maximal segment which ends with iteration $N$:  
    $$J_1 = \{ t : t \leq N, (1-\lambda)\Delta_t \leq \Delta_N \}$$
  - $J_1$ is certain group of subsequent iterations $\{s_1, s_1 + 1, \ldots, N\}$. If $J_1$ differs from $I$: $s_1 > 1$, we define $J_2$ as the maximal segment which ends with iteration $s_1 - 1$:  
    $$J_2 = \{ t : t \leq s_1 - 1, (1-\lambda)\Delta_t \leq \Delta_{s_1-1} \} = \{s_2, s_2 + 1, \ldots, s_1 - 1\}$$
  - If $J_1 \cup J_2$ differs from $I$: $s_2 > 1$, we define $J_3$ as the maximal segment which ends with iteration $s_2 - 1$:  
    $$J_3 = \{ t : t \leq s_2 - 1, (1-\lambda)\Delta_t \leq \Delta_{s_2-1} \} = \{s_3, s_3 + 1, \ldots, s_2 - 1\}$$
  and so on.

• As a result, $I$ will be partitioned “from the end to the beginning” into segments of iterations $J_1, J_2, \ldots, J_m$. Let $d_\ell$ be the gap corresponding to the last iteration from $J_\ell$. By maximality of segments $J_\ell$, we have  
  $$d_1 \geq \Delta_N > \epsilon \& d_{\ell+1} > (1-\lambda)^{-1}d_\ell, \, \ell = 1, 2, \ldots, m - 1$$
whence  
  $$d_\ell > \epsilon(1-\lambda)^{-(\ell-1)}.$$  

We now have  
\[
N = \sum_{\ell=1}^{m} \text{Card}(J_\ell) \leq \sum_{\ell=1}^{m} \frac{\text{Var}_{\|2,x}(f)}{(1-\lambda)^2 d^2_\ell} \leq \frac{\text{Var}_{\|2,x}(f)}{(1-\lambda)^2} \sum_{\ell=1}^{m} (1-\lambda)^{2(\ell-1)} \epsilon^{-2} \leq \frac{\text{Var}_{\|2,x}(f)}{(1-\lambda)^2 \epsilon^2} \sum_{\ell=1}^{\infty} (1-\lambda)^{2(\ell-1)} = \frac{\text{Var}_{\|2,x}(f)}{(1-\lambda)^2 \epsilon^2} \frac{1}{(1-(1-\lambda)^2)\epsilon^2} = N(\epsilon). \]
We have seen that Bundle-Level shares the dimension-independent (and optimal in the “favourable” large-scale case) theoretical complexity bound:

For every \( \epsilon > 0 \), the number of steps before an \( \epsilon \)-solution to convex program \( \min_{x \in X} f(x) \) is found, does not exceed

\[
O(1) \left( \frac{\text{Var}_{\|\cdot\|_2, X(f)}}{\epsilon} \right)^2.
\]

There exists quite convincing experimental evidence that Bundle-Level obeys the optimal in fixed dimension “polynomial time” complexity bound:

For every \( \epsilon \in (0, \text{Var}_X(f) \equiv \max_X f - \min_X f) \), the number of steps before an \( \epsilon \)-solution to convex program \( \min_{x \in X} f(x) \) with \( X \subset \mathbb{R}^n \) is found, does not exceed

\[
n \ln \left( \frac{\text{Var}_X(f)}{\epsilon} \right) + 1.
\]

Experimental rule: When solving convex program with \( n \) variables by BL, every \( n \) steps add new accuracy digit.
Illustration: \[ \min_{x: \|x\|_2 \leq 1} f(x) \equiv \|Ax - b\|_1, \dim x = 50 \] \( f(0) = 2.61, f^* = 0 \)

SD, accuracy vs. iteration count. blue: errors; red: efficiency estimate \( 3 \frac{\text{Var}_{\|x\|_2}(f)}{\sqrt{t}}; \varepsilon_{10000} = 0.084 \)

BL, accuracy vs. iteration count. blue: errors; red: efficiency estimate \( e^{-t/n} \text{Var}_x(f); \varepsilon_{233} < 1.e - 4 \)

6.35
In BL, the number of linear constraints in the auxiliary problems

\[ f_t = \min_{x \in X} f_t(x) \]  

\[ x_{t+1} = \arg \min_{x} \left\{ \|x - x\|_2^2 : x \in X, f_t(x) \leq \ell_t \right\} \]

is equal to the size \( t \) of the current bundle – the collection of affine forms \( g_\tau(x) = f(x_\tau) + (x - x_\tau)^T f'(x_\tau) \) participating in the model \( f_t(\cdot) \). Thus, the complexity of an iteration in BL grows with the iteration number. In order to suppress this phenomenon, one needs a mechanism for shrinking the bundle (and thus – simplifying the models of \( f \)).

The simplest way of shrinking the bundle is to initialize \( d \) as \( \Delta_1 \) and to run plain BL until an iteration \( t \) with \( \Delta_t \leq d/2 \) is met. At such an iteration, we

— shrink the current bundle, keeping in it the minimum number of the forms \( g_\tau \) sufficient to ensure that

\[ f_t \equiv \min_{x \in X} \max_{1 \leq \tau \leq t} g_\tau(x) = \min_{x \in X} \max_{\text{selected } \tau} g_\tau(x) \]

(this number is at most \( n \)),

— reset \( d \) as \( \Delta_t \),

and proceed with plain BL until the gap is again reduced by factor 2, etc.

Computational experience demonstrates that the outlined approach does not slow BL down, while keeping the size of the bundle below the level of about \( 2n \).
Truncated Proximal Level Method for $\min_{x \in X} f(x)$

- In Truncated Proximal Level method, the size of bundle is kept below a given desired level $m$.
- Execution of TLM is split into phases. Phase $s$ is associated with
  - prox-center $c_s \in X$
  - $s$-th upper bound $f^s$ on $f_*$, which is the best value of the objective observed before the phase begins
  - $s$-th lower bound $f_s$ on $f_*$, which is the best lower bound on $f_*$ observed before the phase begins

$f^s$ and $f_s$ define

- $s$-th optimality gap $\Delta_s = f^s - f_s$
- $s$-th level $\ell_s = f_s + \lambda \Delta_s$, where $\lambda \in (0, 1)$ is parameter of the method.
- current model $\tilde{f}^s(\cdot) \leq f(\cdot)$ of $f(\cdot)$, which is the maximum of $\leq m$ affine forms.

To initialize the first phase, we choose $c_1 \in X$, compute $f(c_1), f'(c_1)$ and set

$$\tilde{f}^1(x) = f(c_1) + (x - c_1)^T f'(c_1), \quad f^1 = f(c_1), \quad f_1 = \min_{x \in X} \tilde{f}^1(x).$$
At the beginning of step $t = 1, 2, \ldots$ of phase $s$, we have at our disposal

- upper bound $f_{s,t-1}^{s,t-1} \leq f^s$ on $f^*$, which is the best found so far value of the objective,
- lower bound $f_{s,t-1} \geq f_s$ on $f^*$,
- model $\tilde{f}_{s,t-1}(\cdot) \leq f(\cdot)$ of the objective which is the maximum of $\leq m$ affine forms
- iterate $x_t \in X$ and set

$$H_{t-1} = \{x : \alpha_{t-1}^T x \geq \beta_{t-1}\}$$

such that

$$x \in X, f(x) \leq \ell_s \Rightarrow x \in H_{t-1} \quad (a_t)$$

$$x_t = \arg\min_x \{\|x - c_s\|^2 : x \in X \cap H_{t-1}\} \quad (b_t)$$

To initialize the first step of phase $s$, we set

$$f_{s,0} = f^s, f_{s,0} = f_s, \tilde{f}_{s,0}(\cdot) = \tilde{f}^s(\cdot), \alpha_0 = 0, \beta_0 = 0 \Rightarrow H_0 = \mathbb{R}^n$$

thus ensuring $(a_1)$, and set $x_1 = c_s$, thus ensuring $(b_1)$. 
Step $t$ phase $s$: Given
- bounds $f_{s,t-1} \geq f^*, f_{s,t-1} \leq f^*$,
- model $f_{s,t-1}^*(\cdot) \leq f(\cdot)$,
- $x_t$ and $H_{t-1} = \{ x : \alpha^T_{t-1} x \geq \beta_{t-1} \}$ such that

$$x \in X, f(x) \leq \ell_s \Rightarrow x \in H_{t-1} \ (a_t) \quad \& \quad x_t = \arg\min_x \{ \| x - c_s \|_2^2 : x \in X \cap H_{t-1} \} \quad (b_t)$$

1. we compute $f(x_t), f'(x_t)$ and set $g_t(x) = f(x_t) + (x - x_t)^T f'(x_t)$;
2. we define $\tilde{f}_{s,t}^*(\cdot)$ as the maximum of $g_t(\cdot)$ and affine forms associated with $\tilde{f}_{s,t-1}^*$ (dropping, if necessary, one of the latter forms to make $\tilde{f}_{t,s}^*$ the maximum of at most $m$ forms). If $f(x_t) \leq \ell_s + 0.5(f^s - \ell_s)$ ("significant progress in the upper bound"), we terminate phase $s$ and set

$$f_{s+1} = f_{s,t}, \quad f_{s+1} = f_{s,t-1}, \quad \tilde{f}_{s+1}^*(\cdot) = \tilde{f}_{s,t}^*(\cdot),$$

otherwise we proceed as follows:

3. we compute $f_t = \min_x \{ f_{s,t}(x) : x \in H_{t-1} \cap X \}$. Since $f(x) \geq \ell_s$ in $X \setminus H_{t-1}$, we have $f^* \geq \min[\ell_s, f_t]$, so that $f_{s,t} \equiv \max \{ f_{s,t-1}, \min[\ell_s, f_t] \} \leq f^*$. If $f_{s,t} \geq \ell_s - 0.5(\ell_s - f^*)$ ("significant progress in the lower bound"), we terminate phase $s$ and set

$$f_{s+1} = f_{s,t}, \quad f_{s+1} = f_{s,t}, \quad \tilde{f}_{s+1}^*(\cdot) = \tilde{f}_{s,t}^*(\cdot),$$

otherwise we set

$$x_{t+1} = \arg\min_x \{ \| x - c_s \|_2^2 : x \in X \cap H_{t-1}, \tilde{f}_{s,t}^*(x) \leq \ell_s \}$$

$$H_t = \{ x : (x_{t+1} - c_s)^T (x - x_{t+1}) \geq 0 \}$$

and loop to step $t + 1$ of phase $s$. 

6.39
Step of TPL
\[ x_{t+1} = \arg\min_x \left\{ \|x - c_s\|_2^2 : x \in X \cap H_{t-1}, \tilde{f}^{s,t}(x) \leq \ell_s \right\} \quad (1) \]

\[ H_t = \{ x : (x_{t+1} - c_s)^T(x - x_{t+1}) \geq 0 \} \quad (2) \]

**Note:** When passing to step \( t + 1 \), we have ensured the relations

\[ x \in X, f(x) \leq \ell_s \Rightarrow x \in H_t \quad (a_{t+1}) \]

\[ x_{t+1} = \arg\min_x \left\{ \|x - c_s\|_2^2 : x \in X \cap H_t, \tilde{f}^{s,t}(x) \leq \ell \right\} \quad (b_{t+1}) \]

Indeed, \( x_{t+1} \) is the minimizer of \( \omega_s(x) \equiv \frac{1}{2}\|x - c_s\|_2^2 \) on the set

\[ Y_t = X \cap H_{t-1} \cap \{ x : \tilde{f}^{t,s}(x) \leq \ell_s \} \]

whence

\[ \left[ \omega_s'(x_{t+1}) \right]^T(x - x_{t+1}) \geq 0 \quad \forall x \in Y_t \]

\[ \Downarrow \]

\[ Y_t \subset H_t = \{ x : \left[ \omega_s'(x_{t+1}) \right]^T(x - x_{t+1}) \geq 0 \} \quad (*) \]

Thus,

\[ (x \in X, f(x) \leq \ell_s) \Rightarrow (x \in X \cap H_{t-1}, f(x) \leq \ell_s) \quad (a_t) \]

\[ \Rightarrow (x \in X \cap H_{t-1}, \tilde{f}^{s,t}(x) \leq \ell_s) \Rightarrow x \in H_t \quad x \in Y_t \quad (\ast) \]

as required in \((a_{t+1})\). \((b_{t+1})\) readily follows from the definition of \( H_t \).
Convergence of TPL

♣ Preliminary observations:

• When passing from phase $s$ to phase $s+1$, the optimality gap is decreased at least by the factor

$$\theta(\lambda) = \frac{\min[1 + \lambda, 2 - \lambda]}{2}.$$ 

Indeed, phase $s$ can be terminated at step $t$ due to significant progress either in the upper bound on $f_s$: $f_{s+1} = f_{s,t} \leq \ell_s + \frac{1}{2}(f_s - \ell_s)$

$$\Rightarrow \Delta_{s+1} = f_{s+1} - f_{s+1} \leq \frac{1}{2}\ell_s + \frac{1}{2}f_s - f_s = \frac{1 + \lambda}{2}\Delta_s$$

or in the lower bound: $f_{s+1} = f_{s,t} \geq \ell_s - \frac{1}{2}(\ell_s - f_s)$

$$\Rightarrow \Delta_{s+1} = f_{s+1} - f_{s+1} \leq f_s - \frac{1}{2}f_s - \frac{1}{2}\ell_s = \frac{2 - \lambda}{2}\Delta_s$$
Let $x_t, x_{t+1}$ be two subsequent search points of phase $s$. Then

$$
\|x_t - x_{t+1}\|_2 > \frac{(1 - \lambda) \Delta_s}{2L \| \cdot \|_2(f)}.
$$

Indeed, we have $f(x_t) = g_t(x_t) = \tilde{f}^{s,t}(x_t) \geq \ell_s + \frac{1}{2}(f^s - \ell_s)$, since otherwise phase $s$ would be terminates at step $t$. At the same time, $g_t(x_{t+1}) \leq \tilde{f}^{s,t}(x_{t+1}) \leq \ell_s$. Thus, passing from $x_t$ to $x_{t+1}$, we decrease Lipschitz continuous, with constant $L \| \cdot \|_2(f)$ w.r.t. $\| \cdot \|_2$, function $g_t(\cdot)$ by at least $\frac{1}{2}(f^s - \ell_s) = \frac{1-\lambda}{2} \Delta_s$. 
Main observation: Number of steps at phase \( s \) does not exceed

\[
N_s = \frac{4V^2_{\|\cdot\|_2,X(f)}}{(1 - \lambda)^2 \Delta_s^2} + 1. \tag{*}
\]

Indeed, let the number of steps of the phase be \( > N \). By construction, \( x_{t+1} \in H_{t-1} \) and \( x_t \) is the minimizer of \( \omega_s(x) = \frac{1}{2} \|x - c_s\|_2^2 \) on \( H_{t-1} \), whence

\[
1 \leq t \leq N \implies \omega_s(x_{t+1}) = \omega_s(x_t) + \underbrace{(x_{t+1} - x_t)^T \omega'_s(x_t)}_{\geq 0} + \frac{1}{2} \|x_t - x_{t+1}\|_2^2 \geq \omega_s(x_t) + \frac{1}{2} \|x_t - x_{t+1}\|_2^2.
\]

It follows that \( \sum_{t=1}^N \frac{1}{2} \|x_t - x_{t+1}\|_2^2 \leq \frac{1}{2} \max_{x,y \in X} \|y - x\|_2^2 \), whence \( N \leq \frac{4V^2_{\|\cdot\|_2,X(f)}}{(1 - \lambda)^2 \Delta_s^2} \).

Same as in the case of BL, (*) combines with the relation \( \Delta_{s+1} \leq \theta(\lambda) \Delta_s \) to yield the following

Corollary: For every \( \epsilon \), \( 0 < \epsilon < \Delta_1 \), the total number of TPL steps before a gap \( \leq \epsilon \) is obtained (i.e., before an \( \epsilon \)-solution is found) does not exceed the bound

\[
N(\epsilon) = c(\lambda) \frac{\text{Var}^2_{\|\cdot\|_2,X(f)}}{\epsilon^2}.
\]
\[ f_* = \min_{x \in X} f(x) \]  

\textit{From Gradient to Mirror Descent}

♣ Subgradient Descent method and its bundle versions are “intrinsically adjusted” to problems with Euclidean geometry; this is where the role of the \( \| \cdot \|_2 \)-variation of the objective

\[
\text{Var}_{\| \cdot \|_2, X}(f) = L_{\| \cdot \|_2}(f) \max_{x, x' \in X} \| x - x' \|_2
\]

in the efficiency estimate

\[
\min_{t \leq T} f(x_t) - f_* \leq O(1) \frac{\text{Var}_{\| \cdot \|_2, X}(f)}{\sqrt{T}}
\]

comes from.

♣ An extension of SD and its bundle versions onto problems with “nice non-Euclidean geometry” is offered by the \textit{Mirror Descent} scheme.
Mirror Descent – Building Blocks

♣ Building block #1: Distance-Generating Function.
♠ A SD step

\[ x \mapsto x_+ = \Pi_X(x - \gamma f'(x)) \]  

(1)

can be viewed as follows: given an iterate \( x \in X \), we

1) Compute \( f'(x) \)
2) Perform the prox-step \( x \mapsto x_+ = \text{Prox}_x(\gamma f'(x)) \)

\[
\text{Prox}_x(\xi) := \arg\min_{u \in X} \left[ \langle \xi - \omega'(x), u \rangle + \omega(u) \right] \\
= \arg\min_{u \in X} \left[ \langle \xi, u \rangle + V_x(u) \right], \\
V_x(u) = \omega(u) - \omega(x) - \langle \omega'(x), u - x \rangle
\]

where

\[
\omega(u) = \frac{1}{2} \|u\|_2^2
\]  

(2)

is a specific “distance-generating function.”

Indeed, with the above \( \omega(\cdot) \), we have

\[
V_x(u) := \frac{1}{2} u^T u - x^T (u - x) - \frac{1}{2} x^T x = \frac{1}{2} (u - x)^T (u - x) \Rightarrow \\
\text{Prox}_x(\xi) = \arg\min_{u \in X} \left[ \xi^T u + \frac{1}{2} (u - x)^T (u - x) \right] = \arg\min_{u \in X} \frac{1}{2} [u - (x - \xi)]^T [u - (x - \xi)] = \Pi_x(x - \xi)
\]
\[ \text{Prox}_x(\xi) = \underbrace{\text{argmin}_{u \in X} [\langle \xi - \omega'(x), u \rangle + \omega(u)]}_{\text{underlying all our convergence and rate-of-convergence results is an immediate corollary of the following “Magic Inequality:”}} \]

\[ V_x(u) = \omega(u) - \omega(x) - \langle \omega'(x), u - x \rangle \]

\[ x_+ = \Pi_X(x - \gamma f'(x)) \Rightarrow \forall u \in X : \gamma \langle f'(x), x - u \rangle \leq \frac{1}{2}||x - u||^2 - \frac{1}{2}||x_+ - u||^2 + \frac{1}{2}\gamma^2 ||f'(x)||^2 \]

\[ \text{Proof of (!):} \]

\[ x_+ = \text{argmin}_{u \in X} [\langle \xi - \omega'(x), u \rangle + \omega(u)] \Rightarrow \forall u \in X : \langle \xi - \omega'(x) + \omega'(x_+), u - x_+ \rangle \geq 0 \]

\[ \iff \forall u \in X : \langle \xi, x_+ - u \rangle \leq \langle \omega'(x_+) - \omega'(x), u - x_+ \rangle \]

\[ = \left[ \omega(u) - \omega(x) - \langle \omega'(x), u - x \rangle \right] \]

\[ - \left[ \omega(u) - \omega(x_+) - \langle \omega'(x_+), u - x_+ \rangle \right] \]

\[ - \left[ \omega(x_+) - \omega(x) - \langle \omega'(x), x_+ - x \rangle \right] \]

\[ = V_x(u) - V_{x_+}(u) - V_x(x_+) \]
**Magic Inequality** ⇒ **Main Inequality:** As we know, with \( \omega(u) = \frac{1}{2} \|u\|_2^2 \)
we have \( \Pi_X(x - \xi) = \text{Prox}_x(\xi) \). Thus,

\[
x_+ = \Pi_X(x - \gamma f'(x)) \Rightarrow x_+ = \text{Prox}_x(\gamma f'(x))
\]

\[
\Rightarrow \forall u \in X : \langle \gamma f'(x), x_+ - u \rangle \leq V_x(u) - V_{x_+}(u) - V_x(x_+)
\]

\[
\Rightarrow \forall u \in X : \langle \gamma f'(x), x - u \rangle \leq V_x(u) - V_{x_+}(u) + \left[ \langle \gamma f'(x), x - x_+ \rangle - V_x(x_+) \right]
\]

With our \( \omega(\cdot) \), \( V_x(x_+) = \frac{1}{2} \|x - x_+\|_2^2 \), whence

\[
\delta = \langle \gamma f'(x), x - x_+ \rangle - \frac{1}{2} \|x - x_+\|_2^2 \leq \frac{1}{2} \|\gamma f'(x)\|_2^2,
\]

and we arrive at the Main Inequality.
Now let

- \( \| \cdot \| \) be a norm on \( \mathbb{R}^n \)
- \( \omega(\cdot) \) be a \textit{distance-generating function (DGF)} for \( X \) compatible with \( \| \cdot \| \), meaning that
  - \( \omega(\cdot) : X \to \mathbb{R} \) is convex and continuously differentiable
  - \( \omega(\cdot) \) is strongly convex, modulus 1, w.r.t. \( \| \cdot \| \), meaning that
    \[
    \forall x, y \in X : \langle \omega'(x) - \omega'(y), x - y \rangle \geq \|y - x\|^2
    \]
or, equivalently,
    \[
    \forall (x \in X, u \in X) : V_x(u) := \omega(u) - \omega(x) - \langle \omega'(x), u - x \rangle \geq \frac{1}{2}\|u - x\|^2.
    \]

\textbf{Note:} For every convex compact set \( X \subset \mathbb{R}^n \), the function \( \omega(u) = \frac{1}{2}\|u\|^2 \) restricted to \( X \) is a DGF compatible with \( \| \cdot \| = \| \cdot \|_2 \)
∀(x ∈ X, u ∈ X) : V_x(u) := ω(u) − ω(x) − ⟨ω′(x), u − x⟩ ≥ \frac{1}{2}∥u − x∥^2.

Note: Whenever ω(·) is a DGF for X compatible with ∥ · ∥, for x ∈ X, ξ ∈ R^n, the prox-mapping

\[ x_+ = \text{Prox}_x(ξ) := \arg\min_{u ∈ X} [⟨ξ − ω′(x), u⟩ + ω(u)] \]

is well-defined, belongs to X, and

∀(u ∈ X) : ⟨ξ, x_+ − u⟩ ≤ V_x(u) − V_{x_+}(u) − V_x(x_+), \quad (1)

whence also

∀(u ∈ X) : ⟨ξ, x − u⟩ ≤ V_x(u) − V_{x_+}(u) + \frac{1}{2}∥ξ∥_*^2, \quad (2)

where ∥ · ∥_* is the norm conjugate to ∥ · ∥:

∥ξ∥_* = \max_x \{⟨ξ, x⟩ : ∥x∥ ≤ 1\} .
\[ V_x(u) = \omega(u) - \omega(x) - \langle u - x, \omega'(x) \rangle \geq \frac{1}{2} \|u - x\|^2 \]

\[ x_+ = \text{Prox}_x(\xi) := \arg\min_{u \in X} [\langle \xi - \omega'(x), u \rangle + \omega(u)] = \arg\min_{u \in X} [\langle \xi, u \rangle + V_x(u)] \]

**Claims:**

\[ \forall (u \in X): \langle \xi, x_+ - u \rangle \leq V_x(u) - V_{x_+}(u) - V_x(x_+) \quad (1) \]
\[ \forall (u \in X): \langle \xi, x - u \rangle \leq V_x(u) - V_{x_+}(u) + \frac{1}{2} \|\xi\|_*^2 \quad (2) \]

Indeed, as we have seen, (1) follows from optimality conditions as applied to the problem defining \( x_+ \). To derive (2) from (1), we need to show that

\[ \langle \xi, x - x_+ \rangle - V_x(x_+) \leq \frac{1}{2} \|\xi\|_*^2, \]

which is immediate due to

\[ \langle \xi, x - x_+ \rangle \leq \|\xi\|_* \|x - x_+\| \quad \& \quad V_x(x_+) \geq \frac{1}{2} \|x - x_+\|^2. \]
Conclusion A: Subgradient Descent step

\[ x \mapsto x_+ = \Pi_X(x - \gamma f'(x)) \]  

is the step

\[ x \mapsto x_+ = \arg\min_{y \in X} \left[ \langle \gamma f'(x) - \nabla \omega(x), y \rangle + \omega(y) \right] \]  

associated with the specific distance-generating function

\[ \omega(u) = \frac{1}{2} u^T u \]
\[ x \mapsto x_+ = \underset{y \in X}{\text{argmin}} \left( \langle \gamma f'(x) - \nabla \omega(x), y \rangle + \omega(y) \right) \]  

\[ \text{♣ Building block #2: the potential.} \]  
Convergence analysis of SD was based on the inequality

\[
\forall u \in X : \gamma \langle f'(x), x - u \rangle \leq \frac{1}{2} \| x - u \|_2^2 - \frac{1}{2} \| x_+ - u \|_2^2 + \frac{1}{2} \| \gamma f'(x) \|_2^2 \\
= \left[ \frac{1}{2} x^T x - x^T u \right] - \left[ \frac{1}{2} x_+^T x_+ - x_+^T u \right] \\
= \left[ \langle \nabla \omega(x), x - u \rangle - \omega(x) \right] - \left[ \langle \nabla \omega(x_+), x_+ - u \rangle - \omega(x_+) \right]
\]

where \( \omega(u) = \frac{1}{2} u^T u \), ensured by SD step. This inequality states that when \( \omega(u) = \frac{1}{2} u^T u \), a SD step \( x \mapsto x_+ \) reduces the “potential”

\[ V_x(u) = \omega(u) - \left[ \omega(x) + \langle \omega'(x), u - x \rangle \right] = \frac{1}{2} (u - x)^T (u - x) \]

by at least \( \gamma \langle f'(x), x - u \rangle - O(\gamma^2) \).

\[ \text{♠ We have seen that when } \omega(\cdot) \text{ is continuously differentiable and strongly convex, modulus } 1 \text{ w.r.t. } \| \cdot \|, \text{ on } X: \]

\[ \langle \nabla \omega(u) - \nabla \omega(v), u - v \rangle \geq \| u - v \|^2 \quad \forall u, v \in X \]

step (*) ensures inequality similar to (3):

\[
\gamma \langle f'(x), x - u \rangle \leq V_x(u) - V_{x_+}(u) + \frac{1}{2} \gamma^2 \| f'(x) \|_2^2 \\
\| \| \xi \|_* \| = \max_u \{ \langle \xi, u \rangle : \| u \| \leq 1 \}
\]
Non-Euclidean SD – Mirror Descent

\[
\min_{x \in \mathcal{X}} f(x) \quad \text{(P)}
\]

- \(\mathcal{X}\): compact set in Euclidean space \(\mathcal{E}\)
- \(f\): Lipschitz continuous convex function on \(\mathcal{X}\)

♣ Setup for MD ("Proximal Setup") is given by

— continuously differentiable strongly convex, modulus 1 w.r.t. \(\| \cdot \|\),

function \(\omega(u)\) on \(\mathcal{X}\): 
\[
\langle \nabla \omega(u) - \nabla \omega(v), u - v \rangle \geq \|u - v\|^2 \quad \forall u, v \in \mathcal{X}
\]

— norm \(\| \cdot \|\) on \(\mathcal{E}\)

♣ \(\omega(\cdot)\) and \(\| \cdot \|\) define the important parameter

- \(\Theta = \max_{u,v \in \mathcal{X}} [\omega(u) - \omega(v) - \langle \nabla \omega(v), u - v \rangle] \)

Note: With “Ball setup" \(\omega(u) = \frac{1}{2} \langle u, u \rangle, \|u\| \equiv \|u\|_2 = \sqrt{\langle u, u \rangle}\) one has

\(\Theta = \frac{1}{2} \max_{u,v \in \mathcal{X}} \|u - v\|^2_2\).

♣ As applied to (P), MD generates search points \(x_t\) according to

\[
x_{t+1} = \text{Prox}_{x_t}(\gamma_t f'(x_t)) := \arg\min_{y \in \mathcal{X}} \left[ \langle \gamma_t f'(x_t) - \nabla \omega(x_t), y \rangle + \omega(y) \right] \quad \text{(MD)}
\]

where \(\gamma_t > 0\) are stepsizes.
\[ x_{t+1} = \text{Prox}_{x_t}(\gamma_t f'(x_t)) := \arg\min_{y \in X} \left[ \langle \gamma_t f'(x_t) - \nabla \omega(x_t), y \rangle + \omega(y) \right] \] 

**Note:**

- With Ball setup, (MD) becomes exactly the SD recurrence \( x_{t+1} = \Pi_X(x_t - \gamma_t f'(x_t)) \)
- In order for (MD) to be practical, a step should be easy to implement. This means that \( X \) and \( \omega(\cdot) \) should fit each other in the sense that auxiliary problems

\[ \min_{y \in X} \left[ \langle \zeta, y \rangle + \omega(y) \right] \]

should be easy to solve.
Why and how MD converges?

\[
\{\min_{x \in X} f(x), \omega(\cdot)\} \Rightarrow x_{t+1} = \arg\min_{y \in X} [\langle \gamma_t f'(x_t) - \nabla \omega(x_t), y \rangle + \omega(y)]
\]

We have seen that MD step ensures inequality

\[
\forall u \in X : \gamma_t \langle f'(x_t), x_t - u \rangle \leq V_{x_0}(u) - V_{x_{t+1}}(u) + \frac{1}{2} \gamma_t^2 \| f'(x_t) \|_*^2
\]

\[
[V_x(u) = \omega(u) - \omega(x) - \langle \nabla \omega(x), u - x \rangle]
\]

It follows that for positive integers \( T_0 \leq T \) one has

\[
\sum_{t=0}^{T} \gamma_t \langle f'(x_t), x_t - u \rangle \leq V_{x_0}(u) - V_{x_{T+1}}(u) + \frac{1}{2} \sum_{t=0}^{T} \gamma_t^2 \| f'(x_t) \|_*^2 \leq \Theta + \frac{1}{2} \sum_{t=0}^{T} \gamma_t^2 \| f'(x_t) \|_*^2
\]

For MD, relation (1) plays the same crucial role that the inequality

\[
\sum_{t=0}^{T} \gamma_t \langle f'(x_t), x_t - u \rangle \leq \frac{1}{2} \max_{x, y \in X} \| x - y \|_2^2 + \frac{1}{2} \sum_{t=0}^{T} \gamma_t^2 \| f'(x_t) \|_2^2
\]

played for SD. Specifically, (1) implies that

\[
\epsilon_T \equiv \min_{t \leq T} f(x_t) - f_* \leq \Theta + \frac{1}{2} \sum_{t=0}^{T} \gamma_t^2 \| f'(x_t) \|_*^2 \leq \sum_{t=0}^{T} \gamma_t
\]
\[
\epsilon_T \equiv \min_{t \leq T} f(x_t) - f^* \leq \Theta + \frac{1}{2} \sum_{t=T_0}^{T} \gamma_t^2 \|f'(x_t)\|^2_* \sum_{t=T_0}^{T} \gamma_t
\]

As a result,

♣ [Convergence with “divergent series” stepsizes] Whenever \( 0 < \gamma_t \to 0 \) as \( t \to \infty \) in such a way that \( \sum_t \gamma_t = \infty \), one has \( \epsilon_T \to 0 \) as \( T \to \infty \)

♣ [Optimal stepsize policy] With stepsizes \( \gamma_t = \frac{\sqrt{\Theta}}{\|f'(x_t)\|^*_\sqrt{t}} \), one has

\[
\epsilon_T \equiv \min_{t \leq T} f(x_t) - f^* \leq O(1) \frac{\sqrt{\Theta} L_{\|\cdot\|}(f)}{\sqrt{T}}
\]

where \( L_{\|\cdot\|}(f) \) is the Lipschitz constant of \( f \) w.r.t. the norm \( \| \cdot \| \).
\[
\{ f_\ast = \min_{x \in X} f(x), \omega(\cdot) : X \to \mathbb{R}, \Theta = \max_{u, v \in X} \left[ \omega(u) - \omega(v) - \langle \nabla \omega(v), u - v \rangle \right] \}
\Rightarrow x_{t+1} = \arg\min_{y \in X} \left[ \langle \gamma_t f'(x_t) - \nabla \omega(x_t), y \rangle + \omega(y) \right], \gamma_t = \frac{\sqrt{\Theta}}{\|f'(x_t)\|, \sqrt{t}}
\Rightarrow \min_{t \leq T} f(x_t) - f_\ast \leq O(1) \frac{\sqrt{\Theta L\|\cdot\|_2}(f)}{\sqrt{T}}
\]

♠ To get the usual SD, one uses
♣ Ball setup \( \omega(u) = \frac{1}{2}\|u\|_2^2, \|\cdot\| = \|\cdot\|_2 \) \( X \subset \{ x : \|x\|_2 \leq R \} \Rightarrow \Theta \leq \frac{1}{2}R^2 \)

There are several other important setups:
♣ Simplex setup: \( X \subset \Delta_n = \{ x \in \mathbb{E} = \mathbb{R}^n : x \geq 0, \sum_i x_i \leq 1 \}, \|x\| = \|x\|_1, \omega(x) = O(1) \sum_i (x_i + n^{-1}\delta) \ln(x_i + n^{-1}\delta), \delta = 1.e-16 \)

[\( \Theta(X) \leq O(1) \ln(n + 1) \)]
♣ Spectahedron setup: \( X \subset \Xi_n = \{ x \in \mathbb{E} = S^n : x \geq 0, \text{Tr}(x) \leq 1 \}, \|x\| = |x|_1 \equiv \|\lambda(x)\|_1, \omega(x) = O(1) \sum_i (\lambda_i(x) + n^{-1}\delta) \ln(\lambda_i(x) + n^{-1}\delta) \)

[\( \Theta(X) \leq O(1) \ln(n + 1) \)]

For Simplex and Spectahedron setups, \( \Theta \leq O(\ln n) \), which results in nearly dimension-independent MD efficiency estimate

\[
\min_{t \leq T} f(x_t) - f_\ast \leq O(1) \frac{\sqrt{\ln(n) L\|\cdot\|_2(f)}}{\sqrt{T}}
\]

6.58
\[\ell_1 \text{ setup: } X \subset \mathbb{R}^{k_1} \times \mathbb{R}^{k_2} \times \ldots \times \mathbb{R}^{k_n}, \omega([x^1; \ldots; x^n]) = O(1) \left[ \sum_{i=1}^n \| x^i \|_2 \pi_n \right]^{2/\pi_n},\]
\[\pi_n = 1 + \frac{1}{n}, \| [x^1; \ldots; x^n] \| = \sum_i \| x^i \|_2\]
\[[X \subset \{ x : \| x \| \leq R \} \Rightarrow \Theta \leq O(1) \ln(n + 1) R^2] \]

\[\text{Nuclear norm setup: } X \subset \mathbb{R}^{p \times q}, \omega(x) = O(1) \left[ \sum_{i=1}^n \sigma_i \pi_n(x) \right]^{2/\pi_n},\]
\[n = \min[p, q], \pi_n = 1 + \frac{1}{n}, \sigma_i(x) : \text{singular values of } x, \| x \| = \| x \|_{\text{nuc}} := \sum_i \sigma_i(x)\]
\[[X \subset \{ x : \| x \| \leq R \} \Rightarrow \Theta \leq O(1) \ln(n + 1) R^2] \]
Justifying Simplex setup: It is easily seen that $\omega$ is strongly convex, modulus 1, w.r.t. $\|\cdot\|$, iff

$$\langle \nabla^2 \omega(x) h, h \rangle \geq \|h\|^2 \ \forall x \in X \forall h$$

For $x \in \Delta_n$ and $\bar{\omega}(x) = \sum_i (x_i + n^{-1}\delta) \ln(x_i + n^{-1}\delta)$, setting $\bar{x}_i = x_i + n^{-1}\delta$, one has

$$\|h\|_1^2 = \left[ \sum_i |h_i| \right]^2 = \left[ \sum_i (|h_i|/\sqrt{\bar{x}_i}) \sqrt{\bar{x}_i} \right]^2 \leq \left[ \sum_i h_i^2/\bar{x}_i \right] \left[ \sum_i \bar{x}_i \right] \leq (1 + \delta) \left( \sum_i h_i^2/\bar{x}_i \right) = (1 + \delta) \langle h, \nabla^2 \bar{\omega}(x) h \rangle,$$

whence $\omega(x) := (1 + \delta)\bar{\omega}(x)$ is strongly convex, modulus 1 w.r.t. $\|\cdot\|_1$, on $\Delta_n$.

Next, for $x, y \in \Delta_n$, setting $\bar{y}_i = y_i + \delta n^{-1}$, $\bar{x}_i = x_i + \delta n^{-1}$, we have

$$\omega(y) - \omega(x) - \langle \nabla \omega(x), y - x \rangle = (1 + \delta) \left[ \sum_i \bar{y}_i \ln \bar{y}_i - \sum_i \bar{x}_i \ln \bar{x}_i - \sum_i (1 + \ln \bar{x}_i)(\bar{y}_i - \bar{x}_i) \right]
= (1 + \delta) \left[ \sum_i \bar{y}_i \ln(\bar{y}_i/\bar{x}_i) + \sum_i [\bar{x}_i - \bar{y}_i] \right]
\leq (1 + \delta) \left[ \sum_i \bar{y}_i \ln(n/\delta) + 1 \right] \leq O(1) \ln n.$$
$$f_\ast = \min_{x \in X} f(x) \quad (P)$$

Let us compare the convergence properties of MD with Simplex setup and SD (i.e., MD with Ball setup).

- Observe that in order to apply MD with Simplex setup, $X$ should be a subset of the standard simplex. We can ensure this requirement by scaling and translating the original feasible domain. As a result, MD with Simplex setup becomes applicable to an arbitrary convex problem $(P)$ with compact feasible domain $X$, and the efficiency estimate for the method becomes

$$
\epsilon_T[\text{Simplex setup}] = \min_{t \leq T} f(x_t) - f_\ast \leq O(1) \ln^{1/2}(n) \max_{x, y \in X} \|x - y\|_1 L_{\|\cdot\|_1}(f) / \sqrt{T} \quad (S)
$$

while for SD the efficiency estimate is

$$
\epsilon_T[\text{Ball setup}] = \min_{t \leq T} f(x_t) - f_\ast \leq O(1) \max_{x, y \in X} \|x - y\|_2 L_{\|\cdot\|_2}(f) / \sqrt{T} \quad (B)
$$

The ratio of the estimates is

$$
\frac{\epsilon_T[\text{Simplex setup}]}{\epsilon_T[\text{Ball setup}]} = O(\sqrt{\ln n}) \cdot \left[ \frac{\max_{x, y \in X} \|x - y\|_1}{A} \right] \cdot \left[ \frac{L_{\|\cdot\|_1}(f)}{B \ L_{\|\cdot\|_2}(f)} \right]
$$

6.61
\[ \epsilon_T[\text{Simplex setup}] = O(\sqrt{\ln n}) \cdot \left[ \max_{x,y \in X} \|x - y\|_1 \right] \cdot \left[ \frac{L_{\|\cdot\|_1}(f)}{L_{\|\cdot\|_2}(f)} \right] \]

The factor \( O(\sqrt{\ln n}) \) is “against” Simplex setup; however, in practice this factor is just a moderate absolute constant.

Note that \( \frac{\|u\|_1}{\|u\|_2} \) is always \( \geq 1 \) and, depending on \( x \), can be as large as \( \sqrt{n} \). It follows that

— factor \( A \) is always \( \geq 1 \) (i.e., is “against” Simplex setup) and can be as large as \( \sqrt{n} \)
— factor \( B \) is always \( \leq 1 \) (i.e., is “in favour” of Simplex setup) and can be as small as \( \frac{1}{\sqrt{n}} \). The actual value of \( B \) is

\[ \frac{L_{\|\cdot\|_1}(f)}{L_{\|\cdot\|_2}(f)} = \max_{x \in X} \|f'(x)\|_\infty \max_{x \in X} \|f'(x)\|_2 \]

and depends on the “geometry” of \( f \). For example,

— when all first order partial derivatives of \( f \) in \( X \) are of the same order (“\( f \) is nearly equally sensitive to all variables”), we have

\[ B = O\left( \frac{\|(a,...,a)^T\|_\infty}{\|(a,...,a)^T\|_2} \right) = O(n^{-1/2}) \]

— when just \( O(1) \) first order derivatives of \( f \) on \( X \) are of the same order, and the remaining derivatives are negligible small (“\( f \) is sensitive to just \( O(1) \) variables”), we have

\[ B = O\left( \frac{\|(a,0,...,0)^T\|_\infty}{\|(a,0,...,0)^T\|_2} \right) = O(1) \]

\[ \bullet \text{ Conclusion: } \text{The performance ratio } \chi \text{ depends on the geometry of } X \text{ and } f. \]
\[ \chi = \frac{\epsilon_T[\text{Simplex setup}]}{\epsilon_T[\text{Ball setup}]} = O(\sqrt{\ln n}) \cdot \frac{\max_{x,y \in X} \|x - y\|_1}{\max_{x,y \in X} \|x - y\|_2} \cdot \left[ \frac{L\|\cdot\|_1(f)}{L\|\cdot\|_2(f)} \right] \]

\[ 1 \leq A \leq \sqrt{n} \quad 1 \geq B \geq \frac{1}{\sqrt{n}} \]

**Extreme example I:** \( X \) is a ball. In this case, \( A = \sqrt{n} \), and since \( B \geq \frac{1}{\sqrt{n}} \), \( \chi \geq 1 \) – method with Ball setup (i.e., the classical SD) outperforms the method with Simplex setup by factor which varies from \( O(\sqrt{\ln n}) \) (\( f \) is nearly equally sensitive to all variables) to \( O(\sqrt{n \ln n}) \) (\( f \) is sensitive to just \( O(1) \) variables).

**Extreme example II:** \( X \) is the unit simplex \( \Delta_n \). In this case, \( A = O(1) \), and since \( B \leq 1 \) and \( O(\sqrt{\ln n}) \) in practice a moderate absolute constant, \( \chi \leq O(1) \) – method with Simplex setup outperforms the classical SD by factor which varies from \( O\left(\frac{n}{\ln n}\right) \) (\( f \) is nearly equally sensitive to all variables) to \( O\left(\frac{1}{\ln n}\right) \) (\( f \) is sensitive to just \( O(1) \) variables).

**Conclusion:** Flexibility in setup allows to adjust MD, to some extent, to the geometry of the problem to be solved. Let all flowers blossom!
The Maximum Likelihood estimate of tracer’s density in PET is

\[
\lambda_* = \arg\min_{\lambda \geq 0} \left\{ \sum_{j=1}^{n} p_j \lambda_j - \sum_{i=1}^{m} y_i \ln(\sum_{j=1}^{n} p_{ij} \lambda_j) \right\}
\]

\[y_i \geq 0\text{ are observations, } p_{ij} \geq 0, p_j = \sum_i p_{ij}\]

The KKT optimality conditions read

\[
\lambda_j \left( p_j - \sum_i y_i \frac{p_{ij}}{\sum_{\ell} p_{i\ell} \lambda_{\ell}} \right) = 0 \quad \forall j,
\]

whence, taking sum over \( j \),

\[
\sum_j p_j \lambda_j = B \equiv \sum_i y_i.
\]

Thus, in fact (PET) is the problem of minimizing over a simplex. Passing to the variables \( x_j = p_j B^{-1} \lambda_j \), we end up with the problem

\[
\min_x \left\{ f(x) = -\sum_i y_i \ln(\sum_j q_{ij} x_j) : x \in \Delta_n \right\}
\]

\[q_{ij} = B p_{ij} p_j^{-1}\]

(PET)
Illustration: “Hot Spheres” phantom \( (n = 515,871) \)

<table>
<thead>
<tr>
<th>Itr</th>
<th>( f(x_i) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-4.295</td>
</tr>
<tr>
<td>2</td>
<td>-4.767</td>
</tr>
<tr>
<td>3</td>
<td>-5.079</td>
</tr>
<tr>
<td>4</td>
<td>-5.189</td>
</tr>
<tr>
<td>5</td>
<td>-5.168</td>
</tr>
<tr>
<td>6</td>
<td>-5.230</td>
</tr>
<tr>
<td>7</td>
<td>-5.181</td>
</tr>
<tr>
<td>8</td>
<td>-5.227</td>
</tr>
<tr>
<td>9</td>
<td>-5.189</td>
</tr>
<tr>
<td>10</td>
<td>-5.225</td>
</tr>
</tbody>
</table>

\( f^* \geq -5.283 \)

Simplex setup. Progress in accuracy in 10 iterations by factor 21.4
Simplex setup (left) vs. Ball setup (right) progress in accuracy 21.4 vs. 5.26
Illustration: Brain clinical data \((n = 2,763,635)\)

<table>
<thead>
<tr>
<th>Itr</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
</table>

\([f_* \geq -2.050]\)

Simplex setup. Progress in accuracy in 10 iterations by factor 17.5
Mirror-Level Algorithm

- Same as SD, the general Mirror Descent admits a version with memory – Mirror Level (ML) algorithm. The setup for ML is similar to the one of MD and is given by a strongly convex $C^1$ function $\omega(\cdot)$ on $X$ and a norm $\| \cdot \|$ on $E$.

- At step $t$ of ML, we
  - compute $f(x_t), f'(x_t)$ and build the current model of $f$
    $$f_t(x) = \max_{\tau \leq t} [f(x_\tau) + \langle f'(x_\tau), x - x_\tau \rangle]$$
  which underestimates the objective and is exact at the points $x_1, ..., x_t$;
  - define the best found so far value of the objective $f^t = \min_{\tau \leq t} f(x_\tau)$
  - define the current lower bound $f_t$ on $f^*$ by solving the auxiliary problem
    $$f_t = \min_{x \in X} f_t(x)$$

The current gap $\Delta_t = f^t - f_t$ is an upper bound on the inaccuracy of the best found so far approximate solution;

- compute the current level $\ell_t = f_t + \lambda \Delta_t$ ($\lambda \in (0, 1)$ is a parameter)
- finally, we set
  $$L_t = \{ x \in X : f^t(x) \leq \ell_t \},$$
  $$x_{t+1} = \text{Prox}_{L_t}^x(0) := \arg\min_{x \in L_t} \left[ \langle -\nabla \omega(x_t), x \rangle + \omega(x) \right]$$

and loop to step $t + 1$. 

6.68
♠ With Ball setup,

$$\text{Prox}^{L_t}_{x_t}(0) = \arg\min_{x \in L_t} \left[ -x^T x + \frac{1}{2} x^T x_t \right] = \arg\min_{x \in L_t} \frac{1}{2} \| x - x_t \|^2_2.$$ 

i.e., the method becomes exactly the BL algorithm.
Why and how ML converges?

odynamo Convergence analysis of BL was based on the following fact:

Let $J = \{s, s + 1, ..., r\}$ be a segment of iterations of BL:

$$\Delta_r \geq (1 - \lambda) \Delta_s.$$  

Then the cardinality of $J$ can be upper-bounded as

$$\text{Card}(J) \leq \frac{\left(\max_{x,y \in X} \|x - y\|_2 L_{\|\|}(f)\right)^2}{(1 - \lambda)^2 \Delta^2}.$$  

Similar fact for ML reads:

(1) Let $J = \{s, s + 1, ..., r\}$ be a segment of iterations of ML: $\Delta_r \geq (1 - \lambda) \Delta_s.$

Then the cardinality of $J$ can be upper-bounded as

$$\text{Card}(J) \leq \frac{2 \Theta L^2_{\|\|}(f)}{(1 - \lambda)^2 \Delta^2}.$$  

From (1), exactly as in the case of BL, one derives

**Corollary:** For every $\epsilon$, $0 < \epsilon < \Delta_1$, the number $N$ of steps of ML before a gap $\leq \epsilon$ is obtained (i.e., before an $\epsilon$-solution is found) does not exceed the bound

$$N(\epsilon) = \frac{4 \Theta L^2_{\|\|}(f)}{\lambda (1 - \lambda)^2 (2 - \lambda) \epsilon^2}.$$  

In particular, for Simplex/Spectahedron setup one has

$$N(\epsilon) = O(\ln n) \frac{\left(\max_{x,y \in X} \|x - y\| L_{\|\|}(f)\right)^2}{\lambda (1 - \lambda)^2 (2 - \lambda) \epsilon^2}.$$
(!) Let $J = \{s, s + 1, ..., r\}$ be a segment of iterations of ML: $\Delta_r \geq (1 - \lambda)\Delta_s$.

Then the cardinality of $J$ can be upper-bounded as $\text{Card}(J) \leq \frac{2\Theta L^2(f)}{(1 - \lambda)^2 \Delta_r^2}$.

**Proof.** Same as in the case of BL, we observe that

- For $t$ running through a segment of iterations $J$, the level sets $L_t = \{x \in X : f_t(x) \leq \ell_t\}$ have a point in common, namely, $v \in \text{Argmin}_{x \in X} f_r(x)$;

- When $t \in J$, the distances $\gamma_t = \|x_t - x_{t+1}\|$ are not too small: $\gamma_t \geq \frac{(1 - \lambda)\Delta_r}{L\|\|f)}$.

- As we shall see in a while,

\[
V_{x_{t+1}}(v) \leq V_{x_t}(v) - \frac{1}{2} \gamma_t^2, \quad t \in J
\]

\[
[V_x(y) = \omega(y) - [\langle y - \nabla \omega(x), y - x \rangle + \omega(x)] \geq \frac{1}{2}\|y - x\|^2]
\]

(\#)

Thus, while $t$ stays within $J$, $V_{x_t}(v)$ decrease from step to step by at least $\frac{1}{2} \gamma_t^2$.

Since $0 \leq V_x(y) \leq \Theta$ for all $x, y \in X$, (\#) combines with the lower bound on $\gamma_t, t \in J$, to imply the desired upper bound on the cardinality of $J$.
Proof of (¬). Magic Inequality says that whenever \( x \in X \), \( \xi \in E \) and 
\[
x_+ = \arg \min_{y \in X} [\langle \xi - \nabla \omega(x), y \rangle + \omega(y)],
\]
it holds 
\[
\langle \xi, x_+ - u \rangle \leq V_x(u) - V_{x_+}(u) - V_x(x_+),
\]
This fact admits modification as follows:

($) Let \( Y \subset X \) be nonempty convex compact sets in Euclidean space \( E \) and \( \omega(\cdot) \) be a
DGF for \( X \) compatible with a norm \( \| \cdot \| \) on \( E \). Given \( x \in X \) and \( \xi \in E \), let
\[
x_+ = \arg \min_{y \in Y} [\langle \xi - \nabla \omega(x), y \rangle + \omega(y)]
\]
Then
\[
\forall u \in Y : \langle \xi, x_+ - u \rangle \leq V_x(u) - V_{x_+}(u) - V_x(x_+).
\]
Applying ($) to \( \xi = 0 \), \( x = x_t \), \( Y = L_t \) and \( u = v \), we get (¬).

Proof of Modification repeats the proof of plain Magic Inequality:

\[
x_+ = \arg \min_{y \in Y} [\langle \xi - \omega'(x), y \rangle + \omega(y)] \Rightarrow \forall u \in Y : \langle \xi - \omega'(x) + \omega'(x_+) + u - x_+ \rangle \geq 0
\]
[optimality conditions]
NERML – Non-Euclidean Restricted Memory Level algorithm

\[ \min_{x \in X} f(x) \]

- NERML is a version of ML where bundle size is kept below a given desired level \( m \).
- The setup for NERML, same as those for MD and ML, is given by a continuously differentiable strongly convex on \( X \) function \( \omega(\cdot) \) and a norm \( \| \cdot \| \) on the Euclidean space \( E \) where \( X \) lives.
- Execution of NERML is split into phases. Phase \( s \) is associated with
  - prox-center \( c_s \in X \)
  - \( s \)-th upper bound \( f^s \) on \( f^* \), which is the best value of the objective observed before the phase begins
  - \( s \)-th lower bound \( f_s \) on \( f^* \), which is the best lower bound on \( f^* \) observed before the phase begins
- \( f^s \) and \( f_s \) define \( s \)-th optimality gap \( \Delta_s = f^s - f_s \)
- \( s \)-th level \( \ell_s = f_s + \lambda \Delta_s \), where \( \lambda \in (0, 1) \) is parameter of the method,
- \( s \)-th local distance
  \[ \omega_s(x) = \omega(x) - \langle \nabla \omega(c_s), x \rangle - \omega(c_s) \]
- current model \( \tilde{f}^s(\cdot) \leq f(\cdot) \) of \( f(\cdot) \), which is the maximum of \( \leq m \) affine forms.
To initialize the first phase, we choose $c_1 \in X$, compute $f(c_1), f'(c_1)$ and set
\[\tilde{f}^1(x) = f(c_1) + \langle f'(c_1), x - c_1 \rangle, \quad f^1 = f(c_1), \quad f_1 = \min_{x \in X} \tilde{f}^1(x).\]

At the beginning of step $t = 1, 2, \ldots$ of phase $s$, we have at our disposal
— upper bound $f^{s,t-1} \leq f^s$ on $f^*$, which is the best found so far value of the objective,
— lower bound $f_{s,t-1} \geq f_s$ on $f^*$,
— model $\tilde{f}^{s,t-1}(\cdot) \leq f(\cdot)$ of the objective which is the maximum of $\leq m$ affine forms
— iterate $x_t \in X$ and set
\[H_{t-1} = \{x : \langle \alpha_{t-1}, x \rangle \geq \beta_{t-1}\}\]
such that
\[x \in X, f(x) \leq \ell_s \Rightarrow x \in H_{t-1} \quad (a_t)\]
\[x_t = \arg\min_x \{\omega_s(x) : x \in H_{t-1} \cap X\} \quad (b_t)\]

To initialize the first step of phase $s$, we set
\[f^{s,0} = f^s, \quad f_{s,0} = f_s, \quad \tilde{f}^{s,0}(\cdot) = \tilde{f}^s(\cdot), \quad \alpha_0 = 0, \beta_0 = 0 \quad [\Rightarrow H_0 = E]\]
thus ensuring $(a_1)$, and set $x_1 = c_s$, thus ensuring $(b_1)$. 
Step \( t \) phase \( s \): Given

- bounds \( f_{s,t-1}^{s,t-1} \geq f_\star, f_{s,t-1} \leq f_\star \)
- model \( \tilde{f}_{s,t-1}^{s,t-1}(\cdot) \leq f(\cdot) \)
- \( x_t \) and \( H_{t-1} = \{ x : \langle \alpha_{t-1}, x \rangle \geq \beta_{t-1} \} \) such that

\[
x \in X, f(x) \leq \ell_s \Rightarrow x \in H_{t-1} \quad (a_t) \quad \text{and} \quad x_t = \text{argmin}_x \{ \omega_s(x) : x \in H_{t-1} \cap X \} \quad (b_t)
\]

1. we compute \( f(x_t), f'(x_t) \) and set

\[
  g_t(x) = f(x_t) + \langle f'(x_t), x - x_t \rangle;
\]

2. we define \( \tilde{f}_{s,t}(\cdot) \) as the maximum of \( g_t(\cdot) \) and affine forms associated with \( \tilde{f}_{s,t-1} \) (dropping, if necessary, one of the latter forms to make \( \tilde{f}_{s,t} \) the maximum of at most \( m \) forms). If \( f(x_t) \leq \ell_s + 0.5(f^s - \ell_s) \) ("progress in upper bound"), we terminate phase \( s \) and set

\[
  f_{s+1} = f_{s,t}, \quad f_{s+1} = f_{s,t-1}, \quad \tilde{f}_{s+1}(\cdot) = \tilde{f}_{s,t}(\cdot),
\]

otherwise

3. we compute \( f_t = \text{min}_x \{ \tilde{f}_{s,t}(x) : x \in H_{t-1} \cap X \} \). Since \( f(x) \geq \ell_s \) in \( X \setminus H_{t-1} \), we have \( f_\star \geq \text{min}[\ell_s, f_t] \), so that

\[
  f_{s,t} = \max \{ f_{s,t-1}, \text{min}[\ell_s, f_t] \} \leq f_\star.
\]

If \( f_{s,t} \geq \ell_s - 0.5(\ell_s - f_s) \) ("progress in lower bound"), we terminate phase \( s \) and set

\[
  f_{s+1} = f_{s,t}, \quad f_{s+1} = f_{s,t}, \quad \tilde{f}_{s+1}(\cdot) = \tilde{f}_{s,t}(\cdot)
\]

otherwise we set

\[
  x_{t+1} = \text{argmin}_x \{ \omega_s(x) : x \in X \cap H_{t-1}, \tilde{f}_{s,t}(x) \leq \ell_s \}
\]

\[
  H_t = \{ x : \langle \nabla \omega_s(x_{t+1}), x - x_{t+1} \rangle \geq 0 \}
\]

and loop to step \( t + 1 \) of phase \( s \).
\[ x_{t+1} = \operatorname{argmin}_x \left\{ \omega_s(x) : x \in X \cap H_{t-1}, \tilde{f}^{s,t}(x) \leq \ell_s \right\} \]  \hspace{1cm} (1)
\[ H_t = \{ x : \langle \nabla \omega_s(x_{t+1}), x - x_{t+1} \rangle \geq 0 \} \]  \hspace{1cm} (2)

**Note:** When passing to step \( t + 1 \), it is ensured that
\[ x \in X, f(x) \leq \ell_s \Rightarrow x \in H_t \]  \hspace{1cm} (a_{t+1})
\[ x_{t+1} = \operatorname{argmin}_x \left\{ \omega_s(x) : x \in X \cap H_t, \tilde{f}^{s,t}(x) \leq \ell \right\} \]  \hspace{1cm} (b_{t+1})

Indeed, \( x_{t+1} \) is the minimizer of \( \omega_s(x) \) on the set
\[ Y_t = X \cap H_{t-1} \cap \{ x : \tilde{f}^{t,s}(x) \leq \ell_s \} \]
whence
\[ \langle \nabla \omega_s(x), x - x_{t+1} \rangle \geq 0 \ \forall x \in Y_t \]
\[ \Rightarrow Y_t \subset H_t = \{ x : \langle \nabla \omega_s(x_{t+1}), x - x_{t+1} \rangle \geq 0 \} \]  \hspace{1cm} (*)

Thus,
\[ (x \in X, f(x) \leq \ell_s) \Rightarrow (x \in X \cap H_{t-1}, f(x) \leq \ell_s) \Rightarrow (x \in X \cap H_{t-1}, \tilde{f}^{s,t}(x) \leq \ell_s) \Rightarrow x \in H_t \]  \hspace{1cm} (*)
as required in \((a_{t+1})\). \((b_{t+1})\) readily follows from the definition of \( H_t \). \qed
Convergence of NERML

♣ The efficiency estimate for TLM was a nearly straightforward consequence of the following fact:

(*)  The number of steps of TLM at a phase $s$ does not exceed

$$N_s = \frac{4 \left( \max_{x,y} \|x - y\|_2 L_{\|\cdot\|_2}^2 (f) \right)^2}{(1 - \lambda)^2 \Delta_s^2} + 1.$$  

♠ For NERML, a similar fact is valid:

(!)  The number of steps of NERML at a phase $s$ does not exceed

$$N_s = \frac{8 \Theta L_{\|\cdot\|_2}^2 (f)}{(1 - \lambda)^2 \Delta_s^2} + 1.$$  

♣ The same reasoning as in the case of TLM, with (!) playing the role of (*), yields

Corollary: For every $\epsilon$, $0 < \epsilon < \Delta_1$, the total number of NERML steps before a gap $\leq \epsilon$ is obtained (i.e., before an $\epsilon$-solution is found) does not exceed the bound

$$N(\epsilon) = c(\lambda) \Theta L_{\|\cdot\|_2}^2 (f) \epsilon^{-2}.$$  

6.77
Claim:
(!) The number of steps of NERML at a phase \( s \) does not exceed \( N_s = \frac{8\Theta L^2(f)}{(1-\lambda)^2 \Delta_s^2} + 1 \).

Proof. Let phase \( s \) not be terminated in course of \( N \) steps. By construction, for \( 1 \leq t \leq N \) we have

\[
x_{t+1} \in H_{t-1} \cap X \& x_t = \arg\min_x \{\omega_s(x) : x \in H_{t-1} \cap X\}
\]

\[
\Rightarrow \omega_s(x_{t+1}) \geq \omega_s(x_t) + \langle \nabla \omega(x_t), x_{t+1} - x_t \rangle + \frac{1}{2} \|x_{t+1} - x_t\|^2 \geq \omega_s(x_t) + \frac{1}{2} \|x_{t+1} - x_t\|^2 \quad (1)
\]

Further, when passing from \( x_t \) to \( x_{t+1} = \arg\min_x \{\omega_s(x) : x \in H_{t-1} \cap X, \tilde{f}^{s,t}(x) \leq \ell_s\} \), the function \( g_t(x) \equiv f(x_t) + \langle f'(x_t), x - x_t \rangle \leq \tilde{f}^{s,t} \) varies from the value \( f(x_t) \geq f^{s,t} \) to a value \( \leq \ell_s \) and thus decreases by at least \( 0.5(1-\lambda)\Delta_s \) (otherwise phase \( s \) would be terminated at step \( t \) due to progress in upper bound). Since \( g_t(\cdot) \) is Lipschitz continuous, with constant \( L_{\|\cdot\|}(f) \) w.r.t. \( \|\cdot\| \), we conclude that

\[
0.5(1-\lambda)\Delta_s \leq \|x_t - x_{t+1}\| L_{\|\cdot\|}(f) \Rightarrow \|x_t - x_{t+1}\| \geq \frac{0.5(1-\lambda)\Delta_s}{L_{\|\cdot\|}(f)}.
\]

Applying (1), we arrive at

\[
\omega_s(x_{t+1}) \geq \omega_s(x_t) + \frac{(1-\lambda)^2}{8L^2_{\|\cdot\|}(f)} \Delta_s^2, \quad 1 \leq t \leq N. \quad (2)
\]

Since the function \( \omega_s(x) = \omega(x) - \langle \nabla \omega(c_s), x - c_S \rangle + \omega(c_s) \) maps \( X \) into \([0, \Theta]\), (2) implies (!). \( \square \)
Implementation issues

How to solve auxiliary problems? At a step of NERML, one should solve the auxiliary problems

\[
\begin{align*}
    f_t &= \min_x \{ \tilde{f}^{s,t}(x) : x \in H_{t-1} \cap X \} \quad (L) \\
    x_{t+1} &= \arg\min_x \{ \omega_s(x) : x \in H_{t-1} \cap X, \tilde{f}^{s,t}(x) \leq \ell_s \} \quad (N)
\end{align*}
\]

Formally, both (L) and (N) are problems of the same dimension as the problem of interest.

**Question:** Does it make sense to reduce the large-scale problem of interest to a series of equally large-scale auxiliary problems?

**Answer:** Yes, it does – (L), (N) can be easily reduced to a low-dimensional black-box-represented convex programs.
\[
\min_x \left\{ \tilde{f}^{s,t}(x) : x \in H_{t-1} \cap X \right\} \quad (L)
\]

Assume that \(X\) is a simple polytope. Then \((L)\) is an LP program and can be solved as such, unless the dimension of \(X\) is really large. In the latter case, we can solve \((L)\) via Lagrange Duality. Indeed, the objective in \((L)\) is the maximum of (at most) \(m\) affine functions \(h_i(x), i = 1, \ldots, m\), while \(H_{t-1}\) is given by a single linear inequality \(h_0(x) \leq 0\). Thus, \((L)\) is the problem

\[
f_t = \min_{x \in X} \left\{ \max_{1 \leq i \leq m} h_i(x) : h_0(x) \leq 0 \right\} = \max_\lambda \left\{ F(\lambda) = \min_{x \in X} [\sum_{i=0}^{m} \lambda_i h_i(x)] : \lambda \geq 0, \sum_{i=1}^{m} \lambda_i = 1 \right\}.
\]

In order to compute \(F(\lambda)\) and \(F'(\lambda)\) at a given \(\lambda\), it suffices to minimize over \(X\) the linear function \(h_\lambda(x) = \sum_{i=0}^{m} \lambda_i h_i(x)\). after a minimizer \(x_\lambda\) of \(h_\lambda(\cdot)\) over \(X\) is found, one sets

\[
F(\lambda) = h_\lambda(x_\lambda); \quad F'(\lambda) = (h_1(x_\lambda), \ldots, h_m(x_\lambda))^T. \quad (\ast)
\]

Assuming problems \(\min_{x \in X} [\langle \xi, x \rangle + \omega(x)]\) easily solvable, problem of minimizing linear objective over \(X\) is easily solvable as well. \(\Rightarrow\) it is easy to implement the First Order oracle for \(F\). Thus, we can find \(f_t\) by solving the black-box-represented convex program

\[
\max_\lambda \left\{ F(\lambda) : \lambda \geq 0, \sum_{i=1}^{m} \lambda_i = 1 \right\}
\]

with dimension \(m + 1\) (which is under our full control!) by, say, the Ellipsoid method.

6.80
The second auxiliary problem

\[
x_{t+1} = \arg\min_{x} \left\{ \omega_{s}(x) : x \in X \cap H_{t-1}, \tilde{f}_{s,t} \leq \ell_{s} \right\}
\]

\[
= \arg\min_{x \in X} \left\{ \omega(x) + \langle \xi_{s}, x \rangle : \tilde{h}_{i}(x) \leq 0, i = 1, \ldots, m + 1 \right\}
\]

also can be reduced to \( m + 1 \)-dimensional black-box-represented convex program

\[
\max_{\lambda \geq 0} \Phi(\lambda), \quad \Phi(\lambda) = \min_{x \in X} \left[ \omega(x) + \langle \xi_{s}, x \rangle + \sum_{i=1}^{m+1} \lambda_{i} h_{i}(x) \right]
\]

with First Order oracle readily given by the possibility to solve auxiliary problems

\[
x_{\lambda} = \arg\min_{x \in X} \left[ \omega(x) + \langle \xi_{s}, x \rangle + \sum_{i=1}^{m+1} \lambda_{i} h_{i}(x) \right].
\]

After \( \lambda_{*} \in \text{Argmin}_{\lambda \geq 0} \Phi(\lambda) \) is found by, e.g., the Ellipsoid method, we recover \( x_{t+1} \) as \( x_{\lambda_{*}} \).

**Note:** \( \omega(\cdot) \) is strongly convex, so that high-accuracy approximate solution to \( \max_{\lambda \geq 0} \Phi(\lambda) \) results in high accuracy approximation to \( x_{t+1} \).

\( \Rightarrow \) With the outlined approach MD/ML/NERML become implementable under the only assumption that one can easily solve problems \( \min_{X} [\langle \xi, x \rangle + \omega(x)] \). This indeed is so for

- Ball setup and simple \( X \) (ball, box, positive part of ball, standard simplex,...),
- Simplex setup and simple \( X \) (the entire simplex \( \Delta_{n} \), intersection of \( \Delta_{n} \) and a box,...),
- Spectahedron setup with \( X \) comprised of block-diagonal matrices with diagonal blocks of size \( O(1) \).

In all these cases, \( (*) \) can be solved in \( \leq O(n \ln n) \) a.o.
\[
\min_x \left\{ f(x) = -\sum_{i=1}^m y_i \ln \left( \sum_{j=1}^n q_{ij} x_j \right) : x \in \Delta_n \right\}
\]

We have simulated 2D PET scanner with a single ring of detectors:

- Ring with 360 detectors, field of view and a LOR (ring's radius 1, field of view's radius 0.9)
- and field of view partitioned into pixels by 128 × 128 regular grid. With this setup,
  — the design dimension of the problem is \( n = 10,471 \);
  — the number of log-terms in the objective is 39,784
  — the number of nonzero \( q_{ij} \) is 3,746,832 (the density of the matrix \([q_{ij}]\) is 0.009).

The algorithm: plain NERML with Simplex setup, \( m = 1 \) and \( \lambda = 0.95 \).
Experiment 1: noiseless measurements (brighter pixels correspond to higher tracer’s density):

True image: 10 “hot spots”
\[ f = f_\ast = 2.817 \]

\[ x^1 = n^{-1}(1,\ldots,1)^T \]
\[ f = 3.247 \]

\[ x^2 \text{ – some traces of 8 spots} \]
\[ f = 3.185 \]

\[ x^3 \text{ – traces of 8 spots} \]
\[ f = 3.126 \]

\[ x^5 \text{ – some trace of 9-th spot} \]
\[ f = 3.016 \]

\[ x^8 \text{ – 10-th spot still missing…} \]
\[ f = 2.869 \]

\[ x^{24} \text{ – trace of 10-th spot} \]
\[ f = 2.828 \]

\[ x^{27} \text{ – all 10 spots in place} \]
\[ f = 2.823 \]

\[ x^{31} \text{ – that is it…} \]
\[ f = 2.818 \]
Progress in accuracy, noiseless measurements.

solid line: Relative gap $\frac{\text{Gap}(t)}{\text{Gap}(1)}$ vs. step number $t$; $\text{Gap}(t)$ is the difference between the best found so far value $f(x^t)$ of $f$ and the current lower bound on $f_*$.  
• In 111 steps, the gap was reduced by factor $> 1600$

dashed line: Progress in accuracy $\frac{f(x^t)-f_*}{f(x^1)-f_*}$ vs. step number $t$  
• In 111 steps, the accuracy was improved by factor $> 1080$

• 111 steps of the NERML algorithm took 18'51” on a 350 MHz Pentium II laptop with 96 MB RAM.
Experiment 2: noisy measurements (at average, 40 LOR's per bright pixel, 63,092 LOR's totally):

True image: 10 “hot spots”
\[ f = -0.883 \]

\[ x^1 = n^{-1}(1, \ldots, 1)^T \]
\[ f = -0.452 \]

\( x^2 \) – light traces of 5 spots
\[ f = -0.520 \]

\( x^3 \) – traces of 8 spots
\[ f = -0.585 \]

\( x^5 \) – 8 spots in place
\[ f = -0.707 \]

\( x^8 \) – 10th spot still missing...
\[ f = -0.865 \]

\( x^{12} \) – all 10 spots in place
\[ f = -0.872 \]

\( x^{35} \) – all 10 spots in place
\[ f = -0.886 \]

\( x^{43} \)...
\[ f = -0.896 \]
Progress in accuracy, noisy measurements.

solid line: Relative gap $\frac{\text{Gap}(t)}{\text{Gap}(1)}$ vs. step number $t$
- In 115 steps, the gap was reduced by factor 1580

dashed line: Progress in accuracy $\frac{f(x^t) - f}{f(x^1) - f}$ vs. step number $t$ ($f$ is the last lower bound on $f^*$ built in the run)
- In 115 steps, the accuracy was improved by factor $>460$
Mirror Descent Stochastic Approximation

Consider the case when solving a convex program

$$\text{Opt} = \min_{x \in X} f(x)$$

[$\bullet$ $X \subset \mathbb{R}^n$: convex compact $\bullet$ $f : X \to \mathbb{R}$ convex and Lipschitz]

no precise first order information is available. Specifically, we have at our disposal Stochastic Oracle (SO) as follows: at $t$-th call to the oracle, $x_t$ being the input, the oracle returns

$$g(x_t, \xi_t) \in \mathbb{R}, \quad G(x_t, \xi_t) \in \mathbb{R}^n$$

as random estimates of $f(x_t)$ and $f'(x_t)$, where $\xi_1, \xi_2, ...$ is a sequence of independent realizations of a random variable $\xi$ ("oracle's noise").

We assume that the SO is unbiased:

$$\mathbb{E}\{g(x, \xi)\} = f(x), \quad \mathbb{E}\{G(x, \xi)\} \in \partial f(x).$$

In addition, we assume that

$$\mathbb{E}\{\|G(x, \xi)\|_2^2\} \leq L^2 < \infty \ \forall x \in X$$
**Example:** Our $f$ is given as expectation:

$$f(x) = \int_\Xi F(x, \xi) dP(\xi),$$

where $F$ is convex in $x$ and efficiently computable.

When we cannot compute the expectation in a closed analytic form, but can instead sample from the distribution $P$, we, under mild regularity assumptions on $F$, have at our disposal unbiased Stochastic Oracle

$$g(x, \xi) = F(x, \xi), \quad G(x, \xi) = F'_x(x, \xi).$$

In this case, we can solve the problem with **Mirror Descent Stochastic Approximation** which is completely similar to MD:

$$x_1 \in X; x_{t+1} = \text{Prox}_{x_t}(\gamma_t G(x_t, \xi_t)), 1 \leq t \leq N;$$

$$x^N = \frac{1}{\gamma_1 + \ldots + \gamma_N} \sum_{t=1}^N \gamma_t x_t.$$

Here $\gamma_t > 0$ are deterministic stepsizes, and $\| \cdot \|$ and the function $\omega$ underlying the prox-mapping are given by the MD setup.
\[ x_1 \in X; x_{t+1} = \text{Prox}_{x_t}(\gamma_t G(x_t, \xi_t)), 1 \leq t \leq N; \]
\[ x^N = \frac{1}{\gamma_1 + \ldots + \gamma_N} \sum_{t=1}^{N} \gamma_t x_t. \]

Let us carry out convergence analysis of the algorithm. As always, we have

\[ \sum_{t=1}^{N} \gamma_t \langle G(x_t, \xi_t), x_t - x* \rangle \leq \Theta + \frac{1}{2} \sum_{t=1}^{N} \gamma_t^2 \| G(x_t, \xi_t) \|^2. \]

Taking expectations of both sides and taking into account that \( x_t \) is a deterministic function of \( x_1, \ldots, x_{t-1} \), while \( \xi_1, \ldots, \xi_N \) are independent, we get

\[ \sum_{t=1}^{N} \gamma_t \mathbb{E}\{ \langle f'(x_t), x_t - x* \rangle \} \leq \Theta + \frac{1}{2} \sum_{t=1}^{N} \gamma_t^2 L^2, \]

whence also

\[ \mathbb{E}\{ \sum_{t=1}^{N} \gamma_t [f(x_t) - f(x*)] \} \leq \Theta + \frac{1}{2} \sum_{t=1}^{N} \gamma_t^2 L^2 \]
\[
\sum_{t=1}^{N} \gamma_t \mathbb{E}\{\langle f'(x_t), x_t - x^* \rangle\} \leq \Theta + \frac{1}{2} \sum_{t=1}^{N} \gamma_t^2 L^2,
\]

By convexity,
\[
\mathbb{E}\{f(x^N) - f(x^*)\} \leq \left[\sum_{t=1}^{N} \gamma_t\right]^{-1} \mathbb{E}\left[\sum_{t=1}^{N} \gamma_t [f(x_t) - f(x^*)]\right] \leq \frac{\Theta + \frac{1}{2} \sum_{t=1}^{N} \gamma_t^2 L^2}{\sum_{t=1}^{N} \gamma_t},
\]

that is, we get exactly the same efficiency estimate as in the case of precise First Order oracle, but now – for the expected inaccuracy of the approximate solution \(x^N\) – the weighted sum of the search points we have generated.
Mirror Descent
for
Convex-Concave Saddle Point Problems

Convex-Concave Saddle Point problem is

\[ SV = \min_{x \in X} \max_{y \in Y} \phi(x, y) \]  

where:

- \( X \subset E_x, Y \subset E_y \) are nonempty closed and bounded convex sets in Euclidean spaces \( E_x, E_y \)
- \( \phi(x, y) : Z := X \times Y \to \mathbb{R} \) is the cost function which is Lipschitz continuous, convex in \( x \in X \) and concave in \( y \in Y \).

Solutions to \((SP)\) are, by definition, saddle points of \( \phi \) on \( X \times Y \), that is, points \( (x^*, y^*) \in X \times Y \) where \( \phi \) achieves its minimum in \( x \in X \) and its maximum in \( y \in Y \):

\[ \forall (x \in X, y \in Y) : \phi(x, y^*) \geq \phi(x^*, y^*) \geq \phi(x^*, y). \]
\[ SV = \min_{x \in X} \max_{y \in Y} \phi(x, y) \]  

**Fact:** (SP) gives rise to two optimization problems:

\[
(P) : \quad \text{Opt}(P) = \min_{x \in X} \left[ \phi(x) := \max_{y \in Y} \phi(x, y) \right] \\
= \min_{x \in X} \max_{y \in Y} \phi(x, y)
\]

\[
(D) : \quad \text{Opt}(D) = \max_{y \in Y} \left[ \phi(y) := \min_{x \in X} \phi(x, y) \right] \\
= \max_{y \in Y} \min_{x \in X} \phi(x, y)
\]

- We always have \( \text{Opt}(P) \geq \text{Opt}(D) \) ["weak duality"]
- \( \phi \) has saddle points on \( X \times Y \) iff both (P) and (D) are solvable with equal optimal values: \( \text{Opt}(P) = \text{Opt}(D) \), that is,

\[
\min_{x \in X} \max_{y \in Y} \phi(x, y) = \max_{y \in Y} \min_{x \in X} \phi(x, y)
\]

["strong duality"]. In this case the saddle points are exactly the pairs \( (x \in \text{Argmin}_X \phi, y \in \text{Argmax}_Y \phi) \).
(P): $\text{Opt}(P) = \min_{x \in X} \left[ \overline{\phi}(x) := \max_{y \in Y} \phi(x, y) \right]$

$= \min_{x \in X} \max_{y \in Y} \phi(x, y)$

(D): $\text{Opt}(D) = \max_{y \in Y} \left[ \underline{\phi}(y) := \min_{x \in X} \phi(x, y) \right]$

$= \max_{y \in Y} \min_{x \in X} \phi(x, y)$

- Under our standing assumption ($X, Y$ are nonempty convex compacts, $\phi$ is Lipschitz continuous convex-concave), both (P) and (D) are solvable with equal optimal values, that is, saddle points do exist.

♠ It is natural to quantify the (in)accuracy of an approximate saddle point $(x, y) \in Z := X \times Y$ by its saddle point residual

$$\epsilon_{\text{Sad}}(x, y) = \overline{\phi}(x) - \underline{\phi}(y) = [\overline{\phi}(x) - \text{Opt}(P)] + [\text{Opt}(D) - \underline{\phi}(y)]$$

This residual always is nonnegative and is zero iff $(x, y)$ is a saddle point of $\phi$. 

6.93
Vector field associated with a saddle point problem. Under our standing assumptions, we can associate with a convex-concave saddle point problem

\[ \min_{x \in X} \max_{y \in Y} \phi(x, y) \]

vector field

\[ F(z = [x; y]) = [F_x(x, y); F_y(x, y)] : Z := X \times Y \to E_z := E_x \times E_y \]

with

\[ F_x(x, y) \in \partial_x \phi(x, y), \quad F_y(x, y) \in \partial_y [-\phi(x, y)] \]

Note: In the sequel, we equip \( E_x \) with a norm \( \| \cdot \|_x \), and \( E_y \) with a norm \( \| \cdot \|_y \). Denoting by \( L_x, L_y \) the Lipschitz constants of \( \phi \) w.r.t. these norms:

\[ \forall (x, x' \in X, y, y' \in Y) : |\phi(x, y) - \phi(x', y')| \leq L_x \|x - x'|_x + L_y \|y - y'|_y \]

we assume by default that the field \( F \) satisfies

\[ \forall (x, y) \in X \times Y : \|F_x(x, y)\|_{x,*} \leq L_x, \|F_y(x, y)\|_{y,*} \leq L_y. \]
\[
F(z = [x; y]) = [F_x(x, y); F_y(x, y)] : Z := X \times Y \to E_z := E_x \times E_y \\
F_x(x, y) \in \partial_x \phi(x, y), \quad F_y(x, y) \in \partial_y[-\phi(x, y)]
\]

♦ **Facts:**

- **F is monotone:**
  \[
  \forall (z, z' \in Z := X \times Y) : \langle F(z) - F(z'), z - z' \rangle \geq 0
  \]

  Indeed, setting \( z = (x, y), \ z' = (x', y') \), we have

  \[
  \langle F(z) - F(z'), z - z' \rangle = \langle F_x(x, y) - F_x(x', y'), x - x' \rangle + \langle F_y(x, y) - F_y(x', y'), y - y' \rangle \\
  \geq [\phi(x, y) - \phi(x', y')] + [\phi(x', y') - \phi(x, y')] + [(-\phi)(x, y) - (-\phi)(x, y')] + [(-\phi)(x', y') - (-\phi)(x', y)] \\
  \equiv 0
  \]

- **Saddle points of \( \phi \) on \( Z = X \times Y \) are exactly the points \( z_* \in Z \) such that
  \[
  \langle F(z), z - z_* \rangle \geq 0 \forall z \in Z.
  \]
$$\text{SV} = \min_{x \in X} \max_{y \in Y} \phi(x, y) \quad \text{(SP)}$$

• $X \subset E_x, Y \subset E_y$ are nonempty closed and bounded convex sets in Euclidean spaces $E_x, E_y$
• $\phi(x, y) : Z := X \times Y \to \mathbb{R}$ is the cost function which is Lipschitz continuous, convex in $x \in X$ and concave in $y \in Y$.

ющее Problems (SP) arise in a wide spectrum of applications. Our major interest in these problems stems from the fact that numerous "complex" and nonsmooth convex functions $f(x)$ admit saddle point representation:

$$f(x) = \max_{y \in Y} \phi(x, y)$$

with convex-concave and smooth functions $\phi$, which allows to reduce a nonsmooth minimization problem

$$\min_{x \in X} f(x)$$

to a smooth convex-concave saddle point problem

$$\min_{x \in X} \max_{y \in Y} \phi(x, y)$$

and this “gain in smoothness” possesses dramatic potential as far as computationally cheap First Order methods are concerned.

6.96
Examples of saddle point reformulations:

- **Maximum of smooth convex functions:**
  \[
  f(x) := \max_{1 \leq i \leq m} f_i(x) = \max_{y \in Y} [\phi(x, y) := \sum_i y_i f_i(x)] \\
  [Y = \{y \geq 0, \sum_i y_i = 1\}]
  \]
  When \( f_i \) are smooth, so is \( \phi \); when \( f_i \) are linear, \( \phi \) is just bilinear.

- **Norm-type functions:**
  \[
  \|Ax - b\| = \max_{y : \|y\| \leq 1} [\phi(x, y) = \langle y, Ax - b \rangle]
  \]

- **Maximal eigenvalue of a symmetric matrix:**
  \[
  \lambda_{\text{max}}(x) = \max_{y \in Y} [\phi(x, y) = \text{Tr}(xy)] \\
  Y = \{y \geq 0 : \text{Tr}(y) = 1\}
  \]

**Note:** Smooth/bilinear saddle point representations admit fully algorithmic calculus. For example,

\[
\begin{align*}
\hat{f}_i(x) &= \max_{y_i \in Y_i} [\langle a_i, x \rangle + \langle b_i, y_i \rangle + \langle x, A_i y_i \rangle], \quad \lambda_i \geq 0 \\
\Rightarrow \sum_i \lambda_i \hat{f}_i(x) &= \max_{y = [y_1; \ldots; y_k] \in Y_1 \times \ldots \times Y_k} [\sum_i \langle \lambda_i a_i, x \rangle + \langle \lambda_i b_i, y_i \rangle + \langle x, \lambda_i A_i y_i \rangle] \\
&= \max_{y = [y_1; \ldots; y_k] \in Y_1 \times \ldots \times Y_k} [\langle \sum_i \lambda_i a_i, x \rangle + \langle [\lambda_1 b_1; \ldots; \lambda_k b_k], y \rangle + \langle x, [\lambda_1 A_1, \ldots, \lambda_k A_k] y \rangle]
\end{align*}
\]
\[ SV = \min_{x \in X} \max_{y \in Y} \phi(x, y) \quad \text{(SP)} \]
\[ \Rightarrow F(z = [x; y]) = [F_x(x, y) \in \partial_x \phi(x, y); F_y(x, y) \in \partial_y [-\phi(x, y)]] \]

- \( X \subset E_x, Y \subset E_y \) are nonempty closed and bounded convex sets in Euclidean spaces \( E_x, E_y \)
- \( \phi(x, y) : Z := X \times Y \rightarrow \mathbb{R} \) is the cost function which is Lipschitz continuous, convex in \( x \in X \) and concave in \( y \in Y \).

\( \spadesuit \) (SP) can be solved by MD. Indeed, let \( \| \cdot \| \) be a norm on \( E = E_x \times E_y \) and \( \omega(\cdot) \) be a DGF for \( Z = X \times Y \) which is compatible with \( \| \cdot \| \). Consider the process

\[ z_1 \in Z; z_{t+1} = \text{Prox}_{z_t}(\gamma_t F(z_t)); z^t = [\sum_{\tau=1}^t \gamma_\tau]^{-1} \sum_{\tau=1}^t \gamma_\tau z_\tau \]
\[ [z_\tau = [x_\tau; y_\tau]] \]
\[ z_1 \in Z; z_{t+1} = \text{Prox}_{z_t}(\gamma_t F(z_t)); z^t = \left[ \sum_{\tau=1}^{t} \gamma_\tau \right]^{-1} \sum_{\tau=1}^{t} \gamma_\tau z_\tau \]

As always, we have
\[
\forall u = [\xi; \eta] \in Z : \sum_{\tau=1}^{t} \gamma_\tau \langle F(z_\tau), z_\tau - u \rangle \leq \Theta + \frac{1}{2} \sum_{\tau=1}^{T} \gamma_\tau^2 \| F(z_\tau) \|_*^2
\]

and
\[
\langle F(z_\tau), z_\tau - u \rangle = \langle \phi'_x(x_\tau, y_\tau), x_\tau - \xi \rangle + \langle -\phi'_y(x_\tau, y_\tau), y_\tau - \eta \rangle \\
\geq [\phi(x_\tau, y_\tau) - \phi(\xi, y_\tau)] + [-\phi(x_\tau, y_\tau) + \phi(x_\tau, \eta)] \\
= \phi(x_\tau, \eta) - \phi(\xi, y_\tau)
\]

⇒ setting \( \Gamma_t = \sum_{\tau=1}^{t} \gamma_\tau \) and \( \lambda_\tau = \gamma_\tau / \Gamma_t \), we get
\[
\sum_{\tau=1}^{t} \lambda_\tau [\phi(x_\tau, \eta) - \phi(\xi, y_\tau)] \leq \frac{\Theta + \frac{1}{2} \sum_{\tau=1}^{t} \gamma_\tau^2 \| F(z_\tau) \|_*^2}{\sum_{\tau=1}^{t} \gamma_\tau}.
\]

⇒ \( \forall ([\xi; \eta] \in X \times Y) : \phi(x^t, \eta) - \phi(\xi, y^t) \leq \frac{\Theta + \frac{1}{2} \sum_{\tau=1}^{t} \gamma_\tau^2 \| F(z_\tau) \|_*^2}{\sum_{\tau=1}^{t} \gamma_\tau} \).

The supremum of the left hand side in \( \xi \in X, \eta \in Y \) is \( \epsilon_{\text{Sad}}(x^t, y^t) \), and we arrive at
\[
\epsilon_{\text{Sad}}(x^t, y^t) \leq \frac{\Theta + \frac{1}{2} \sum_{\tau=1}^{T} \gamma_\tau^2 \| F(z_\tau) \|_*^2}{\sum_{\tau=1}^{T} \gamma_\tau},
\]
with all consequences related to the rate of convergence, stepsize policies, etc.
Consider the *extragradient* Saddle Point MD:

\[ z_1 \in Z; w_t = \text{Prox}_{z_t}(\gamma_tF(z_t)); z_{t+1} = \text{Prox}_{z_t}(\gamma_tF(w_t)); \]

\[ z^t = \left[ \sum_{t=1}^{\infty} \gamma_t \right]^{-1} \sum_{t=1}^{\infty} \gamma_t w_t \]

Magic Inequality states

(a) \quad \forall u \in Z : \langle \gamma_tF(w_t), z_{t+1} - u \rangle \leq V_{z_t}(u) - V_{z_{t+1}}(u) - V_{z_t}(z_{t+1})

(b) \quad \forall v \in Z : \langle \gamma_tF(z_t), w_t - v \rangle \leq V_{z_t}(v) - V_{w_t}(v) - V_{z_t}(w_t)

Applying (b) to \( v = z_{t+1} \), we get

\[ \langle \gamma_tF(z_t), w_t - z_{t+1} \rangle \leq V_{z_t}(z_{t+1}) - V_{w_t}(z_{t+1}) - V_{z_t}(w_t), \]

while (a) implies

\[ \langle \gamma_tF(w_t), w_t - u \rangle \leq V_{z_t}(u) - V_{z_{t+1}}(u) - V_{z_t}(z_{t+1}) + \gamma_t \langle F(w_t), w_t - z_{t+1} \rangle \]

\[ \Rightarrow \langle \gamma_tF(w_t), w_t - u \rangle \leq V_{z_t}(u) - V_{z_{t+1}}(u) - V_{z_t}(z_{t+1}) + \gamma_t \langle F(z_t), w_t - z_{t+1} \rangle + \gamma_t \langle F(w_t) - F(z_t), w_t - z_{t+1} \rangle \]

Taken together, these inequalities imply that

\[ \langle \gamma_tF(w_t), w_t - u \rangle \leq V_{z_t}(u) - V_{z_{t+1}}(u) + \left[ \gamma_t \langle F(w_t) - F(z_t), w_t - z_{t+1} \rangle - V_{w_t}(z_{t+1}) - V_{z_t}(w_t) \right] \]

\[ \leq V_{z_t}(u) - V_{z_{t+1}}(u) + \left[ \frac{1}{2} \gamma_t^2 \| F(z_t) - F(w_t) \|^2 - V_{z_t}(w_t) \right] \]

Now let \( F \) be Lipschitz: \( \| F(z) - F(z') \|_* \leq L \| z - z' \|. \) Since \( V_{z_t}(w_t) \geq \frac{1}{2} \| w_t - z_t \|^2 \), we get

\[ \langle \gamma_tF(w_t), w_t - u \rangle \leq V_{z_t}(u) - V_{z_{t+1}}(u) + \frac{1}{2} \| w_t - z_t \|^2 \left[ L^2 \gamma_t^2 - 1 \right], \]

and we end up with

\[ \gamma_t \equiv \gamma = \frac{1}{L} \forall t \Rightarrow \gamma \langle F(w_t), w_t - u \rangle \leq V_{z_t}(u) - V_{z_{t+1}}(u) \forall u \in Z, \]

whence, like above,

\[ \epsilon_{Sad}(z^t) \leq \Theta \frac{\Theta L}{t}, t = 1, 2, \ldots \]

[1/t rate!!!]
Acceleration by Randomization

Consider a convex-concave saddle point problem

\[ SV = \min_{x \in X} \max_{y \in Y} \phi(x, y) \quad \text{(SP)} \]

\[ \Rightarrow F(z = (x, y)) = [F_x(x, y) \in \partial_x \phi(x, y); F_y(x, y) \in \partial_y [−\phi(x, y)]] \]

- \( X \subset E_x, Y \subset E_y \) are nonempty closed and bounded convex sets in Euclidean spaces \( E_x, E_y \), \( \phi \) is Lipschitz continuous and convex-concave

Assume that the field \( F \) is given by Stochastic Oracle: when calling the oracle at step \( t \), the query point being \( z_t = (x_t, y_t) \), the oracle returns a random estimate \( G(z_t, \xi_t) \) of \( F(z_t) \) which is unbiased and “stochastically bounded”:

\[ \forall z \in Z = X \times Y : E\{G(z, \xi)\} = F(z) \land E\{|G(z, \xi)|^2\} \leq L^2. \]

As always, \( \xi_1, \xi_2, \ldots \) are independent realizations of a random variable \( \xi \).
SV = \min_{x \in X} \max_{y \in Y} \phi(x, y) \quad \text{(SP)}

♠ When using MD:

\[ z_1 = z_0; \quad z_{t+1} = \text{Prox}_{z_t}(\gamma_t G(x_t, \xi_t)); \quad z^t = \left[ \sum_{\tau=1}^{t} \gamma_{\tau} \right]^{-1} \sum_{\tau=1}^{T} \gamma_{\tau} z_{\tau}. \]

it is easy to arrive at

**Theorem:** One has \( E\{\epsilon_{\text{Sad}}(z^t)\} \leq 3 \frac{2\Theta + L^2 \sum_{\tau=1}^{t} \gamma_{\tau}^2}{\sum_{\tau=1}^{t} \gamma_{\tau}} \). In particular, given a number \( N \) of iterations and setting

\[ \gamma_t = \frac{\sqrt{2\Theta}}{L \sqrt{N}}, \quad 1 \leq t \leq N, \]

we ensure that

\[ E\{\epsilon_{\text{Sad}}(z^T)\} \leq 6 \frac{\sqrt{2\Theta} L}{\sqrt{N}}. \]

Here, as always, \( \Theta \) is the capacity of \( Z \) w.r.t. the distance-generating function underlying the algorithm.

**Note:** Similar results hold true for Mirror Prox.
Application: Matrix Game. Matrix Game problem is as follows:

\[
SV = \min_{x \in \Delta_n} \max_{y \in \Delta_m} y^T Ax \\
[\Delta_p = \{ u \in \mathbb{R}^p : u \geq 0, \sum_i u_i = 1 \}]
\]

Interpretation: Two players are playing an antagonistic game; the first selects a \( j \in \{1, \ldots, n\} \), the second selects an \( i \in \{1, \ldots, m\} \). The loss of the first player (i.e., the profit of the second player) is \( A_{ij} \), where \( A \) is a given \( m \times n \) matrix. Naturally, the first player is interested to reduce his losses, while the second player has the opposite interest.
When players make their choices simultaneously, there is no natural definition of “equilibrium,” unless the matrix has a “saddle point” – some entry $A_{i,j}$ is minimal in its column and is maximal in its row.

In the general case, the concept of a solution to the game, going back to von Neumann and Morgenstern, is to look what happens when the players repeat the matrix game many times, drawing their choices at random independently of each other and across the time. Denoting by $x \in \Delta_n$ the probability distribution from which the first player draws his choices, and by $y \in \Delta_m$ similar distribution for the second player, the expected loss of the first player (expected profit of the second player) will be

$$y^T A x$$

Thus, (MG) can be thought of as the problem of finding the best randomized policies of the players (called their mixed strategies); if both players are interested in their long run losses and profits, sticking to the mixed strategies given by a saddle point of the bilinear (and thus convex-concave) game (MG) will be optimal policies for every one of them.
SV = \min_{x \in \Delta_n} \max_{y \in \Delta_m} y^T A x \quad \text{(MG)}

[\Delta_p = \{u \in \mathbb{R}^p : u \geq 0, \sum_i u_i = 1\}]

(MG) is just a primal-dual pair of LP programs:

\begin{align*}
\text{Opt}(P) &= \min_{x \in \Delta_n} \max_i \text{Row}_i^T[A] x \\
\text{Opt}(D) &= \max_{y \in \Delta_m} \min_j \text{Col}_j^T[A] y
\end{align*}

where Row$_i^T[A]$, is $i$-th row, and Col$_j[A]$ is $j$-th column in $A$.

⇒ (MG) can be solved by interior point LP methods.
\[ SV = \min_{x \in \Delta_n} \max_{y \in \Delta_m} y^T A x \quad \text{(MG)} \]

\[
\Delta_p = \{ u \in \mathbb{R}^p : u \geq 0, \sum_i u_i = 1 \} \]

\[ \epsilon_{\text{Sad}}(x^N, y^N) \leq O(1) \sqrt{\frac{\ln(n) \ln(m)}{N}} \| A \|_{1 \rightarrow \infty}, \quad \| A \|_{1 \rightarrow \infty} = \max_{i,j} |A_{ij}| \]

In the large-scale case, (MG) can be solved by Mirror Prox; with appropriate setup, MP yields the efficiency estimate

The complexity of a step is \( O(m + n) \) plus the complexity of two matrix-vector multiplications:

\[ \Delta_n \ni x \mapsto Ax, \quad \Delta_m \ni y \mapsto A^T y \]

needed to compute the associated with (MG) vector field

\[ F(x, y) = \begin{bmatrix} A^T \\ -A \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \]

When \( A \) is a general-type dense matrix, the complexity of finding and \( \epsilon \)-solution to the problem is therefore

\[ C_{\text{detem}}(\epsilon) = O(1) \sqrt{\ln(m) \ln(n)} mn \frac{\| A \|_{1 \rightarrow \infty}}{\epsilon} \text{ flop.} \]

Can we do better?
Observation: Computing matrix-vector multiplication
\[ \mathbb{R}^p \ni u \mapsto Bu \in \mathbb{R}^q \]
is easy to randomize:
— the vector \( v = \text{abs}[u]/\|u\|_1 \) (abs acts coordinatewise) is a probabilistic vector (non-negative entries summing up to 1). Treating \( v \) as a probability distribution on \( \{1, 2, \ldots, p\} \), we draw at random an index \( j \) from this distribution and return
\[ \eta = \|u\|_1 \text{sign}(u_j) \text{Col}_j(B), \]
thus ensuring that \( \mathbb{E}\{\eta\} = Bu \).
— generating a realization of \( \eta \) is cheap:
— drawing \( j \) costs \( O(p) \) flop: in \( O(p) \) flop one computes the “cumulative distribution”
\[ U_j = \|u\|_1^{-1} \sum_{k<j} |u_k|, \quad 1 \leq j \leq p, \]
of the probabilistic vector, generates \( \zeta \sim \text{Uniform}[0, 1] \) and needs \( O(\ln(p)) \) comparisons to find by Bisection \( j \) such that
\[ U_{j-1} < \zeta \leq U_j \]
— after \( j \) is generated, computing \( \eta \) takes just \( O(m) \) flop
whatever be a norm \( \| \cdot \| \), the noise of our oracle is under control:
\[ \|\eta\| \leq \|u\|_1 \max_j \|\text{Col}_j[B]\|. \]
The situation is especially nice when \( \|u\|_1 \) can be bounded in advance.
\[ SV = \min_{x \in \Delta_n} \max_{y \in \Delta_m} y^T Ax \]

\[ [\Delta_p = \{ u \in \mathbb{R}^p : u \geq 0, \sum_i u_i = 1\}] \Rightarrow F(x, y) = \begin{bmatrix} -A \\ A^T \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \]

Applying the above approach to (MG), we get a cheap randomized oracle for \( F \); a call to this oracle costs just \( O(m + n) \) flop, vs. the cost \( O(mn) \) of the precise computation of \( F \).

\[ \Rightarrow \text{Utilizing the cheap stochastic oracle in MD, we get an algorithm for solving (MG) which ensures} \]

\[ E\{\epsilon_{\text{Sad}}(x^N, y^N)\} \leq O(1) \sqrt{\ln(m) \ln(n)} \left( \frac{\|A\|_{1 \to \infty}}{\sqrt{N}} \right), \]

with \( O(m + n) \) flop per step.

\[ \Rightarrow \text{For every } \epsilon > 0, \delta \in (0, 1), \text{ one can build in } (1 - \delta)\text{-reliable fashion an } \epsilon\text{-solution to (MG) at the cost of} \]

\[ C_{\text{rand}}(\epsilon) = O(1) \ln(n) \ln(m) (m + n) \frac{1}{\chi^2} \text{ flop} \quad \text{[} \chi = \epsilon/\|A\|_{1 \to \infty} : \text{relative accuracy} \text{]} \]

which for fixed relative accuracy \( \chi \) and large \( m, n \) is by orders of magnitude better than the best known “deterministic price”

\[ C_{\text{determ}}(\epsilon) = O(1) \sqrt{\ln(m) \ln(n)} mn \frac{1}{\chi} \text{ flop.} \]

of \( \epsilon\)-solution to (MG).
$$C_{\text{rand}}(\epsilon) = O(1) \ln(n) \ln(m)(m + n)\frac{1}{\chi^2} \text{ flop} \quad [\chi = \epsilon/\|A\|_{1\rightarrow\infty} : \text{relative accuracy}]$$

**Note:** Our algorithm exhibits *sublinear time behavior:* for fixed $\chi$ and large $m, n$, reliable building of $\epsilon$-solution requires inspection of a negligibly small, going to 0 as $m, n$ grow, randomly selected fraction of the data.

An “ad hoc” algorithm with this property (in retrospect, pretty similar to Stochastic MD Approximation) was discovered in 1995 by Grigoriadis and Khachiyan.
Illustration: There are $N$ houses in a city, $i$-th with wealth $w_i$. Every evening, Burglar selects a house $i$ to be attacked, and Policeman selects his location by a house $j$. When the burglary starts, the probability for Policeman to react to alarm and to prevent the burglary is $\exp\{-\theta d(i,j)\}$, where $d(i,j)$ is the distance between locations $i$ and $j$, so that the expected profit of Burglar is $A_{ij} = w_i[1 - \exp\{-\theta d(i,j)\}]$. Our goal is to solve in mixed strategies the resulting game

$$\max_{y \in \Delta_N} \min_{x \in \Delta_N} x^T Ay.$$ 

Assuming an $n \times n$ equidistant grid of houses with wealth decreasing from the downtown to outskirts, the resulting $(N := n^2) \times N$ matrix game was solved by the state-of-the-art commercial LP Interior Point Method (IPM) mosekopt, by the Deterministic Mirror Prox and by the randomized MD seeking $\epsilon_{sad} < 0.001$, with CPU limit of 5,300 sec. Here are the results:

<table>
<thead>
<tr>
<th>$N$</th>
<th>Steps</th>
<th>CPU</th>
<th>Gap</th>
<th>Steps</th>
<th>CPU</th>
<th>Gap</th>
<th>Steps</th>
<th>CPU</th>
<th>Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1600</td>
<td>21</td>
<td>120</td>
<td>6.0e-9</td>
<td>78</td>
<td>6</td>
<td>1.0e-3</td>
<td>10556</td>
<td>264</td>
<td>1.0e-3</td>
</tr>
<tr>
<td>6400</td>
<td>21</td>
<td>6930</td>
<td>1.1e-8</td>
<td>80</td>
<td>31</td>
<td>1.0e-3</td>
<td>10408</td>
<td>796</td>
<td>1.0e-3</td>
</tr>
<tr>
<td>14400</td>
<td>not tested</td>
<td>95</td>
<td>171</td>
<td>1.0e-3</td>
<td>9422</td>
<td>1584</td>
<td>1.0e-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40000</td>
<td>out of memory</td>
<td>15</td>
<td>5533</td>
<td>0.022</td>
<td>10216</td>
<td>4931</td>
<td>1.0e-3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Policeman vs. Burglar, $N$ houses
Wealth

Policeman

Burglar

Duality gap vs. iteration count

Policeman vs. Burglar, $N = 40,000$. RMD with 10,216 steps (4931 sec)

6.111
Smooth Convex Minimization: 
Nesterov’s Fast Gradient Method

♣ Problem of interest: Composite minimization

\[
\text{Opt} = \min_{x \in X} \{ \phi(x) = \Psi(x) + f(x) \}
\]

- \( X \): closed convex nonempty subset in Euclidean space \( E \)
  
  \( (X, E) \) is equipped with proximal setup \( (\omega(\cdot), \|\cdot\|) \)

- \( \Psi : X \to \mathbb{R} \cup \{+\infty\} \): convex lower semicontinuous
  function which is finite on the relative interior of \( X \)

- \( f : X \to \mathbb{R} \): represented by FO oracle convex function
  with Lipschitz continuous gradient:
  \[
  \forall x, y \in X : \| \nabla f(x) - \nabla f(y) \|_* \leq L_f \| x - y \|
  \]

♣ Main Assumption: We are able to compute composite prox-mappings, i.e., solve auxiliary problems

\[
\min_{x \in X} \{ \omega(x) + \langle h, x \rangle + \alpha \Psi(x) \} \quad [\alpha \geq 0]
\]
**Example:** LASSO problem

\[
\min_{x \in X} \left\{ \lambda \|x\|_E + \frac{1}{2} \|A(x) - b\|_2^2 \right\}
\]

\[
\begin{cases}
(a) \text{ block } \ell_1 \text{ norm } \sum_{j=1}^{n} \|x^j\|_2 \text{ on } E = \mathbb{R}^{k_1} \times \ldots \times \mathbb{R}^{k_n} \ (\ell_1 \ case) \\
(b) \text{ nuclear norm on the space } E \text{ of block diagonal matrices of a given block diagonal structure} \ (\text{nuclear norm case})
\end{cases}
\]

- \[\| \cdot \|_E: \]
- \[A(\cdot): E \rightarrow \mathbb{R}^m: \text{ linear mapping}\]
- \[X: \text{ either the unit } \| \cdot \|_E-\text{ball, or the entire } E\]

For properly chosen proximal setup, Main Assumption is satisfied: computing composite prox mapping

\[
\min_{x \in X} \{ \omega(x) + \langle h, x \rangle + \alpha \Psi(x) \} \quad [\alpha \geq 0]
\]

takes \(O(\dim E)\) a.o. in the case of (a) and reduces to finding svd of a matrix from \(E\) in the case of (b).
Nesterov’s Fast Gradient algorithm for Composite Minimization

♣ Problem:

\[ \text{Opt} = \min_{x \in X \subset E} \{ \phi(x) := \Psi(x) + f(x) \} \]

- \( \Psi, f \): convex and
- \( \forall x, y \in X : \| \nabla f(x) - \nabla f(y) \|_* \leq L_f \| x - y \| \)  

(\( CP \))

♣ Assumptions: \( L_f \) is known and \( (CP) \) is solvable with an optimal solution \( x_* \).

♣ The algorithm is described in terms of proximal setup \( (\omega(\cdot), \| \cdot \|) \) for \( X \) and auxiliary sequence

\[ \{ L_t \in (0, L_f] \}_{t=0}^\infty \]

which can be adjusted on-line.

Recall that DGF \( \omega \) defines Bregman distance

\[ V_x(y) = \omega(y) - \omega(x) - \langle \omega'(x), y - x \rangle \] \( [x \in X^o, y \in X] \)
\[\text{Opt} = \min_{x \in X \subset E} \{ \phi(x) := \Psi(x) + f(x) \}\]

\[\text{♣ Algorithm:}\]
\[\text{♠ Initialization: Set}\]
\[A_0 = 0, \ y_0 = x_\omega = \arg\min_X \omega, \ \psi_0(x) = V_{x_\omega}(x)\]

and select \(y_0^+ \in X\) such that \(\phi(y_0^+) \leq \phi(y_0)\).

\[\text{♣ Step } t = 0, 1, 2, \ldots: \text{ Given } \psi_t(\cdot) = \omega(\cdot) + \alpha \Psi(\cdot) + \langle \text{affine form} \rangle \ [\alpha \geq 0], \ y_t^+ \in X \text{ and } L_t, 0 < L_t \leq L_f,\]

- Compute \(z_t = \arg\min_{x \in X} \psi_t(x)\) (reduces to computing composite prox-mapping)
- Find the positive root \(a_{t+1}^+\) of the equation \(L_t a_{t+1}^2 = A_t + a_{t+1}\) and set \(A_{t+1} = A_t + a_{t+1}, \ \tau_t = a_{t+1}/A_{t+1} \in (0, 1]\)
- Set \(x_{t+1} = \tau_t z_t + (1 - \tau_t) y_t^+\) and compute \(f(x_{t+1}), \nabla f(x_{t+1})\)
- Compute \(\hat{x}_{t+1} = \arg\min_{x \in X} \left\{ \langle \nabla f(x_{t+1}), x \rangle + \psi(x) + \frac{1}{a_{t+1}} V_{z_t}(x) \right\}\) (reduces to computing composite prox-mapping)
- Set

\[y_{t+1} = \tau_t \hat{x}_{t+1} + (1 - \tau_t) y_t^+\]

\[\psi_{t+1}(x) = \psi_t(x) + a_{t+1} \left[ f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1} \rangle + \Psi(x) \right]\]

Step \(t\) is completed; go to step \(t + 1\).
Theorem [Yu. Nesterov ’83, ’07] Assume that \( \{L_t \in (0, L_f]\} \) is such that

\[
\frac{V_t(\hat{x}_{t+1})}{A_{t+1}} + \langle \nabla f(x_{t+1}), y_{t+1} - x_{t+1} \rangle + f(x_{t+1}) \geq f(y_{t+1})
\]

(this for sure is the case when \( L_t \equiv L_f \)). Then

\[
\phi(y_t^+) \leq A_t^{-1} V_{x_\omega}(x^*) \leq \frac{4L_f}{\mu^2} V_{x_\omega}(x^*), \quad t = 1, 2, ...
\]
Illustration: As applied to a solvable LASSO problem

\[ x^*_t = \arg\min_x \left\{ \phi(x) := \lambda \|x\|_E + \frac{1}{2} \|A(x) - b\|_2^2 \right\} \]

with \( \| \cdot \|_E \) either (a) block \( \ell_1 \) norm on \( E = \mathbb{R}^{k_1} \times \ldots \times \mathbb{R}^{k_n} \), or (b) nuclear norm on \( E = \mathbb{R}^{p \times q} \) with \( n = \min[p, q] \), the Fast Gradient method in \( t = 1, 2, \ldots \) steps ensures

\[ \phi(y^+_t) \leq \text{Opt} + O(\ln(n + 1)) \frac{\|A\|_{E,2}^2}{t^2} \|x^*\|_E^2 \]

where \( \|A\|_{E,2} = \max\{\|A(x)\|_2 : \|x\|_E \leq 1\} \)
Note: $O(1/t^2)$ rate of convergence is, seemingly, the best one can expect from oracle-based methods in the large scale case.

The precise statement is as follows:

Let $n$ be a positive integer. Consider Least Squares problems

$$\text{Opt} = \min_x \|Ax - b\|_2^2$$

(QP)

with $n \times n$ symmetric matrices $A$.

For every positive reals $R, L$ and every number $t \leq n/4$ of steps, for every $t$-step solution algorithm $B$ operating with the “multiplication oracle” $u \mapsto Au$ one can find an instance of (QP) such that

- the spectral norm of $A$ does not exceed $L$,
- $\text{Opt} = 0$, and the $\| \cdot \|_2$-norm of some optimal solution does not exceed $R$,
- the approximate solution $y$ generated by $B$, as applied to the instance, after $t$ calls to the oracle, satisfies

$$\|Ay - b\|_2^2 \geq O(1) \frac{L^2R^2}{t^2}$$
How it Works:
Fast Composite Minimization for LASSO

♣ Test problem:

\[
\text{Opt} = \min_x \left\{ \phi(x) := 0.01 \|x\|_1 + \frac{1}{2} \|Ax - b\|_2^2 \right\}
\]

with \(4096 \times 2048\) randomly generated matrix \(A\).

<table>
<thead>
<tr>
<th>Method</th>
<th>Setup</th>
<th>Iterations</th>
<th>CPU, sec</th>
<th>Nonoptimality</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPM</td>
<td>—</td>
<td>11</td>
<td>103.1</td>
<td>&lt;1.e-12</td>
</tr>
<tr>
<td>FGr</td>
<td>Ball setup</td>
<td>512</td>
<td>36.3</td>
<td>2.4e-6</td>
</tr>
<tr>
<td>FGr</td>
<td>Simplex setup</td>
<td>512</td>
<td>36.5</td>
<td>1.2e-7</td>
</tr>
</tbody>
</table>

Platform: \(2 \times 3.40\) GHz CPU, 16.0 GB RAM, 64-bit Windows 7
Beyond the Scope of Proximal Algorithms: Conditional Gradients

\[ \text{Opt} = \min_{x \in X} f(x) \]

♣ Fact: All considered so far “computationally cheap” large scale alternatives to IPM’s were proximal type First Order methods

♠ But: In order to be computationally cheap, a proximal type method should operate with problems on Favorable Geometry domains \( X \) (those with “moderate” \( \Theta \), in order to have a reasonable iteration count in the large scale case) admitting easy to compute prox-mappings (“Simple Geometry”, otherwise an iteration becomes expensive).
Both Favorable and Simple Geometry requirements can be violated. For example,

- when $X$ is a box, Favorable Geometry is missing
- when $X$ is a nuclear norm ball in $\mathbb{R}^{n \times n}$ or a spectahedron in $\mathbb{S}^n$, we do have Favorable Geometry, but computing the associated prox-mapping requires singular value decomposition of $n \times n$ matrix (or the eigenvalue decomposition of a symmetric $n \times n$ matrix), and both these computations require

\[ O(n^3) = O((\dim X)^{3/2}) \text{ a.o.} \]

While much cheaper than the cost $O((\dim X)^3) = O(n^6)$ a.o. of an IPM iteration, $O(n^3)$ a.o. prox-mapping for large $n$ becomes prohibitively time consuming.

**Note:** nuclear norm balls/spectahedrons arise naturally in many important applications, including, but not reducing to, low rank matrix recovery, multi-class classification in Machine Learning and high dimensional Statistics (and more generally – large scale Semidefinite programming).
Another important example of generic problem with Complex Geometry is Total Variation based Image Reconstruction

\[ \min_{x \in \mathbb{R}^{m \times n}} \left\{ \lambda \cdot TV(x) + \frac{1}{2} \| A(x) - b \|_2^2 \right\}, \]

where \( x = [x_{ij}] \in \mathbb{R}^{m \times n} \) is an \((m \times n)\)-pixel image, and \( TV(x) \) is the Total Variation:

\[ TV(x) = \sum_{i=1}^{m-1} \sum_{j=1}^{n} |x_{i+1,j} - x_{i,j}| + \sum_{i=1}^{m} \sum_{j=1}^{n-1} |x_{i,j+1} - x_{i,j}| \]

— the \( \ell_1 \)-norm of the discrete gradient of \( x = [x_{ij}] \). Restricted to the space \( M_{0}^{m,n} \) of \( m \times n \) images with zero mean, TV becomes a norm. For the unit TV-ball, no DGF compatible with the TV norm and leading to easy-to-compute prox mapping is known...
Linear Minimization Oracle

♣ Observation: When \( X \subset E \) admits a proximal setup with easy-to-compute prox-mapping, \( X \) definitely admits a computationally cheap Linear Minimization Oracle (LMO) — a procedure which, given on input a linear form \( \langle \eta, \cdot \rangle \), returns \( x[\eta] \in \text{Argmin}_{x \in X} \langle \eta, x \rangle \).

Indeed, the optimization program

\[
\min_{x \in X} \langle \eta, x \rangle
\]

is the “limiting case,” as \( \theta \to 0^+ \), of the programs

\[
\min_{x \in X} \{ \theta \omega(x) + \langle \eta, x \rangle \}.
\]

♠ Fact: Admitting a cheap LMO is a much weaker requirement than admitting proximal setup with cheap prox-mapping, and there are important domains \( X \) with Complex Geometry admitting relatively cheap Linear Minimization Oracle.
Examples:

A: Nuclear Norm ball $X = \{ x \in \mathbb{R}^{m \times n} : \| x \|_{\text{nuc}} \leq 1 \}$. Here computing $x[\eta]$ reduces to finding the left and the right leading singular vectors of $\eta \in \mathbb{R}^{m \times n}$, i.e., to solving the problem

$$\max_{\| u \|_2 \leq 1, \| v \|_2 \leq 1} u^T \eta v.$$  

For large $m, n$, this is incomparably easier than the full svd of $\eta$ required when computing prox-mapping.

B: Spectahedron $X = \{ x \in S^n : x \geq 0, \text{Tr}(x) = 1 \}$. Here computing $x[\eta]$ reduces to finding the leading eigenvector of $-\eta$, i.e., to solving the problem

$$\min_{\| u \|_2 = 1} u^T \eta u.$$  

For large $n$, this is incomparably easier than the full eigenvalue decomposition of $\eta$ required when computing prox-mapping.

C: Unit TV-ball $X = \{ x \in M_0^{m,n} : \text{TV}(x) \leq 1 \}$: For $\eta \in M_0^{m,n}$, a point $x[\eta] \in \text{Argmin}_{x \in X} \text{Tr}(\eta x^T)$ is readily given by the optimal Lagrange multipliers for the capacitated network flow problem

$$\max_{t,f} \{ t : \Gamma f = t\eta, \| f \|_\infty \leq 1 \}$$

$\Gamma$: incidence matrix of the network with nodes $(i,j)$, $1 \leq i \leq m$, $1 \leq j \leq n$, and arcs $(i,j) \rightarrow (i+1,j)$, $(i,j) \rightarrow (i,j+1)$.
A: CPU ratio “full svd” / “finding leading singular vectors” for $n \times n$ matrix vs. $n$

<table>
<thead>
<tr>
<th>$n$</th>
<th>1024</th>
<th>2048</th>
<th>4096</th>
<th>8192</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU ratio</td>
<td>0.5</td>
<td>2.6</td>
<td>4.5</td>
<td>7.5</td>
</tr>
</tbody>
</table>

Full svd for $n = 8192$ takes 475.6 sec!

B: CPU ratio “full evd” / “finding leading eigenvector” for $n \times n$ symmetric matrix vs. $n$

<table>
<thead>
<tr>
<th>$n$</th>
<th>1024</th>
<th>2048</th>
<th>4096</th>
<th>8192</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU ratio</td>
<td>2.0</td>
<td>4.1</td>
<td>7.9</td>
<td>13.0</td>
</tr>
</tbody>
</table>

Full evd for $n = 8192$ takes 142.1 sec!

C: CPU ratio “metric projection” / “LMO computation” for TV ball in $M_0^{n,n}$ vs. $n$

<table>
<thead>
<tr>
<th>$n$</th>
<th>129</th>
<th>256</th>
<th>512</th>
<th>1024</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU ratio</td>
<td>10.8</td>
<td>8.8</td>
<td>11.3</td>
<td>20.6</td>
</tr>
</tbody>
</table>

Metric projection onto TV ball for $n = 1024$ takes 1062.1 sec!

Platform: 2 x 3.40 GHz CPU, 16.0 GB RAM, 64-bit Windows 7
Conditional Gradient Algorithm

$$\text{Opt} = \min_{x \in X} f(x)$$

[$$X \subset E$$: convex compact set $$f : X \rightarrow \mathbb{R}$$: convex]

(W.l.o.g. we assume that $$X$$ linearly spans the embedding Euclidean space $$E$$.

When $$X$$ is given by Linear Minimization oracle and $$f$$ is smooth, ($$CM$$) can be solved by Conditional Gradient (CndG), a.k.a. Frank-Wolfe, algorithm given by the recurrence

$$x_1 \in X, \quad x_{t+1} \in X : f(x_{t+1}) \leq f \left( x_t + \frac{2}{t+1}(x_t^+ - x_t) \right),$$

$$\begin{aligned}
 x_t^+ &= x[\nabla f(x_t)] \in \text{Argmin}_{y \in X} \langle \nabla f(x), y \rangle \\
 f_t^* &= \max_{\tau \leq t} [f(x_\tau) + \langle \nabla f(x_\tau), x_\tau^+ - x_\tau \rangle] \leq \text{Opt}
\end{aligned}$$

Theorem: Let $$f : X \rightarrow \mathbb{R}$$ be convex and $$(\kappa, L)$$-smooth:

$$\forall x, y \in X : f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{\kappa} \|x - y\|_X^\kappa$$

[$$\bullet \quad L < \infty, \, \kappa \in (1, 2]: \text{parameters}$$

$$\bullet \quad \|\cdot\|_X: \text{norm with the unit ball } \frac{1}{2}[X - X]$$

When solving ($$CP$$) by CndG, one has for $$t = 2, 3, ...$$

$$f(x_t) - \text{Opt} \leq f(x_t) - f_t^* \leq \frac{2^\kappa}{\kappa(3 - \kappa)} \cdot \frac{L}{(t + 1)^{\kappa-1}}$$
∀x, y ∈ X : f(y) ≤ f(x) + ⟨∇f(x), y − x⟩ + \frac{L}{κ}∥x − y∥^κ

[ • L < ∞, κ ∈ (1, 2]: parameters ]

Note: When X is convex, a sufficient condition for (!) is Hölder continuity of ∇f(x):
∥∇f(x) − ∇f(y)∥_* ≤ L∥x − y∥^{κ−1} ∀x, y ∈ X

For convex f and κ = 2, this condition is also necessary for (!).
Example: Minimization over a Box

Typically, the CndG rate of convergence $O(1/T^{\kappa-1})$ is not the best we can hope for. For example, when $\kappa = 2$ and $X$ is either

- the unit $\| \cdot \|_p$ ball in $\mathbb{R}^n$ with $p = 1$ or $p = 2$ (in fact, with $1 \leq p \leq 2$), or
- the unit nuclear norm ball in $\mathbb{R}^{n \times n}$,

Nesterov’s Fast Gradient method converges at the rate $O(1) \ln(n + 1)L^2/t^2$, and CndG only at the rate $O(1)L/t$. In fact,

In Favorable Geometry case, the only, if any, disadvantage of proximal algorithms as compared to CndG is the necessity to compute prox mappings, which could be expensive for problems with Complex Geometry.
Beyond the case of Favorable Geometry, CndG can be optimal.

Fact: Let $X$ be $n$-dimensional box:

\[ X = \{ x \in \mathbb{R}^n : \| x \|_\infty \leq 1 \} . \]

Then for every $t \leq n$, $L < \infty$, $\kappa \in (1, 2]$, and every utilizing local oracle $t$-step method $B$ for minimizing $(\kappa, L)$-smooth convex functions over $X$ there exists a function $f$ in the family such that for the approximate minimizer $x_B$ of $f$ generated by $B$ it holds

\[ f(x_B) - \min_x f \geq \frac{O(1)}{\ln(n)} \frac{L}{t^{\kappa-1}} \]

⇒ When minimizing smooth convex functions, represented by a local oracle, over an $n$-dimensional box, $t$-step CndG cannot be accelerated by more than $O(\ln(n))$ factor, provided $t \leq n$.

• The result remains true when replacing $n$-dimensional box $X$ with its matrix analogy

\[ \{ x \in \mathbb{R}^{n \times n} : \text{spectral norm of } x \text{ is } \leq 1 \} \]

• When minimizing $(\kappa, L)$-smooth functions over $n$-dimensional $\| \cdot \|_p$-balls with $2 \leq p \leq \infty$, the rate-of-convergence advantages of proximal algorithms over CndG rapidly deteriorate as $p$ grows and disappears (up to $O(\ln(n))$-factor) when $p$ becomes as large as $O(\ln(n))$. 

6.129
Proof of Theorem

(a) \( f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{\kappa} \| y - x \|_X \)

(b) \( f(x_{t+1}) \leq f(x_t + \gamma_t (x_t^+ - x_t)) \),

\[ \gamma_t = \frac{2}{t+1}, \quad x_t^+ \in \text{Argmin}_{y \in X} \langle \nabla f(x_t), y \rangle \]

\[ f^*_t := \max_{\tau \leq t} [f(x_\tau) + \langle \nabla f(x_\tau), x_\tau^+ - x_\tau \rangle] \]

? \( \Rightarrow ? \)

\( f(x_t) - f^*_t \leq \frac{2^t L}{\kappa (3-\kappa)} \gamma_t^{\kappa-1} \) \( (\ast_t), t \geq 2 \)

Let

\( \epsilon_t = f(x_t) - f^*_t, \quad e_t = \langle \nabla f(x_t), x_t - x_t^+ \rangle \)

\( f^*_t \geq f(x_t) + \langle \nabla f(x_t), x_t^+ - x_t \rangle \Rightarrow e_t \geq \epsilon_t \)

We have

(c) \( \| x_t - x_t^+ \|_X \leq 2 \)

\( \Rightarrow f(x_{t+1}) \leq f(x_t + \gamma_t (x_t^+ - x_t)) \) \( [\text{by (b)}] \)

\[ \leq f(x_t) + \gamma_t \langle \nabla f(x_t), x_t^+ - x_t \rangle + \frac{L}{\kappa} [2 \gamma_t]^{\kappa} \]

\[ = f(x_t) - \gamma_t e_t + \frac{2^t L}{\kappa} \gamma_t^{\kappa} \]

\[ \leq f(x_t) - \gamma_t e_t + \frac{2^t L}{\kappa} \gamma_t^{\kappa} \] \( [\text{since } e_t \geq \epsilon_t] \)

\( \Rightarrow \epsilon_{t+1} = f(x_{t+1}) - f^*_{t+1} \leq f(x_{t+1}) - f^*_t \)

\[ \leq \epsilon_t (1 - \gamma_t) + \frac{2^t L}{\kappa} \gamma_t^{\kappa} \] \( [\text{since } f^*_{t+1} \geq f^*_t] \)

6.130
\[0 \leq \epsilon_{t+1} \leq \epsilon_t (1 - \gamma_t) + \frac{2^\kappa L}{\kappa} \gamma_t^{k-1} \quad (*)_t\]

? \Rightarrow? \quad \epsilon_t \leq \frac{2^{k+1} L}{\kappa (3-\kappa)} \gamma_t^{k-1}, \quad t \geq 2 \quad [\gamma_t = \frac{2}{t+1}] \quad (!_t)

- By \((*)_2\), we have \(\epsilon_2 \leq \frac{2^\kappa L}{\kappa} \Rightarrow \epsilon_2 \leq \frac{2^{k+1} L}{\kappa (3-\kappa)} (2/3)^{k-1}\) due to \(1 < \kappa \leq 2 \Rightarrow (!_2)\) holds true.
- Assuming \(!_t\) true for some \(t \geq 2\), we have

\[
\epsilon_{t+1} \leq \frac{2^{k+1} L}{\kappa (3-\kappa)} \gamma_t^{k-1} (1 - \gamma_t) + \frac{2^\kappa L}{\kappa} \gamma_t^{k} \quad [\text{by } (*)_t \text{ and } (!_t)]
\]

\[
= \frac{2^{k+1} L}{\kappa (3-\kappa)} \left[ \gamma_t^{k-1} - \frac{\kappa-1}{2} \gamma_t^k \right]
\]

\[
= \frac{2^{k+1} L}{\kappa (3-\kappa)} 2^{k-1} \left[ (t+1)^{1-\kappa} + (1 - \kappa)(t+1)^{-\kappa} \right]
\]

\[
\leq \frac{2^{k+1} L}{\kappa (3-\kappa)} 2^{k-1} (t+2)^{1-\kappa} \quad [\text{by convexity of } (t+1)^{1-\kappa}]
\]

\[
= \frac{2^{k+1} L}{\kappa (3-\kappa)} \gamma_t^{k-1} \Rightarrow (!_{t+1}) \text{ holds true.}
\]

Thus, \(!_t\) holds true for all \(t\), Q.E.D.
Conditional Gradient Algorithm for Norm-regularized Smooth Convex Minimization

♣ “As is”, CndG is applicable only to minimizing smooth convex functions on bounded and closed convex domains.

**Question:** How to apply CndG to Composite Minimization problem

\[
\text{Opt} = \min_{x \in K} \{\lambda \|x\| + f(x)\}
\]

- \(K\): closed convex cone in Euclidean space \(E\)
- \(\|\cdot\|\): norm on \(E\)
- \(\lambda > 0\): penalty
- \(f : K \to \mathbb{R}\): convex function with Lipschitz continuous gradient:

\[
\|\nabla f(x) - \nabla f(y)\|_* \leq L_f \|x - y\|, \ x, y \in K
\]

♠ **Main Assumption:** We have at our disposal LMO oracle for \((\|\cdot\|, K)\). Given on input a linear form \(\langle \eta, \cdot \rangle\) on \(E\), the oracle returns

\[x[\eta] \in \text{Argmin}_x \{\langle \eta, x \rangle : x \in K, \|x\| \leq 1\}\]

**Examples:**

- **A.** \(E = \mathbb{R}^{m \times n}, \|\cdot\| = \|\cdot\|_{\text{nuc}}, K = E\)
- **B.** \(E = S^n, \|\cdot\| = \|\cdot\|_{\text{nuc}}, K = S^n_+ = \{x \in E : x \succeq 0\}\)
- **C.** \(E = M_0^{m,n}, \|\cdot\| = \text{TV}(\cdot), K = E\).

6.132
We can reformulate the problem of interest as

\[
\text{Opt} = \min_{[x;r] \in K^+} \{ \phi(x, r) := \lambda r + f(x) \}
\]

\[K^+ = \{ [x; r] \in E^+ := E \times \mathbb{R} : \|x\| \leq r \}\]

**Assumption:** There exists \( D_* < \infty \) such that

\[y := [x; r] \in K^+ \& r > D_* \Rightarrow \phi(y) > \phi(0),\]

and we are given a finite upper bound \( D^+ \) on \( D_* \).

**Note:** The efficiency estimate for the forthcoming method depends on \( D_* \), and not on \( D^+ \)!

**Algorithm:**

- **Initialization:** Set \( y_1 = 0 \in K^+ \)
- **Step** \( t = 1, 2, \ldots \) Given \( y_t = [x_t; r_t] \in K^+ \),
  - compute \( \nabla f(x_t) \)
  - compute
    \[
    x_t^+ = x[\nabla f(x_t)] \\
    \in \text{Argmin}_x \{ \langle \nabla f(x_t), x \rangle : x \in K, \|x\| \leq 1 \}
    \]
  - set \( \Delta_t = \text{Conv} \{ y_t, 0, D^+[x_t^+; 1] \} \subset K^+ \) and find
    \[y_{t+1} \in K^+ : \phi(y_{t+1}) \leq \min_{y \in \Delta_t} \phi(y)\]

Step \( t \) is completed; pass to step \( t + 1 \).

**Note:** One can set \( y_{t+1} \in \text{Argmin}_y \phi(y) \). With this policy, a step requires minimizing \( \phi \)
over a \textit{2D triangle} \( \Delta_t \), which can be done within machine precision in \( O(1) \) steps (e.g.,
by the Ellipsoid method).
\[ \text{Opt} = \min_{[x; r] \in K^+} \{ \phi(x, r) := \lambda r + f(x) \} \]
\[ K^+ = \{ [x; r] \in E^+ := E \times \mathbb{R} : x \in K, \| x \| \leq r \} \]

\[ \blacklozenge \text{ Theorem: For the outlined algorithm,} \]
\[ \phi(y_t) - \text{Opt} \leq \frac{8L_fD_*^2}{t + 14}, \quad t = 2, 3, \ldots \]

\[ \spadesuit \text{ Bundle Implementation: We can set} \]
\[ y_{t+1} \in \text{Argmin}_y \{ \phi(y) : y \in \text{Conv}\{0 \cup Y_t\} \} \quad (\ast) \]
\[ Y_t \subset K^+: \text{finite set containing } y_t = [x_t; r_t] \text{ and } D^+[x_t^+; 1], \text{ with} \]
\[ x_t^+ \in \text{Argmin}_x \{ \langle \nabla f(x_t), x \rangle : x \in K, \| x \| \leq 1 \} \]

For example, we can comprise \( Y_t \) of \( y_t, D^+[x_t^+; 1] \) and several of the previous iterates \( y_1, \ldots, y_{t-1} \).

\[ \heartsuit \text{ Bundle approach is especially attractive when} \]
\[ f(x) = \Psi(Ax + b) \]

for easy to compute \( \Psi, \) like \( \Psi(u) = \frac{1}{2}u^Tu. \) Here computing \( f, \nabla f \) at a convex (or linear) combination \( x = \sum \lambda_i x_i \) of points \( x_i \) with already computed \( Ax_i \) becomes cheap: \( Ax = \sum \lambda_i (Ax_i). \)

\[ \Rightarrow \text{ the FO oracle for (}\ast\text{) is computationally cheap} \]
\[ y_{t+1} \in \text{Argmin}_y \{ \phi(y) : y \in \text{Conv}\{0 \cup Y_t\} \} \quad (\ast) \]

\( Y_t \subset K^+ \): finite set containing \( y_t = [x_t; r_t] \) and \( D^+[x_t^+; 1] \), with
\[ x_t^+ \in \text{Argmin}_x \{ \langle \nabla f(x_t), x \rangle : x \in K, \|x\| \leq 1 \} \]

- For example, with \( f(x) = \frac{1}{2}\|Ax - b\|_2^2 \), solving \((\ast)\) reduces to solving \( k_t = \text{Card}(Y_t)\)-dimensional convex quadratic problem

\[
\min_{\lambda \in \mathbb{R}^{k_t}} \left\{ \frac{1}{2} \lambda^T Q_t \lambda + 2 q_t^T \lambda : \lambda \geq 0, \sum_j \lambda_j \leq 1 \right\},
\]

\[ Q_t = [x_i^T A^T A x_j]_{i,j} \]

where \( x_j, 1 \leq j \leq k_t \), are the \( x \)-components of the points from \( Y_t \).

⇒ Assuming that \( Y_t \) is a set of moderate cardinality (say, few tens) obtained from \( Y_{t-1} \) by discarding several “old” points and adding the new points \( y_t = [x_t; r_t], D^+[x_t^+; 1] \), updating

\[
[Q_{t-1}, q_{t-1}] \mapsto [Q_t, q_t]
\]

basically reduces to computing matrix-vector products \( Ax_t \) and \( Ax_t^+ \). After \( Q_t, q_t \) are computed, \((!\) can be solved “in no time” by an IPM.

**Note:** \( Ax_t \) is computed anyway when computing \( \nabla f(x_t) \).
How It Works: TV-based Image Reconstruction

Bundle CndG, 256 × 256 image (65,536 variables)
Recovery in 13 CndG iterations, CPU time 50.0 sec
Error removal: 98.5%, $\phi(y_{13})/\phi(0) < 4.6e-5$

Bundle CndG, 512 × 512 image (262,144 variables)
Recovery in 18 CndG iterations, CPU time 370.3 sec
Error removal: 98.2%, $\phi(y_{18})/\phi(0) < 1.3e-4$

**Platform:** 2 × 3.40 GHz CPU with 16.0 GB RAM and 64-bit operating system
♠ Note: We used 15-element bundle, adding to it at step $t$ the points $y_t = [x_t; r_t], D^+[x^+_t; 1]$ and $[\nabla f(x_t); TV(\nabla f(x_t))]$ and removing (up to) 3 old points according to “first in — first out.” Adding $[\nabla f(x_t); TV(\nabla f(x_t))]$ to the bundle dramatically accelerated the algorithm.
How It Works:
Low Rank Matrix Completion

♠ Problem:

\[
\text{Opt} = \min_{x \in \mathbb{R}^{n \times n}} \left\{ 0.1 \|x\| + \|x - a\|_F^2 \right\}
\]

- \(\|\cdot\|\): nuclear norm
- \(\|\cdot\|_F\): Frobenius norm
- \(a = \bar{x} + \xi\)

\[
\text{Rank}(\bar{x}) \approx \sqrt{n}, \|\bar{x}\| \approx \sqrt{2n/\pi}, \|\xi\|_F \approx 0.1\|\bar{x}\|_F \text{ with i.i.d. Gaussian } \xi_{ij}
\]

• Required relative inaccuracy 0.01

<table>
<thead>
<tr>
<th>(n)</th>
<th>Method</th>
<th>CPU, sec</th>
<th>Iterations</th>
<th>Relative inaccuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>128</td>
<td>CndG</td>
<td>4.5</td>
<td>42</td>
<td>&lt;1.3e-6</td>
</tr>
<tr>
<td></td>
<td>IPM</td>
<td>2675.0</td>
<td>31</td>
<td>&lt;1.e-10</td>
</tr>
<tr>
<td>1024</td>
<td>CndG</td>
<td>44.2</td>
<td>31</td>
<td>&lt;0.008</td>
</tr>
<tr>
<td></td>
<td>IPM</td>
<td>not tested</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4096</td>
<td>CndG</td>
<td>1997.7</td>
<td>87</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td></td>
<td>IPM</td>
<td>not tested</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8192</td>
<td>CndG</td>
<td>1364.5</td>
<td>36</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td></td>
<td>IPM</td>
<td>not tested</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

† \(\text{Rank}(\bar{x}) = 32\)

Platform: 2 \times 3.40 GHz CPU with 16.0 GB RAM and 64-bit operating system

Note: CPU time in 8192 \times 8192 example is less than needed to compute just 3 full svd's of a 8192 \times 8192 matrix ⇒ The time taken by 36 steps of CndG is less than needed to perform just 3 steps of the simplest proximal algorithm, or just 2 steps of Nesterov's Fast Gradient method for Composite minimization!