Convex Analysis, Duality and Optimization

Yao-Liang Yu
yaoliang@cs.ualberta.ca
Dept. of Computing Science
University of Alberta

March 7, 2010

Outline

Notations Used Throughout

- $\mathcal{C}$ for convex set, $\mathcal{S}$ for arbitrary set, $\mathcal{K}$ for convex cone,
- $g(\cdot)$ is for arbitrary functions, *not* necessarily convex,
- $f(\cdot)$ is for convex functions, for simplicity, we assume $f(\cdot)$ is closed, proper, continuous, and differentiable when needed,
- min (max) means inf (sup) when needed,
- w.r.t.: with respect to; w.l.o.g.: without loss of generality; u.s.c.: upper semi-continuous; l.s.c.: lower semi-continuous; int: interior point; RHS: right hand side; w.p.1: with probability 1.

Historical Note

- 60s: Linear Programming, Simplex Method
- 70s-80s: (Convex) Nonlinear Programming, Ellipsoid Method, Interior-Point Method
- 90s: Convexification *almost everywhere*
- Now: Large-scale optimization, First-order gradient method

But...

Neither of poly-time solvability and convexity implies the other.

NP-Hard **convex** problems abound.

Outline

Convex Sets and Functions

Definition (Convex set)

A point set $\mathcal{C}$ is said to be convex if $\forall\ \lambda \in [0,1], x, y \in \mathcal{C}$, we have $\lambda x + (1-\lambda) y \in \mathcal{C}$.

Definition (Convex function)

A function $f(\cdot)$ is said to be convex if

1. $\mathrm{dom} f$ is convex, and,
2. $\forall\ \lambda \in [0,1], x, y \in \mathrm{dom} f$, we have $f(\lambda x + (1-\lambda) y) \leq \lambda f(x) + (1-\lambda) f(y)$;

Or equivalently, $f(\cdot)$ is convex if its epigraph $\left\{ \binom{x}{t} : f(x) \leq t \right\}$ is a convex set.

- Function $h(\cdot)$ is concave iff $-h(\cdot)$ is convex,
- $h(\cdot)$ is called affine (linear) iff it's both convex and concave,
- No concave set. Affine set: drop the constraint on λ.

More on Convex functions

Definition (Strongly Convex Function)

$f(x)$ is said to be μ-strongly convex with respect to a norm $\|\cdot\|$ iff $\mathrm{dom}\ f$ is convex and $\forall \lambda \in [0,1]$,

$$f(\lambda x + (1-\lambda) y) + \mu \cdot \frac{\lambda(1-\lambda)}{2}\|x-y\|^2 \leq \lambda f(x) + (1-\lambda) f(y).$$

Proposition (Sufficient Conditions for μ-Strong Convexity)

1. *Zero Order: definition*
2. *First Order:* $\forall x, y \in \mathrm{dom}\ f$,

$$f(y) \geq f(x) + \langle \nabla f(x), x - y \rangle + \frac{\mu}{2}\|x-y\|^2.$$

3. *Second Order:* $\forall x, y \in \mathrm{dom}\ f$,

$$\langle \nabla^2 f(x) y, y \rangle \geq \mu \|y\|^2.$$

Elementary Convex Functions (to name a few)

- Negative entropy $x \log x$ is convex on $x > 0$,
- ℓ_p-norm $\|x\|_p := \left[\sum_i |x_i|^p\right]^{1/p}$ is convex when $p \geq 1$, concave otherwise (except $p = 0$),
- Log-sum-exp function $\log \sum_i \exp(x_i)$ is convex, same is true for the matrix version $\log \mathrm{Tr} \exp(X)$ on symmetric matrices,
- Quadratic-over-linear function $x^T Y^{-1} x$ is *jointly* convex in x and $Y \succ 0$, what if $Y \succeq 0$?
- Log-determinant $\log \det X$ is concave on $X \succ 0$, what about $\log \det X^{-1}$?
- $\mathrm{Tr}\ X^{-1}$ is convex on $X \succ 0$,
- The largest element $x_{[1]} = \max_i x_i$ is convex; moreover, sum of largest k elements is convex; what about smallest analogies?
- The largest eigenvalue of *symmetric* matrices is convex; moreover, sum of largest k eigenvalues of *symmetric* matrices is also convex; can we drop the condition *symmetric*?

Compositions

Proposition (Affine Transform)

$A\mathcal{C} := \{Ax : x \in \mathcal{C}\}$ and $A^{-1}\mathcal{C} := \{x : Ax \in \mathcal{C}\}$ are convex sets. Similarly, $(Af)(x) := \min_{Ay=x} f(y)$ and $(fA)(x) := f(Ax)$ are convex.

Proposition (Sufficient but NOT Necessary)

$f \circ g$ is convex if

- *$g(\cdot)$ is convex and $f(\cdot)$ is non-decreasing, or*
- *$g(\cdot)$ is concave and $f(\cdot)$ is non-increasing.*

Proof.

For simplicity, assume $f \circ g$ is twice differentiable. Use the second-order sufficient condition. □

Remark: One needs to check if $\mathrm{dom} f \circ g$ is convex! However, this is unnecessary if we consider *extended-value* functions.

Operators Preserving Convexity

Proposition (Algebraic)

For $\theta > 0$, $\lambda\mathcal{C} := \{\theta x : x \in \mathcal{C}\}$ *is convex;* $\theta f(x)$ *is convex; and* $f_1(x) + f_2(x)$ *is convex.*

Proposition (Intersection v.s. Supremum)

- *Intersection of arbitrary collection of convex sets is convex;*
- *Similarly, pointwise supremum of arbitrary collection of convex functions is convex.*

Proposition (Sum v.s. Infimal Convolution)

- $\mathcal{C}_1 + \mathcal{C}_2 := \{x_1 + x_2 : x_i \in \mathcal{C}_i\}$ *is convex;*
- *Similarly,* $(f_1 \square f_2)(x) := \inf_y\{f_1(y) + f_2(x - y)\}$ *is convex.*

Proof.

Consider affine transform. □

What about union v.s. infimum? Needs extra convexification.

Convex Hull

Definition (Convex Hull)

The convex hull of $\mathcal{S}$, denoted $\mathrm{conv}\mathcal{S}$, is the smallest convex set containing $\mathcal{S}$, i.e. the intersection of all convex sets containing $\mathcal{S}$.

Similarly, the convex hull of $g(x)$, denoted $\mathrm{conv}g$, is the greatest convex function dominated by g, i.e. the pointwise supremum of all convex functions dominated by g.

Theorem (Carathéodory, 1911)

The convex hull of any set $\mathcal{S} \in \mathbb{R}^n$ is:

$$\{x : x = \sum_{i=1}^{n+1} \lambda_i x_i, x_i \in \mathcal{S}, \lambda_i \geq 0, \sum_{i=1}^{n+1} \lambda_i = 1\}.$$

We will see how to compute $\mathrm{conv}g$ later.

Cones and Conic Hull

Definition (Cone and Positively Homogeneous Function)

A set $\mathcal{S}$ is called a cone if $\forall x \in \mathcal{S}, \theta \geq 0$, we have $\theta x \in \mathcal{S}$.
Similarly, a function $g(x)$ is called positively homogeneous if $\forall \theta \geq 0, g(\theta x) = \theta g(x)$.

$\mathcal{K}$ is a convex cone if it is a cone and is convex, specifically, if $\forall x_1, x_2 \in \mathcal{K}, \theta_1, \theta_2 \geq 0, \Rightarrow \theta_1 x_1 + \theta_2 x_2 \in \mathcal{K}$.
Similarly, $f(x)$ is positively homogeneous convex if it is positively homogeneous and convex, specifically, if $\forall x_1, x_2 \in \mathrm{dom} f, \theta_1, \theta_2 \geq 0, \Rightarrow f(\theta_1 x_1 + \theta_2 x_2) \leq \theta_1 f(x_1) + \theta_2 f(x_2)$.

Remark: *Under the above definitions, cones always contain the origin, and positively homogeneous functions equal 0 at the origin.*

Definition (Conic Hull)

The conic hull of $\mathcal{S}$ is the smallest convex cone containing $\mathcal{S}$.
Similarly, the conic hull of $g(x)$, denoted $\mathrm{cone} g$, is the greatest positively homogeneous convex function dominated by g.

Conic Hull cont'

Theorem (Carathéodory, 1911)

The conic hull of any set $\mathcal{S} \in \mathbb{R}^n$ *is:*

$$\{x : x = \sum_{i=1}^{n} \theta_i x_i, x_i \in \mathcal{S}, \theta_i \geq 0, \}.$$

For *convex* function $f(x)$, its conic hull is:

$$(\mathrm{cone} f)(x) = \min_{\theta \geq 0} \theta \cdot f(\theta^{-1} x).$$

How to compute $\mathrm{cone} g$? Hint: $\mathrm{cone} g = \mathrm{cone}\ \mathrm{conv} g$, why?

Elementary Convex Sets (to name a few)

- Hyperplane $a^T x = \alpha$ is convex,
- Half space $a^T x \leq \alpha$ is convex,
- Affine set $Ax = b$ is convex (proof?),
- Polyhedra set $Ax \leq b$ is convex (proof?),

Proposition (Level sets)

(Sub)level sets of $f(x)$, defined as $\{x : f(x) \leq \alpha\}$ are convex.

Proof.

Consider the intersection of the epigraph of $f(x)$ and the hyperplane $\binom{\mathbf{0}}{1}^T \binom{x}{t} = \alpha$. □

Remark: *A function, with all level sets being convex, is not necessarily convex! We call such functions, with convex domain, quasi-convex.*

Convince yourself the ℓ_0-norm, defined as $\|x\|_0 = \sum_i \mathbb{I}[x_i \neq 0]$, is not convex. Show that -$\|x\|_0$ on $x \geq 0$ is quasi-convex.

Elementary Convex Sets cont'

- Ellipsoid $\{x : (x - x_c)^T P^{-1}(x - x_c) \leq 1, P \succ 0\}$ or $\{x_c + P^{1/2}u : \|u\|_2 \leq 1\}$ is convex,
- Nonnegative orthant $x \geq 0$ is a convex cone,
- All positive (semi)definite matrices compose a convex cone (positive (semi)definite cone) $X \succ 0$ $(X \succeq 0)$,
- All norm cones $\{\binom{x}{t} : \|x\| \leq t\}$ are convex, in particular, for the Euclidean norm, the cone is called second order cone or Lorentz cone or ice-cream cone.

  Remark: This is essentially saying that all norms are convex. ℓ_0-norm is not convex? No, but it's not a "norm" either. People call it "norm" unjustly.

Outline

Unconstrained

Consider the simple problem:

$$\min_{x} \ f(x), \tag{1}$$

where $f(\cdot)$ is defined in the whole space.

Theorem (First-order Optimality Condition)

A sufficient and necessary condition for $x^\star$ to be the minimizer of (1) is:

$$0 \in \partial f(x^\star). \tag{2}$$

When $f(\cdot)$ is differentiable, (2) reduces to $\nabla f(x^\star) = 0$.

Remark:

- The minimizer is unique when $f(\cdot)$ is strictly convex,
- For general nonconvex functions $g(\cdot)$, the condition in (2) gives only critical (stationary) points, which could be minimizer, maximizer, or nothing (saddle-point).

Simply Constrained

Consider the constrained problem:

$$\min_{x \in \mathcal{C}} \; f(x), \tag{3}$$

where $f(\cdot)$ is defined in the whole space.
Is $\nabla f(x^\star) = 0$ still the optimality condition? If you think so, consider the example:

$$\min_{x \in [1,2]} \; x.$$

Theorem (First-order Optimality Condition)

A sufficient and necessary condition for $x^\star$ to be the minimizer of (3) is (assuming differentiability):

$$\forall x \in \mathcal{C}, \quad (x - x^\star)^T \nabla f(x^\star) \geq 0. \tag{4}$$

Verify this condition is indeed satisfied by the example above.

General Convex Program

We say a problem is convex if it is of the following form:

$$\begin{aligned} \min_{x \in \mathcal{C}} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, i = 1, \ldots, m \\ & Ax = b, \end{aligned}$$

where $f_i(x), i = 0, \ldots, m$ are all convex.

Remark:

- The equality constraint must be affine! But affine functions are free to appear in inequality constraints.
- The objective function, being convex, is to be minimized. Sometimes we see maximizing a concave function, no difference (why?).
- The inequality constraints are $\leq$, which lead to a convex feasible region (why?).
- To summarize, convex programs are to minimize a convex function over a convex feasible region.

Two Optimization Strategies

Usually, unconstrained problems are easier to handle than constrained ones, and there are two typical ways to convert constrained problems into unconstrained ones.

Example (Barrier Method)

Given the convex program, determine the feasible region (needs to be compact), and then construct a barrier function, say $b(x)$, which is convex and quickly grows to ∞ when x, the decision variable, approaches the boundary of the feasible region. Consider the following composite problem:

$$\min_{x} \quad f(x) + \lambda \cdot b(x).$$

If we initialize with an interior point of the feasible region, we will stay within the feasible region (why?). Now minimizing the composite function and gradually decreasing the parameter λ to 0. The so-called interior-point method in each iteration takes a Newton step w.r.t. x and then updates λ in a clever way.

Two Optimization Strategies cont'

Example (Penalty Method)

While the barrier method enforces feasibility in each step, the penalty method penalizes the solver if any equality constraint is violated, hence we first convert any inequality constraint $f_i(x) \leq 0$ to an equality one by the trick $\Big[h(x) := \max\{f_i(x), 0\}\Big] = 0$ (convex?). Then consider, similarly, the composite problem:

$$\min_{x} \quad f(x) + \lambda \cdot h(x).$$

Now minimizing the composite function and gradually increasing the parameter λ to ∞. Note that the max function is not smooth, usually one could square the function $h(\cdot)$ to get some smoothness.

Remark: The bigger λ is, the harder the composite problem is, so we start with a gentle λ, gradually increase it while using the x we got from previous λ as our initialization, the so-called "warm-start" trick. How about the λ in the barrier method?

Linear Programming (LP)

Standard Form

$$\begin{aligned} \min_x \quad & c^T x \\ \text{s.t.} \quad & x \geq 0 \\ & Ax = b \end{aligned}$$

General Form

$$\begin{aligned} \min_x \quad & c^T x \\ \text{s.t.} \quad & Bx \leq d \\ & Ax = b \end{aligned}$$

Example (Piecewise-linear Minimization)

$$\min_x \; f(x) := \max_i a_i^T x + b_i$$

This does not look like an LP, but can be equivalently reformulated as one:

$$\min_{x,t} \; t \quad \text{s.t.} \quad a_i^T x + b_i \leq t, \forall i.$$

Remark: Important trick, learn it!

Quadratic Programming (QP)

Standard Form

$$\min_x \quad \frac{1}{2}x^T P x + q^T x + r$$
$$\text{s.t.} \quad x \geq 0$$
$$Ax = b$$

General Form

$$\min_x \quad \frac{1}{2}x^T P x + q^T x + r$$
$$\text{s.t.} \quad Bx \leq d$$
$$Ax = b$$

Remark: P must be positive semidefinite! Otherwise the problem is non-convex, and in fact NP-Hard.

Example (LASSO)

$$\min_w \quad \frac{1}{2}\|Aw - b\|_2^2 + \lambda\|w\|_1$$

Example (Compressed Sensing)

$$\min_w \quad \frac{1}{2}\|Aw - b\|_2^2 \quad \text{s.t.} \quad \|w\|_1 \leq C$$

Reformulate them as QPs (but never solve them as QPs!).

More QP Examples

Example (Support Vector Machines)

$$\min_{w,b} \quad \sum_i \Big[y_i(w^T x_i + b) - 1\Big]_+ + \frac{\lambda}{2}\|w\|_2^2$$

Reformulate it as a QP.

Example (Fitting data with Convex functions)

$$\min_f \quad \frac{1}{2}\sum_i [f(x_i) - y_i]^2 \quad \text{s.t.} \quad f(\cdot) \text{ is convex}$$

Using convexity, one can show that the optimal $f(\cdot)$ has the form:

$$f(x) = \max_i y_i + g_i^T(x - x_i).$$

Turn the functional optimization problem into finite dimensional optimization w.r.t. g_i's. Show that it is indeed a QP.
Fitting with *monotone* convex functions? Overfitting issues?

Quadratically Constrained Quadratic Programming (QCQP)

General Form

$$
\begin{aligned}
\min_{x} \quad & \frac{1}{2} x^T P_0 x + q_0^T x + r_0 \\
\text{s.t.} \quad & \frac{1}{2} x^T P_i x + q_i^T x + r_i \leq 0, i = 1, \ldots, m \\
& Ax = b
\end{aligned}
$$

Remark: $P_i, i = 0, \ldots, m$ must be positive semidefinite! Otherwise the problem is non-convex, and in fact NP-Hard.

Example (Euclidean Projection)

$$
\min_{\|x\|_2 \leq 1} \quad \frac{1}{2} \|x - x_0\|_2^2
$$

We will use Lagrangian duality to solve this trivial problem.

Second Order Cone Programming (SOCP)

Standard Form

$$\begin{aligned} \min_x \quad & c^T x \\ \text{s.t.} \quad & \|B_i x + d_i\|_2 \leq f_i^T x + \gamma_i, i = 1, \ldots, m \\ & Ax = b \end{aligned}$$

Remark: It's the ℓ_2-norm, not *squared*, in the inequality constraints (otherwise the problem is a ?).

Example (Chance Constrained Linear Programming)

Oftentimes, our data is corrupted by noise and we might want a probabilistic (v.s. deterministic) guarantee:

$$\min_x \quad c^T x \quad \text{s.t.} \quad \mathbb{P}_{a_i}(a_i^T x \leq 0) \geq 1 - \epsilon$$

Assume a_i's follow the normal distribution with known mean $\bar{a}_i$ and covariance matrix Σ_i, can reformulate the problem as an SOCP:

SOCP Examples cont'

Example (CCLP cont')

$$
\begin{aligned}
\min_{x} \quad & c^T x \\
\text{s.t.} \quad & \bar{a}_i^T x + \Phi^{-1}(1-\epsilon)\|\Sigma_i^{1/2} x\|_2 \leq 0
\end{aligned}
$$

What if the distribution is not normal? Not known beforehand?

Example (Robust LP)

Another approach is to construct a robust region and optimize w.r.t. the worst-case scenario:

$$
\begin{aligned}
\min_{x} \quad & c^T x \\
\text{s.t.} \quad & \left[\max_{a_i \in \mathcal{E}_i} a_i^T x\right] \leq 0
\end{aligned}
$$

Popular choices for $\mathcal{E}_i$ are the box constraint $\|a_i\|_\infty \leq 1$ or the ellipsoid constraint $(a_i - \bar{a}_i)^T \Sigma_i^{-1} (a_i - \bar{a}_i) \leq 1$.

We will use Lagrangian duality to turn the latter case to an SOCP. How about the former case?

Semidefinite Programming (SDP)

Standard Primal Form

$$\begin{aligned} \min_x \quad & c^T x \\ \text{s.t.} \quad & \sum_i x_i F_i + G \preceq 0 \\ & Ax = b \end{aligned}$$

Standard Dual Form

$$\begin{aligned} \min_X \quad & \mathrm{Tr}(GX) \\ \text{s.t.} \quad & \mathrm{Tr}(F_i X) + c_i = 0 \\ & X \succeq 0 \end{aligned}$$

Remark: We will learn how to transform *primal* problems into *dual* problems (and vice versa) later.

Example (Largest Eigenvalue)

Let S_i's be symmetric matrices, consider

$$\min_x \quad \lambda_{\max}\left[\sum_i x_i S_i\right]$$

Reformulate:

$$\min_{x,t} t \quad \text{s.t.} \quad \sum_i x_i S_i \preceq tI$$

SDP Examples

Example (2nd Smallest Eigenvalue of Graph Laplacian)

We've seen the graph Laplacian $L(x)$. In some applications, we need to consider the following problem:

$$\max_{x \geq 0} \quad \lambda_2[L(x)],$$

where $\lambda_2(\cdot)$ means the second smallest eigenvalue. Does this problem belong to convex optimization? Reformulate it as an SDP. Hint: The smallest eigenvalue of a Laplacian matrix is always 0.

Before moving on to the next example, we need another theorem, which is interesting in its own right:

Theorem (Maximizing Convex Functions)

$$\max_{x \in S} f(x) = \max_{x \in \mathrm{conv}\, S} f(x).$$

Remark: We are talking about maximizing a convex function now!

SDP Examples cont'

Example (Yet Another Eigenvalue Example)

We know the largest eigenvalue (of a symmetric matrix) can be efficiently computed. We show that it can in fact be reformulated as an SDP (illustration only, do NOT compute eigenvalues by solving SDPs!)
The largest eigenvalue problem, mathematically, is:

$$\max_{x^T x=1} \quad x^T A x,$$

where A is assumed to be symmetric.
Use the previous cool theorem to show that the following reformulation is equivalent:

$$\max_{M \succeq 0} \quad \mathrm{Tr}(AM) \quad \text{s.t.} \quad \mathrm{Tr}(M) = 1$$

Generalization to the sum of k largest eigenvalues? Smallest ones?

NP-Hard Convex Problem

Consider the following problem:

$$\max_{x} \quad x^T A x \quad \text{s.t.} \quad x \in \Delta_n, \tag{5}$$

where $\Delta_n := \{x : x_i \geq 0, \sum_i x_i = 1\}$ is the standard simplex. (5) is known to be NP-Hard since it embodies the maximum clique problem. It is trivial to see (5) is the same as

$$\max_{X,x} \quad \mathrm{Tr}(AX) \quad \text{s.t.} \quad X = xx^T, x \in \Delta_n, \tag{6}$$

which is further equivalent to

$$\max_{X} \quad \mathrm{Tr}(AX) \quad \text{s.t.} \quad \sum_{ij} X_{ij} = 1, X \in \mathcal{K}, \tag{7}$$

where $\mathcal{K} := \mathrm{conv}\{xx^T : x \geq 0\}$ is the so-called completely positive cone. Verify by yourself $\mathcal{K}$ is indeed a convex cone.

Remark: The equivalence of (6) and (7) comes from the fact that the extreme points of their feasible regions agree, hence the identity of convex hulls.

Geometric Programming (mainly based on Ref. 5)

Notice that during this subsection, we always assume x_i's are positive.

Definition (Monomial)

We call $c \cdot x_1^{a_1} x_2^{a_2} \ldots x_n^{a_n}$ monomial when $c > 0$ and $a_i \in \mathbb{R}$.

Definition (Posynomial)

The sum (product) of finite number of monomials.

Remark: Posynomial = Positive + Polynomial.

Definition (Generalized Posynomial)

Any function formed from addition, multiplication, positive fractional power, pointwise maximum of (generalized) posynomials.

Example (Simple Instances)

- $0.5,\ x,\ x_1/x_2^3,\ \sqrt{x_1/x_2}$ are monomials;
- $(1+x_1x_2)^3,\ 2x_1^{-3}+3x_2$ are posynomials;
- $x_1^{-1.1}+(1+x_2/x_3)^{3.101},\ \max\{((x_2+1)^{1.3}+x_3^{-1})^{1.92}+x_1^{0.7},\ 2x_1+x_2^{0.9}x_3^{-3.9}\}$ are generalized posynomials;
- $-0.11,\ x_1-x_2,\ x^2+\cos(x),\ (1+x_1/x_2)^{-1.1},\ \max\{x^{0.7},-1.1\}$ are not generalized posynomials;

Generalized Geometric Programming (GGP)

Let $p_i(\cdot), i = 0, \ldots, m$ be generalized posynomials and $m_j(\cdot)$ be monomials.

Standard Form

$$\begin{aligned} \min_{x} \quad & p_0(x) \\ \text{s.t.} \quad & p_i(x) \leq 1, i = 1, \ldots, m \\ & m_j(x) = 1, j = 1, \ldots, n, \end{aligned}$$

Convex Form

$$\begin{aligned} \min_{y} \quad & \log p_0(e^y) \\ \text{s.t.} \quad & \log p_i(e^y) \leq 0, i = 1, \ldots, m \\ & \log m_j(e^y) = 0, j = 1, \ldots, n \end{aligned}$$

GPP does not look like convex in its standard form, however, using the following proposition, it can be easily turned into convex (by changing variables $x = e^y$ and applying the monotonic transform $\log(\cdot)$):

Proposition (Generalized Log-Sum-Exp)

If $p(x)$ is a generalized posynomial, then $f(y) := \log p(e^y)$ is convex. Immediately, we know $p(e^y)$ is also convex.

A Nice Trick

One can usually reduce GPPs to programs that only involve posynomials. This is best illustrated by an example. Consider, say, the constraint:

$$(1+\max\{x_1,x_2\})(1+x_1+(0.1x_1x_3^{-0.5}+x_2^{1.6}x_3^{0.4})^{1.5})^{1.7} \leq 1$$

By introducing new variables, this complicated constraint can be simplified to:

$$\begin{aligned} t_1 t_2^{1.7} &\leq 1 \\ 1+x_1 \leq t_1, \quad 1+x_2 &\leq t_1 \\ 1+x_1+t_3^{1.5} &\leq t_2 \\ 0.1x_1x_3^{-0.5}+x_2^{1.6}x_3^{0.4} &\leq t_3 \end{aligned}$$

Through this example, we see monotonicity is the key guarantee of the applicability of our trick. Interestingly, this monotonicity-based trick goes even beyond GPPs, and we illustrate it by more examples.

More GPP Examples

Example (Fraction)

Consider first the constraints:

$$\frac{p_1(x)}{m(x) - p_2(x)} + p_3(x) \leq 1 \text{ and } p_2(x) < m(x),$$

where $p_i(x)$ are generalized posynomials and $m(x)$ is a monomial. Obviously, they do not fall into GPPs. However, it is easily seen that the two constraints are equivalent to

$$t + p_3(x) \leq 1 \text{ and } \frac{p_2(x)}{m(x)} < 1 \text{ and } \frac{p_2(x)}{m(x)} + \frac{p_1(x)}{t \cdot m(x)} \leq 1,$$

which indeed fall into GPPs.

More GPP Examples cont'

Example (Exponential)

Suppose we have an exponential constraint $e^{p(x)} \leq t$, this clearly does not fall into GPPs. However, by changing variables, we get $e^{p(e^y)} \leq e^s$, which is equivalent to $p(e^y) \leq s$. This latter constraint is obviously convex since $p(e^y)$ is a convex function, according to our generalized log-sum-exp proposition.

Example (Logarithmic)

Instead if we have a logarithmic constraint $p_1(x) + \log p_2(x) \leq 1$, we can still convert it into GPPs. Changing variables we get $p_1(e^y) + \log p_2(e^y) \leq 1$, which is clearly convex since both functions on the LHS are convex.

Summary

We have seen six different categories of general convex problems, and in fact they form a hierarchy (exclude GPPs):

- The power of these categories monotonically increases, that is, every category (except SDP) is a special case of the later one. Verify this by yourself;
- The computational complexity monotonically increases as well, and this reminds us that whenever possible to formulate our problem as an instance of lower hierarchy, never formulate it as an instance of higher hierarchy;
- We've seen that many problems (including non-convex ones) do not seem like to fall into these five categories at first, but can be (equivalently) reformulated as so. This usually requires some efforts but you have learnt some tricks.

Outline

Fenchel Conjugate

Definition

The Fenchel conjugate of $g(x)$ (not necessarily convex) is:

$$g^*(x^*) = \max_x \; x^T x^* - g(x).$$

Fenchel inequality: $g(x) + g^*(x^*) \geq x^T x^*$ (when equality holds?).

Remark: $(f_1 + f_2)^* = f_1^* \Box f_2^* \neq f_1^* + f_2^*$, assuming closedness.

Proposition

Fenchel conjugate is always (closed) convex.

Theorem (Double Conjugation is the Convex Hull)

$$g^{**} = \mathrm{cl}\ \mathrm{conv}\ g.$$

Special case: $f^{**} = \mathrm{cl}\ f$.

Remark: A special case of Fenchel conjugate is called Legendre conjugate, where $f(\cdot)$ is restricted to be differentiable and strictly convex (i.e. both $f(\cdot)$ and $f^*(\cdot)$ are differentiable).

Fenchel Conjugate Examples

Quadratic function

Let $f(x) = 1/2x^TQx + a^Tx + \alpha, Q \succ 0$, what is $f^*(\cdot)$?
Want to solve $\max_x \; x^Tx^* - 1/2x^TQx - a^Tx - \alpha$, set the derivative to zero (why?), get $x = Q^{-1}(x^* - a)$. Plug in back, $f^*(x^*) = 1/2(x^* - a)^TQ^{-1}(x^* - a) + a^TQ^{-1}(x^* - a) + \alpha$.

Norms

Set $Q = I, a = 0, \alpha = 0$ in the above example, we know the Euclidean norm $\|\cdot\|_2$ is self-conjugate. More generally, the conjugate of $\|\cdot\|_p$ is $\|\cdot\|_q$ if $1/p + 1/q = 1, p \geq 1$. Specifically, $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are conjugate pairs.

Matrix norms are similar to their vector cousins. In particular, Frobenius norm is self-conjugate, and the conjugate of the spectral norm (largest singular value) is the trace norm (sum of singular values).

More Interesting Examples

In many cases, one really needs to minimize the ℓ_0-norm, which is unfortunately non-convex. The remedy is to instead minimize the so-called tightest convex approximation, namely, $\mathrm{conv}\|\cdot\|_0$.

We've seen that $g^{**} = \mathrm{conv}\, g$, so let's compute $\mathrm{conv}\|\cdot\|_0$.

Step 1: $(\|\cdot\|_0)^*(x^*) = \max_x x^T x^* - \|x\|_0 = \begin{cases} 0, & x^* = 0 \\ \infty, & \text{otherwise} \end{cases}$

Step 2: $(\|\cdot\|_0)^{**}(x) = \max_{x^*} x^T x^* - (\|\cdot\|_0)^*(x^*) = 0.$

Hence, $(\mathrm{conv}\|\cdot\|_0)(x) = 0$! Is this correct? Draw a graph to verify. Is this a meaningful surrogate for $\|\cdot\|_0$? Not really...

Stare at the graph you drew. What prevents us from obtaining a meaningful surrogate? How to get around?

More Interesting Examples cont'

Yes, we need some kind of truncation! Consider the ℓ_0-norm restricted to the ℓ_∞-ball region $\|x\|_\infty \leq 1$. Redo it.

Step 1: $(\|\cdot\|_0)^*(x^*) = \max_{\|x\|_\infty \leq 1} x^T x^* - \|x\|_0 = \sum_i (|x_i^*| - 1)_+$

Step 2:

$$(\|\cdot\|_0)^{**}(x) = \max_{x^*} x^T x^* - (\|\cdot\|_0)^*(x^*) = \begin{cases} \|x\|_1, & \|x\|_\infty \leq 1 \\ \infty, & \text{otherwise} \end{cases}.$$

Does the result coincide with your intuition? Check your graph.

Remark: Use Von Neumann's lemma to prove the analogy in the matrix case, i.e. the rank function.

We will see another interesting connection when discussing the Lagrangian duality.

More Interesting Examples cont'[2]

Let us now truncate the ℓ_0-norm differently. To simplify the calculations, we can w.l.o.g. assume below $x \geq 0$ (or $x^* \geq 0$) and its components are ordered in decreasing manner. Consider first restricting the ℓ_0-norm to the ℓ_1-ball $\|x\|_1 \leq 1$.

Step 1: $(\|\cdot\|_0)^*(x^*) = \max_{\|x\|_1 \leq 1} x^T x^* - \|x\|_0 = (\|x^*\|_\infty - 1)_+$

Step 2:

$$(\|\cdot\|_0)^{**}(x) = \max_{x^*} x^T x^* - (\|\cdot\|_0)^*(x^*) = \begin{cases} \|x\|_1, & \|x\|_1 \leq 1 \\ \infty, & \text{otherwise} \end{cases}.$$

Notice that the maximizer of x^* is at $\mathbb{1}$.

Next consider the general case, that is, restricting the ℓ_0-norm to the ℓ_p-ball $\|x\|_p \leq 1$. Assume of course, $p \geq 1$, and let $1/p + 1/q = 1$.

Step 1: $(\|\cdot\|_0)^*(x^*) = \max_{\|x\|_p \leq 1} x^T x^* - \|x\|_0 = \max_{0 \leq k \leq n} \|x^*_{[1:k]}\|_q - k$,

where $x_{[1:k]}$ denotes the largest (in terms of absolute values) k components of x.

More Interesting Examples cont'[3]

Convince yourself the RHS (on the previous slide), which has to be convex, is indeed convex. Also you can verify that this formula is correct for the previous two special examples $p = 1, \infty$.

Step 2:

$$(\|\cdot\|_0)^{**}(x) = \max_{x^*} x^T x^* - (\|\cdot\|_0)^*(x^*) = \begin{cases} \|x\|_1, & \|x\|_p \leq 1 \\ \infty, & \text{otherwise} \end{cases}.$$

To see why, suppose first $\|x\|_p > 1$, set $y/a = \arg \max_{\|x^*\|_q \leq 1} x^T x^*$, then $(\|\cdot\|_0)^{**}(x) \geq x^T y - (\|\cdot\|_0)^*(y) \geq a\|x\|_p - a$, letting $a \to \infty$ proves the otherwise case. Since the ℓ_q-norm is decreasing as a function of q, we have the inequality (for any $q \geq 1$):

$$x^T x^* - \left[\max_{0 \leq k \leq n} \|x^*_{[1:k]}\|_q - k\right] \leq x^T x^* - \left[\max_{0 \leq k \leq n} \|x^*_{[1:k]}\|_\infty - k\right]$$

Maximizing both sides (w.r.t. x^*) gives us $(\|\cdot\|_0)^{**}(x) \leq \|x\|_1$, for any truncation $p \geq 1$, and the equality is indeed attained, again, at $\mathbb{1}$.

Outline

Weak Duality

Theorem (Weak Duality)

$$\min_{x \in M} \max_{y \in N} g(x, y) \geq \max_{y \in N} \min_{x \in M} g(x, y).$$

Interpretation: It matters who plays first in games (but not always).

Proof.

Step 1: $\forall x_0 \in M, y_0 \in N$, we have $g(x_0, y_0) \geq \min_{x \in M} g(x, y_0)$;

Step 2: Maximize w.r.t. y_0 on both sides:

$$\forall x_0 \in M, \quad \max_{y_0 \in N} g(x_0, y_0) \geq \max_{y_0 \in N} \min_{x \in M} g(x, y_0)$$

Step 3: Minimize w.r.t. x_0 on both sides, but note that the RHS does not depend on x_0 at all. $\square$

Strong Duality

Theorem (Sion, 1958)

Let $g(x, y)$ be l.s.c. and quasi-convex on $x \in M$, u.s.c. and quasi-concave on $y \in N$, while M and N are convex sets and one of them is compact, then

$$\min_{x \in M} \max_{y \in N} g(x, y) = \max_{y \in N} \min_{x \in M} g(x, y).$$

Remark: Don't forget to check the **crucial** "compact" assumption!
Note: Sion's original proof used the KKM lemma and Helly's theorem, which is a bit advanced for us. Instead, we consider a rather elementary proof provided by Hidetoshi Komiya (1988).
Advertisement: Consider seriously reading the proof, since this's probably the only chance in your life to fully appreciate this celebrated theorem. Oh, math!

Proof: We need only to show $\min \max g(x, y) \leq \max \min g(x, y)$, and we can w.l.o.g. assume M is compact (otherwise consider $-g(x, y)$). We prove two technical lemmas first.

Proof cont'

Lemma (Key)

If $y_1, y_2 \in N$ and $\alpha \in \mathbb{R}$ satisfy $\alpha < \min_{x \in M} \max\{g(x, y_1), g(x, y_2)\}$, then $\exists y_0 \in N$ with $\alpha < \min_{x \in M} g(x, y_0)$.

Proof: Assume to the contrary, $\min_{x \in M} g(x, y) \geq \alpha, \forall y \in N$. Let $C_z = \{x \in M : g(x, z) \leq \alpha\}$. Notice that $\forall z \in [y_1, y_2]$, C_z is closed (l.s.c.), convex (quasi-convexity) and non-empty (otherwise we are done). We also know C_{y_1}, C_{y_2} are disjoint (given condition). Because of quasi-concavity, $g(x, z) \geq \min\{g(x, y_1), g(x, y_2)\}$, hence C_z belongs to either C_{y_1} or C_{y_2} (convex sets must be connected), which then divides $[y_1, y_2]$ into two disjoint parts. Pick any part and choose two points z', z'' in it. For any sequence $\lim z_n = z$ in this part, using quasi-concavity again and u.s.c. we have $g(x, z) \geq \limsup g(x, z_n) \geq \min\{g(x, z'), g(x, z'')\}$. Thus both parts are closed, which is impossible. $\square$

Proof cont'[2]

Lemma (Induction)

If $\alpha < \min_{x \in M} \max_{1 \le i \le n} g(x, y_i)$, *then* $\exists y_0 \in N$ *with* $\alpha < \min_{x \in M} g(x, y_0)$.

Proof: Induction from the previous lemma. □

Now we are ready to prove Sion's theorem. Let $\alpha < \min \max g$ (what if such α does not exist?) and let M_y be the **compact** set $\{x \in M : g(x, y) \le \alpha\}$ for each $y \in N$. Then $\bigcap_{y \in N} M_y$ is empty, and hence by the compactness assumption on M, there are finite points $y_1, \dots y_n \in N$ such that $\bigcap_{y_i} M_{y_i}$ is empty, that is $\alpha < \min_{x \in M} \max_{1 \le i \le n} g(x, y_i)$. By the induction lemma, we know $\exists y_0$ such that $\alpha < \min_{x \in M} g(x, y_0)$, and hence $\alpha < \max \min g$. Since α can be chosen arbitrarily, we get $\min \max g \le \max \min g$. □

Remark: We used u.s.c., quasi-concavity, quasi-convexity in the key lemma, l.s.c. and compactness in the main proof. It can be shown that neither of these assumptions can be appreciably weakened.

Variations

Theorem (Von Neumann, 1928)

$$\min_{x\in\Delta_m}\max_{y\in\Delta_n} x^T Ay = \max_{y\in\Delta_n}\min_{x\in\Delta_m} x^T Ay,$$

where $\Delta_m := \{x : x_i \geq 0, \sum_{i=1}^{m} x_i = 1\}$ *is the standard simplex.*

Theorem (Ky Fan, 1953)

Let $g(x, y)$ *be convex-concave-like on* $M \times N$*, where i).* M *any space,* N *compact on which* g *is u.s.c.; or ii).* N *any space,* M *compact on which* g *is l.s.c., then*

$$\min_{x\in M}\max_{y\in N} g(x, y) = \max_{y\in N}\min_{x\in M} g(x, y).$$

Remark: We can apply either Sion's theorem or Ky Fan's theorem when $g(x, y)$ is convex-concave, however, note that Ky Fan's theorem does not require (explicitly) any convexity on the domain M and N!

Proof: We resort to an elementary proof based on the separation theorem, appeared first in J. M. Borwein and D. Zhuang (1986).

Variations cont'

Let $\alpha < \min \max g$, as in the proof of Sion's theorem, $\exists$ finite points $y_1, \dots y_n \in N$ such that $\alpha < \min_{x \in M} \max_{1 \le i \le n} g(x, y_i)$. Now consider the set

$$\mathcal{C} := \{(z, r) \in \mathbb{R}^{n+1} \mid \exists x \in M, g(x, y_i) \le r + z_i, i = 1, \dots, n\}.$$

$\mathcal{C}$ is obviously convex since g is convex-like (in x). Also by construction, $(0_n, \alpha) \notin \mathcal{C}$. By the separation theorem, $\exists\ \theta_i, \gamma$ such that

$$\sum_i \theta_i z_i + \gamma r \ge \gamma \alpha,\ \forall (z, r) \in \mathcal{C}.$$

Notice that $\mathcal{C} + \mathbb{R}_+^{n+1} \subseteq \mathcal{C}$, therefore $\theta_i, \gamma \ge 0$. Moreover, $\forall x \in M$, the point $(0_n, \max_{1 \le i \le n} g(x, y_i) + 1) \in \mathrm{int}\ \mathcal{C}$, meaning that $\gamma \ne 0$ (otherwise contradicting the separation). Consider the point $(g(x, y_1) + r, \dots, g(x, y_n) + r, -r) \in \mathcal{C}$, we know $\sum_i \theta_i [g(x, y_i) + r] - \gamma r \ge \gamma \alpha \Rightarrow \sum_i \frac{\theta_i}{\gamma} g(x, y_i) + r(\sum_i \frac{\theta_i}{\gamma} - 1) \ge \alpha$. Since r can be chosen arbitrarily in $\mathbb{R}$, we must have $\sum_i \frac{\theta_i}{\gamma} = 1$. Hence by concave-like, $\exists y_0$ such that $g(x, y_0) \ge \alpha, \forall x$. $\square$

Minimax Examples

Example (It matters a lot who plays first!)

$$\min_x \max_y \quad x + y = \infty,$$

$$\max_y \min_x \quad x + y = -\infty.$$

Example (It does not matter who plays first!)

Let's assure compactness on the y space:

$$\min_x \max_{0 \le y \le 1} \quad x + y = -\infty,$$

do we still need to compute $\max \min$ in this case?

Example (Sion's theorem is not necessary)

$$\min_x \max_{y \le 0} \quad x + y = -\infty,$$

No compactness, but strong duality still holds.

Alternative Optimization

A simple strategy for the following problem

$$\min_{x \in M} \min_{y \in N} f(x, y)$$

is to alternatively fix one of x and y while minimize w.r.t the other. *Under appropriate conditions*, this strategy, called decomposition method or coordinate descent or Gauss-Seidel update etc., converges to optimum.

Remark: To understand "under appropriate conditions", consider:

$$\min_{x} \min_{y} \quad x^2 \quad \text{s.t.} \quad x + y = 1.$$

Initialize x_0 randomly, will the alternative strategy converge to optimum? So the minimum requirement is decision variables do not interact through constraints.

Can we apply this alternative strategy to minimax problems?

Think...

Alternative Optimization cont'

The answer is NO. Consider the following trivial example:

$$\min_{-1\leq x\leq 1} \max_{-1\leq y\leq 1} xy$$

The true saddle-point is obviously (0,0). However, if we use alternative strategy, suppose we initialize x_0 randomly, w.p.1 $x_0 \neq 0$, assume $x_0 > 0$:
Maximize w.r.t. y gives $y_0 = 1$;
Minimize w.r.t. x gives $x_1 = -1$;
Maximize w.r.t. y again gives $y_1 = -1$;
Minimize w.r.t. x again gives $x_2 = 1$;
and oscillate so on.

The analysis is similar when $x_0 < 0$, hence w.p.1 the alternative strategy does not converge!

Outline

Kuhn-Tucker (KT) Vector

Recall the convex program (which we call primal from now on):

$$\min_{x \in \mathcal{C}} \quad f_0(x) \tag{8}$$

$$\text{s.t.} \quad f_i(x) \leq 0, i = 1, \ldots, m \tag{9}$$

$$a_j^T x = b_j, j = 1, \ldots, n \tag{10}$$

Assume you are given a KT vector, $\mu_i \geq 0, \nu_j$, which ensure you the minimum (being finite) of

$$\min_{x \in \mathcal{C}} \quad L(x, \mu, \nu) := f_0(x) + \sum_i \mu_i f_i(x) + \sum_j \nu_j (a_j^T x - b_j) \tag{11}$$

equals that of the primal (8). We will call $L(x, \mu, \nu)$ the Lagrangian from now on. Obviously, any minimizer of (8) must be also a minimizer of (11), therefore if we were able to collect all minimizers of (11), we can pick those of (8) by simply verifying constraints (9) and (10). Notice that the KT vector turns the constrained problem (8) into an unconstrained one (11)!

Existence and KKT Conditions

Before we discuss how to find a KT vector, we need to be sure about its existence.

Theorem (Slater's Condition)

Assume the primal (8) is bounded from below, and $\exists x_0$*, in the relative interior of the feasible region, satisfies the (non-affine) inequalities strictly, then a KT vector (not necessarily unique) exists.*

Let $x^\star$ be any minimizer of primal (8), and $(\mu^\star, \nu^\star)$ be any KT vector, then they must satisfy the KKT conditions:

$$f_i(x^\star) \leq 0, a_j^T x^\star = b_j \tag{12}$$

$$\mu_i^\star \geq 0 \tag{13}$$

$$0 \in \partial f_0(x^\star) + \sum_i \mu_i^\star \partial f_i(x^\star) + \sum_j \nu_j^\star a_j \tag{14}$$

The remarkable thing is KKT conditions, being necessary for non-convex problems, are sufficient as well for convex programs!

How to find a KT vector?

A KT vector, when exists, can be found, simultaneously with the minimizer $x^\star$ of primal, by solving the saddle-point problem:

$$\min_{x \in \mathcal{C}} \max_{\mu \geq 0, \nu} L(x, \mu, \nu) = \max_{\mu \geq 0, \nu} \min_{x \in \mathcal{C}} L(x, \mu, \nu). \tag{15}$$

Remark: The strong duality holds from Sion's theorem, but notice that we need compactness on one of the domains, and here existence of a KT vector ensures this (why?).

Denote $g(\mu, \nu) := \min_{x \in \mathcal{C}} L(x, \mu, \nu)$, show by yourself it is always concave even for non-convex primals, hence the RHS of (15) is always a convex program, and we will call it the dual problem.

Remark: The Lagragian multipliers method might seem "stupid" since we are now doing some extra work in order to find $x^\star$, however, the catch is the dual problem, compared to the primal, has very simple constraints. Moreover, since the dual problem is always convex, a common trick to solve (to some extent) non-convex problems is to consider their duals.

The Decomposition Principle (taken from Ref. 2)

Most times the complexity of our problem is *not* linear, hence by decomposing the problem into small pieces, we could reduce (oftentimes significantly) the complexity. We now illustrate the decomposition principle by a simple example:

$$\min_{x \in \mathbb{R}^n} \quad \sum_i f_i(x_i) \quad \text{s.t.} \quad \sum_i x_i = 1.$$

Wouldn't it be nice if we had a KT vector λ? Since the problem

$$\min_x \quad \sum_i [f_i(x_i) + \lambda x_i] - \lambda$$

can be solved separably for each x_i. Consider the dual:

$$\max_\lambda \quad \min_x \sum_i [f_i(x_i) + \lambda x_i] - \lambda.$$

Using Fenchel conjugates of $f_i(x)$, the dual can be written compactly as:

$$\min_\lambda \quad \lambda + \sum_i f_i^*(-\lambda),$$

hence we've reduced a convex program in $\mathbb{R}^n$ into $n + 1$ convex problems in $\mathbb{R}$.

Primal-Dual Examples

Let us finish this mini-tutorial by some promised examples.

Example (Primal-Dual SDPs)

Consider the primal SDP:

$$\begin{array}{ll} \min\limits_{x} & c^T x \\ \text{s.t.} & \sum_i x_i F_i + G \preceq 0 \end{array}$$

The dual problem is

$$\max_{X \succeq 0} \quad \min_{x} c^T x + \mathrm{Tr}\Big[X(\sum_i x_i F_i + G)\Big],$$

solving the inner problem (i.e. setting derivate w.r.t. x_i to 0) gives the standard dual SDP formulation.

Remark: Using this example to show that the double dual of a convex program is itself.

Primal-Dual Examples cont'

Example (Euclidean Projection Revisited)

$$\min_{\|x\|_2^2 \le 1} \quad \|x - x_0\|_2^2$$

Assume $\|x_0\| > 1$, otherwise the minimizer is x_0 itself. The dual is:

$$\max_{\lambda \ge 0} \quad \min_x \left[\|x - x_0\|_2^2 + \lambda(\|x\|_2^2 - 1) \right].$$

Solving the inner problem ($x^\star = \frac{x_0}{1+\lambda}$) simplifies the dual to:

$$\max_{\lambda \ge 0} \quad \|x_0\|_2^2 \cdot \frac{\lambda}{1+\lambda} - \lambda.$$

Solving this 1-dimensional problem (just setting the derivative to 0, why?) gives $\lambda^\star = \|x_0\|_2 - 1$, hence $x^\star = x_0 / \|x_0\|_2$. Does the solution coincide with your geometric intuition? Of course, there is no necessity to use the powerful Lagrangian multipliers to solve this trivial problem, but the point is we can now start to use the same procedure to solve slightly harder problems, such as projection to the ℓ_1 ball.

Primal-Dual Examples cont'[2]

Example (Robust LP Revisited)

$$\begin{aligned} \min_x \quad & c^T x \\ \text{s.t.} \quad & \left[\max_{a \in \mathcal{E}} a^T x\right] \leq 0 \end{aligned}$$

We use Lagrangian multipliers to solve the red:

$$\max_a \quad \min_{\lambda \leq 0} \quad a^T x + \lambda \cdot [(a - \bar{a})^T \Sigma^{-1} (a - \bar{a}) - 1]$$

Swap max and min, solve $a^\star = \bar{a} - \frac{1}{2\lambda} \Sigma x$, plug in back, we get

$$\min_{\lambda \leq 0} -\lambda - \frac{1}{4\lambda} x^T \Sigma x + \bar{a}^T x.$$

Solving $\lambda^\star = -\frac{\|\Sigma^{1/2} x\|_2}{2}$, plug in back, we get

$$\left[\max_{a \in \mathcal{E}} a^T x\right] = \|\Sigma^{1/2} x\|_2 + \bar{a}^T x,$$

which confirms the robust LP is indeed an SOCP.

Outline

References

1. Introductory convex optimization book:
   Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.
2. Great book on convex analysis:
   Ralph T. Rockafellar. *Convex Analysis.* Princeton University Press, 1970.
3. Nice introduction of optimization strategies:
   Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Kluwer Academic Publishers, 2003.
4. The NP-Hard convex example is taken from:
   Mirjam Dür. *Copositive Programming: A Survey.* Manuscript, 2009.
5. The GPP subsection are mainly based on:
   Stephen Boyd, Seung-Jean Kim, Lieven Vandenberghe and Arash Hassibi. *A Tutorial on Geometric Programming*. Optimization & Engineering. vol. 8, pp. 67-127, 2007.
6. The proof of Sion's theorem is mainly taken from:
   Hidetoshi Komiya. *Elementary proof for Sion's minimax theorem.* Kodai Mathematical Journal. vol. 11, no. 1, pp. 5-7, 1988.
7. The proof of Ky Fan's theorem is mainly taken from:
   J. M. Bowrein and D. Zhuang. *On Fan's Minimax Theorem.* Mathematical Programming. vol. 34, pp. 232-234, 1986.