SDSC6015 - Semester A, 2025

Stochastic Optimization for Machine Learning

2. Convex Optimization Problems

1 Convex Optimization

1.1 Basic Terminology in Optimization

$$\begin{aligned} \mathrm{P}: \quad & \underset{\boldsymbol{x}}{\text{minimize}} && f_0(\boldsymbol{x}) \\ & \text{subject to} && \boldsymbol{x} \in \mathcal{X} \end{aligned}$$

- $\boldsymbol{x}$ is the *decision variable.*
- f_0 is the *objective function.*
- $\mathcal{X}$ is the *feasible set*, which is usually defined through constraints

$$\begin{aligned} f_i(\boldsymbol{x}) &\leq 0, \quad \forall i \in [m], \\ h_i(\boldsymbol{x}) &= 0, \quad \forall i \in [p]. \end{aligned}$$

— Global Minimizers

- The *infimum* of problem P, inf P, is the largest number ℓ such that

$$\ell \leq f_0(\boldsymbol{x}), \qquad \forall \boldsymbol{x} \in \mathcal{X}.$$

- If there is $\boldsymbol{x}^\star \in \mathcal{X}$ with $f_0(\boldsymbol{x}^\star) \leq f_0(\boldsymbol{x})$, $\forall \boldsymbol{x} \in \mathcal{X}$, then $\boldsymbol{x}^\star$ is a (global) minimizer of P, and $f_0(\boldsymbol{x}^\star)$ is the (global) minimum of P.
- If P has a minimizer, then the minimum is equal to the infimum.
- P always has an infimum (∞ if $\mathcal{X} = \emptyset$) but may have no minimum, *e.g.*

$$\begin{aligned} & \underset{x \in \mathbb{R}}{\text{minimize}} && 1/x \\ & \text{subject to} && x \geq 1 \end{aligned}$$

— **Feasibility**

- Problem P is called *feasible* if $\mathcal{X} \neq \emptyset$. Otherwise, the problem is called *infeasible*.

- Problem P is called *unbounded* if $\inf \text{P} = -\infty$.

- Problem P is called *solvable* if it has at least one minimizer.

- Suppose $\mathcal{X} = \{\boldsymbol{x} \mid f_i(\boldsymbol{x}) \leq 0,\ h_j(\boldsymbol{x}) = 0,\ \forall i \in [m],\ \forall j \in [p]\}$. The constraints $f_i(\boldsymbol{x}) \leq 0$ and $h_j(\boldsymbol{x}) = 0$ are *explicit* constraints. The *implicit* constraint is

$$\boldsymbol{x} \in \mathcal{D} = \left(\bigcap_{i=0}^{m} \mathbf{dom}(f_i)\right) \cap \left(\bigcap_{j=1}^{p} \mathbf{dom}(h_i)\right),$$

where $\mathcal{D}$ is called the domain of the problem.

- Problem P is **unconstrained** if it has no *explicit* constraint, *i.e.* $m = p = 0$. For example,

$$\underset{\boldsymbol{x} \in \mathbb{R}^n}{\text{minimize}} \quad -\sum_{i=1}^{k} \log\left(b_i - \boldsymbol{a}_i^\top \boldsymbol{x}\right)$$

is unconstrained problem with $\mathcal{D} = \{\boldsymbol{x} \mid \boldsymbol{a}_i^\top \boldsymbol{x} < b_i,\ \forall i \in [k]\}$.

— **Local Minimizers & $\epsilon-$Optimal Solutions**

Definition 1 (Local minimizer)**.** *If there is* $\boldsymbol{x}^\star \in \mathcal{X}$ *and* $\delta > 0$ *such that*

$$f_0(\boldsymbol{x}^\star) \leq f_0(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathcal{X} \cap \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2 \leq \delta\},$$

then $\boldsymbol{x}^\star$ *is called a local minimizer and* $f_0(\boldsymbol{x}^\star)$ *is a local minimum of P.*

Definition 2 ($\epsilon-$optimal solution*)**.** $\boldsymbol{x}_\epsilon \in \mathcal{X}$ *is called an* $\epsilon-$*optimal solution of P if* $\epsilon > 0$ *and*

$$f_0(\boldsymbol{x}_\epsilon) - \inf P < \epsilon.$$

— **The Weierstrass Theorem**

Theorem 1 (The Weierstrass Theorem)**.** *Let $\mathcal{D} \subset \mathbb{R}^n$ be compact (closed and bounded), and let $f : \mathcal{D} \to \mathbb{R}$ be a continuous function on $\mathcal{D}$. Then f attains a maximum and a minimum on $\mathcal{D}$, i.e., there exists points $\boldsymbol{x}_{low}, \boldsymbol{x}_{up} \in \mathcal{D}$ such that*

$$f(\boldsymbol{x}_{low}) \leq f(\boldsymbol{x}) \leq f(\boldsymbol{x}_{up}), \quad \forall \boldsymbol{x} \in \mathcal{D}.$$

Examples:

- $\mathcal{D} = \mathbb{R}$ and $f(x) = x^3$
- $\mathcal{D} = (0, 1)$ and $f(x) = x$
- $\mathcal{D} = [-1, 1]$ and

$$f(x) = \begin{cases} 0, & \text{if } x = -1 \text{ or } x = 1 \\ x, & \text{if } -1 < x < 1 \end{cases}$$

- $\mathcal{D} = \mathbb{R}_{++}$ and

$$f(x) = \begin{cases} 1, & \text{if } x \text{ is rational} \\ 0, & \text{otherwise.} \end{cases}$$

— **Quick Summary**

1.2 Basic in Convex Optimization

The standard form of convex optimization problem:

$$
\begin{aligned}
\mathrm{P}: \quad & \underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} && f_0(\boldsymbol{x}) \\
& \text{subject to} && f_i(\boldsymbol{x}) \le 0, \quad \forall i \in [m] \\
& && \boldsymbol{a}_i^\top \boldsymbol{x} = b_i, \quad \forall i \in [p]
\end{aligned}
$$

The objective and inequality constraint functions are convex, while all equality constraint functions are linear (affine).

Proposition 1. *The feasible set of a convex optimization problem is convex.*

Question: If the feasible set and objective function are convex, then it is a convex optimization problem?

Theorem 2. *Consider an optimization problem*

$$\begin{aligned} \underset{\boldsymbol{x} \in \mathbb{R}^n}{\textit{minimize}} \quad & f_0(\boldsymbol{x}) \\ \textit{subject to} \quad & \boldsymbol{x} \in \mathcal{X}, \end{aligned}$$

where $f_0 : \mathbb{R}^n \to \mathbb{R}$ *is a convex function and* $\mathcal{X}$ *is a convex set. Then any local minimum is also a global minimum.*

Theorem 3. *Suppose the objective function* $f_0 : \mathbb{R}^n \to \mathbb{R}$ *is differentiable. For* P *convex,* $\boldsymbol{x}^\star$ *is optimal if and only if* $\boldsymbol{x}^\star \in \mathcal{X}$ *(feasible set) and*

$$\nabla f_0(\boldsymbol{x}^\star)^\top (\boldsymbol{x} - \boldsymbol{x}^\star) \geq 0, \qquad \forall \boldsymbol{x} \in \mathcal{X}.$$

Special Case: Suppose P is unconstrained: $\boldsymbol{x}^\star$ is optimal if and only if $\boldsymbol{x}^\star \in \mathrm{dom}(f_0)$ and $\nabla f_0(\boldsymbol{x}^\star) = 0$

— **Convexity After Minimization**

Proposition 2. *If $f(\boldsymbol{x}, \boldsymbol{y}) : \mathbb{R}^{n_x \times n_y} \to \mathbb{R}$ is a convex function in $(\boldsymbol{x}, \boldsymbol{y})$ and C is nonempty convex set. Then*

$$g(\boldsymbol{x}) = \inf_{y \in C} \; f(\boldsymbol{x}, \boldsymbol{y})$$

is a convex function.

Example: Schur's Lemma

Lemma 1 (Schur's Lemma)**.** *Consider* $\boldsymbol{X} \in \mathbb{S}^n$ *partitioned as* $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{C} \end{bmatrix}$, *where* $\boldsymbol{C} \succ \boldsymbol{0}$. *Then*

$$\boldsymbol{X} \succeq \boldsymbol{0} \qquad \Longleftrightarrow \qquad \boldsymbol{A} - \boldsymbol{B}\boldsymbol{C}^{-1}\boldsymbol{B}^\top \succeq \boldsymbol{0}$$

The matrix $\boldsymbol{A} - \boldsymbol{B}\boldsymbol{C}^{-1}\boldsymbol{B}^\top$ *is called Schur complement of* $\boldsymbol{C}$.

Example: Distance Function The distance of $\boldsymbol{x}$ to a fixed nonempty convex set C is convex in x.

$$f(\boldsymbol{x}) = \inf_{\boldsymbol{y} \in C} \|\boldsymbol{x} - \boldsymbol{y}\|_2$$

— Equivalent Optimization Problems

"Definition": Two problems P and P' are *equivalent* (written as P $\Longleftrightarrow$ P') if the solution of P' is obtained from that of P via "elementary transformations" and vice versa.

Example: Epigraphical Reformulation The standard convex optimization problem is equivalent to

$$
\begin{array}{llll}
\text{P'}: & \underset{\boldsymbol{x}\in\mathbb{R}^n,\, t\in\mathbb{R}}{\text{minimize}} & t & \\
& \text{subject to} & f_0(\boldsymbol{x}) - t \leq 0 & \\
& & f_i(\boldsymbol{x}) \leq 0, & \forall i \in [m] \\
& & \boldsymbol{a}_i^\top \boldsymbol{x} = b_i, & \forall i \in [p]
\end{array}
$$

Question: what does this mean?

Example: Partial Minimization

$$
\begin{array}{llll}
\text{P}: & \underset{\boldsymbol{x}_1,\boldsymbol{x}_2}{\text{minimize}} & f_0(\boldsymbol{x}_1, \boldsymbol{x}_2) & \\
& \text{subject to} & f_i(\boldsymbol{x}_1) \leq 0, & \forall i \in [m] \\
& & g_i(\boldsymbol{x}_2) \leq 0, & \forall i \in [p]
\end{array}
$$

is equivalent to

$$
\begin{array}{llll}
\text{P'}: & \underset{\boldsymbol{x}_1}{\text{minimize}} & \hat{f}_0(\boldsymbol{x}_1) & \\
& \text{subject to} & f_i(\boldsymbol{x}_1) \leq 0, & \forall i \in [m]
\end{array}
$$

where $\hat{f}_0(\boldsymbol{x}_1) = \inf_{\boldsymbol{x}_2} \{f_0(\boldsymbol{x}_1, \boldsymbol{x}_2) \mid g_i(\boldsymbol{x}_2) \leq 0,\ \forall i \in [p]\}$

Example: Slack Variables

$$
\begin{array}{llll}
\text{P}: & \underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} & f_0(\boldsymbol{x}) & \\
& \text{subject to} & \boldsymbol{a}_i^\top \boldsymbol{x} \leq b_i, & \forall i \in [p]
\end{array}
$$

is equivalent to

$$
\begin{array}{llll}
\text{P}: & \underset{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{s}\in\mathbb{R}^p}{\text{minimize}} & f_0(\boldsymbol{x}) & \\
& \text{subject to} & \boldsymbol{a}_i^\top \boldsymbol{x} + s_i = b_i, & \forall i \in [p] \\
& & s_i \geq 0, & \forall i \in [p]
\end{array}
$$

Example: Nonconvex Instance

$$\begin{aligned} \mathrm{P}: \quad & \underset{x_1, x_2 \in \mathbb{R}}{\text{minimize}} && x_1^2 + x_2^2 \\ & \text{subject to} && \frac{x_1}{1 + x_2^2} \leq 0 \\ & && (x_1 + x_2)^2 = 0 \end{aligned}$$

- The LHS in the first constraint is not a convex function.
- The LHS in the second constraint is not linear.

2 Linear Programming (LP)

The standard form of **linear program** (see week 1 material for alternative forms) is the following

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \quad & \boldsymbol{c}^\top \boldsymbol{x} + d \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \\ & \boldsymbol{C}\boldsymbol{x} \leq \boldsymbol{g} \end{aligned}$$

where $\boldsymbol{c} \in \mathbb{R}^n$, $d \in \mathbb{R}$, $\boldsymbol{A} \in \mathbb{R}^{m_1\times n}$, $\boldsymbol{b} \in \mathbb{R}^{m_1}$, $\boldsymbol{C} \in \mathbb{R}^{m_2\times n}$, $\boldsymbol{g} \in \mathbb{R}^{m_2}$.

— **Example: Transportation**

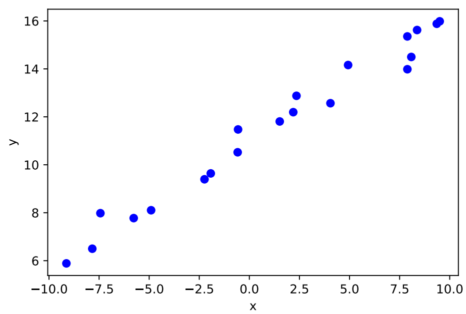
- Shipping goods from different sources to different destinations
- a_i = the amount of goods in source i, $i \in [m]$
- b_j = the amount of goods needed in destination j, $j \in [n]$
- c_{ij} = the costs of shipping 1 unit from source i to j

2.1 Application: L_1-Norm Regression

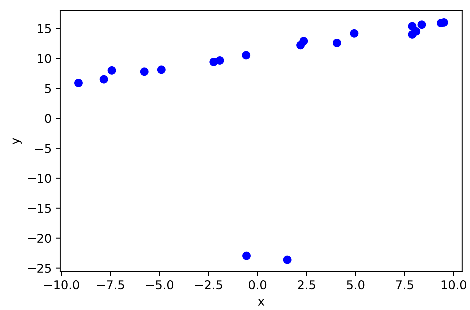
We consider a linear model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$

- $\boldsymbol{y} = [y_1, y_2, \ldots, y_m]^\top \in \mathbb{R}^m$
- $\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & \cdots & x_{mn} \end{bmatrix} \in \mathbb{R}^{m \times (n+1)}$
- $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_n]^\top \in \mathbb{R}^{n+1}$

If we apply least squares...



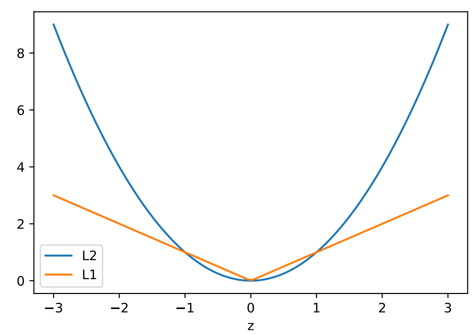
It works well



It does not work well

How about..

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 \quad \text{v.s.} \quad \min_{\boldsymbol{\beta}} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_1$$



Question: *How do we solve* $\min_{\boldsymbol{\beta}} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_1$*?*

.

2.2 Solving the Problem with Solvers

- *Programming Languages:* C++, Python, Julia, MATLAB
- *Solvers:* CPLEX, Gurobi, MOSEK, IPOPT
- *Modeling Languages*
  - Python: Pyomo, CVXPY, RSOME
  - Julia: JuMP, Convex.jl
  - MATLAB: YALMIP, CVX, RSOME

```
import cvxpy as cp

# define variable
H = cp.Variable()
M = cp.Variable()

# define constraints
constraint = []
constraint.append(1.5*H + M <= 27)
constraint.append(H + M <= 21)
constraint.append(0.3*H + 0.5*M <= 9)
constraint.append(H <= 15)
constraint.append(M <= 16)
constraint.append(H >= 0)
constraint.append(M >= 0)

# Solve the CVXPY problem.
prob = cp.Problem(cp.Maximize(0.13*H + 0.1*M),constraint)
prob.solve()

# Print result.
print("The optimal value is", prob.value)
print("Optimal H is", H.value)
print("Optimal M is", M.value)
```

3 Quadratic Programming (QP) & Quadratically Constrained Quadratic Programming (QCQP)

3.1 QP

The standard form of **quadratic program** is the following

$$
\begin{aligned}
\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \quad & \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{P}\boldsymbol{x} + \boldsymbol{q}^\top \boldsymbol{x} + r \\
\text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \\
& \boldsymbol{C}\boldsymbol{x} \leq \boldsymbol{g}
\end{aligned}
$$

where $\boldsymbol{P} \in \mathbb{S}_+^n$ ($\boldsymbol{P}$ can always be reformulated to be within $\mathbb{S}^n$), $\boldsymbol{q} \in \mathbb{R}^n$, $r \in \mathbb{R}$, $\boldsymbol{A} \in \mathbb{R}^{m_1 \times n}$, $\boldsymbol{b} \in \mathbb{R}^{m_1}$, $\boldsymbol{C} \in \mathbb{R}^{m_2 \times n}$, $\boldsymbol{g} \in \mathbb{R}^{m_2}$.

— Example: Least-squares Problem

$$
\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \; \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2
$$

Optimality Condition (If A is full column rank):

3.2 QCQP

The standard form of **quadratically constrained quadratic program** is the following

$$
\begin{aligned}
\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \quad & \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{P}_0\boldsymbol{x} + \boldsymbol{q}_0^\top \boldsymbol{x} + r_0 \\
\text{subject to} \quad & \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{P}_i\boldsymbol{x} + \boldsymbol{q}_i^\top \boldsymbol{x} + r_i \leq 0, \quad \forall i \in [m]
\end{aligned}
$$

where $\boldsymbol{P}_i \in \mathbb{S}_+^n$, $\boldsymbol{q}_i \in \mathbb{R}^n$, $r_i \in \mathbb{R}$, $\forall i \in [m]$.

— **Example: Distance Between Ellipsoids**

- Ellipsoid 1: $\left\{\boldsymbol{x} \in \mathbb{R}^n \mid (\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) \leq 1\right\}$
- Ellipsoid 2: $\left\{\boldsymbol{x} \in \mathbb{R}^n \mid (\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) \leq 1\right\}$

3.3 Application: Portfolio Optimization and Markowitz Model

- Suppose n different risky assets are available.
- The rate of return of assets i is denoted as r_i. That is, if we invest \$$x$ in asset i, this asset will worth \$$x \cdot (1+r)$ in the next time period.
- Suppose r_i is random with expected value $\bar{r}_i$
- We denote our portfolio as $\boldsymbol{w} \in \Delta_n$

Question: What is the expected rate of return of our portfolio $\boldsymbol{w}$?

Question: What is the variance of the rate of return of our portfolio $\boldsymbol{w}$?

The Markowitz Model:

Another Formulation:

4 Second-Order Cone Programming (SOCP)

The standard form of **second-order cone program** is the following

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \quad & \boldsymbol{c}_0^\top \boldsymbol{x} \\ \text{subject to} \quad & \|\boldsymbol{A}_i\boldsymbol{x} + \boldsymbol{b}_i\|_2 \le \boldsymbol{c}_i^\top \boldsymbol{x} + d_i, \quad \forall i \in [m] \\ & \boldsymbol{F}\boldsymbol{x} = \boldsymbol{g} \end{aligned}$$

where $\boldsymbol{c}_0 \in \mathbb{R}^n$, $\boldsymbol{A}_i \in \mathbb{R}^{n_i\times n}$, $\boldsymbol{b}_i \in \mathbb{R}^{n_i}$, $\boldsymbol{c}_i \in \mathbb{R}^n$, $d_i \in \mathbb{R}$, $\boldsymbol{F} \in \mathbb{R}^{p\times n}$, $\boldsymbol{g} \in \mathbb{R}^p$, $\forall i \in [m]$.

Note: Naive approach may make it nonconvex!

4.1 Application: Robust Linear Programming

Assume that the objective/constraint functions are affine in $\boldsymbol{x}$, *i.e.*

$$
\begin{array}{ll}
\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} & \max_{\boldsymbol{z}\in\mathcal{U}} \boldsymbol{a}_0(\boldsymbol{z})^\top \boldsymbol{x} + b_0(\boldsymbol{z}) \\
\text{subject to} & \max_{\boldsymbol{z}\in\mathcal{U}} \boldsymbol{a}_i(\boldsymbol{z})^\top \boldsymbol{x} + b_i(\boldsymbol{z}) \le 0, \quad \forall i \in [m],
\end{array}
$$

where all the coefficients are affine in $\boldsymbol{z} \in \mathbb{R}^k$, *i.e.*

- $\boldsymbol{a}_i(\boldsymbol{z}) = \boldsymbol{A}_i \boldsymbol{z} + \boldsymbol{a}_i^0$ for $\boldsymbol{A}_i \in \mathbb{R}^{n\times k}$ and $\boldsymbol{a}_i^0 \in \mathbb{R}^n$
- $b_i(\boldsymbol{z}) = \boldsymbol{b}_i^\top \boldsymbol{z} + b_i^0$ for $\boldsymbol{b}_i \in \mathbb{R}^k$ and $b_i^0 \in \mathbb{R}$

What does it mean?

Suppose the uncertainty set is ellipsoidal, *i.e.*

$$
\mathcal{U} = \left\{ \boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{D}\boldsymbol{u} \mid \|\boldsymbol{u}\|_2 \le 1, \ \forall \boldsymbol{u} \in \mathbb{R}^\ell \right\},
$$

for some $\boldsymbol{\mu} \in \mathbb{R}^k$ and $\boldsymbol{D} \in \mathbb{R}^{k\times \ell}$. *How do we solve this problem?*

4.2 Hidden SOCP Constraints

Hyperbolic constraint: Consider $\boldsymbol{x} \in \mathbb{R}^n$, $s, t \in \mathbb{R}$,

$$\|\boldsymbol{x}\|_2^2 \le s \cdot t, \;\; s \ge 0, \;\; t \ge 0 \qquad \Longleftrightarrow \qquad \left\| \begin{pmatrix} 2\boldsymbol{x} \\ s - t \end{pmatrix} \right\|_2 \le s + t, \;\; s \ge 0, \;\; t \ge 0$$

Special Case – Quadratic constraint: Let $s = 1$,

$$\|\boldsymbol{x}\|_2^2 \le t, \;\; t \ge 0 \qquad \Longleftrightarrow \qquad \left\| \begin{pmatrix} 2\boldsymbol{x} \\ t - 1 \end{pmatrix} \right\|_2 \le t + 1, \;\; t \ge 0$$

4.3 Equivalence

- Every LP is an SOCP (with $\boldsymbol{A}_i = \boldsymbol{0}, \;\; \boldsymbol{b}_i = \boldsymbol{0}$)
- Every SOCP with $\boldsymbol{c}_i = \boldsymbol{0}$ is a QCQP.
- Every QCQP is an SOCP.