CS257 Linear and Convex Optimization

Lecture 7

Bo Jiang

John Hopcroft Center for Computer Science
Shanghai Jiao Tong University

October 19, 2020

Recap: Convex Optimization Problem

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & g_i(\boldsymbol{x}) \le 0, \quad i = 1, 2, \ldots, m \\ & h_i(\boldsymbol{x}) = 0, \quad i = 1, 2, \ldots, k \end{aligned}$$

1. f, g_i are convex functions
2. h_i are affine functions, i.e. $h_i(\boldsymbol{x}) = \boldsymbol{a}_i^T \boldsymbol{x} - b_i$

Domain. $D = \operatorname{dom} f \cap \left(\bigcap_{i=1}^m \operatorname{dom} g_i\right)$

Feasible set. $X = \{\boldsymbol{x} \in D : g_i(\boldsymbol{x}) \le 0, 1 \le i \le m; h_i(\boldsymbol{x}) = 0, 1 \le i \le k\}$

Optimal value. $f^* = \inf_{\boldsymbol{x} \in X} f(\boldsymbol{x})$

Optimal point. $x^* \in X$ and $f(\boldsymbol{x}^*) = f^*$, i.e. $f(\boldsymbol{x}^*) \le f(\boldsymbol{x}), \forall \boldsymbol{x} \in X$

First-order optimality condition

$$\nabla f(\boldsymbol{x}^*)^T (\boldsymbol{x} - \boldsymbol{x}^*) \ge 0, \quad \forall \boldsymbol{x} \in X$$

Recap: LP

General form

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \boldsymbol{c}^T\boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{B}\boldsymbol{x} \leq \boldsymbol{d} \\ & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \end{aligned}$$

Standard form

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

Inequality form

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \boldsymbol{c}^T\boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \end{aligned}$$

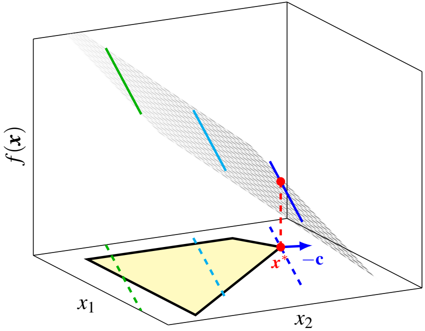
Conversion to equivalent problems

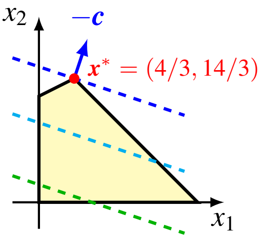
- introducing slack variables
- eliminating equality constraints
- epigraph form
- representing a variable by two nonnegative variables, $x = x^+ - x^-$

Recap: Geometry of LP

$$\min_{\boldsymbol{x}} \quad \boldsymbol{c}^T\boldsymbol{x}$$
$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$$

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & -x_1 - 3x_2 \\ \text{s.t.} \quad & x_1 + x_2 \leq 6 \\ & -x_1 + 2x_2 \leq 8 \\ & x_1, x_2 \geq 0 \end{aligned}$$





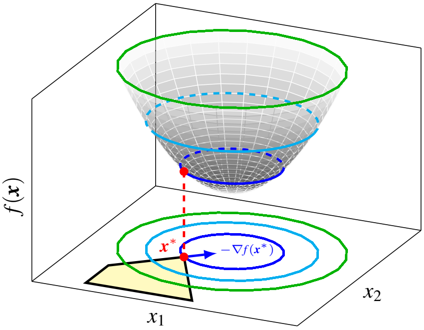
- optimization of a linear function over a polyhedron
- graphic solution of simple LP

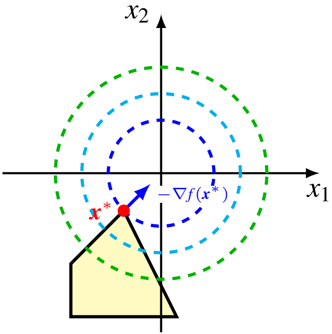
Contents

Quadratic Program (QP)

$$\begin{aligned}\min_{\boldsymbol{x}} \quad & \frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q}\boldsymbol{x} + \boldsymbol{c}^T\boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{B}\boldsymbol{x} \le \boldsymbol{d} \\ & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\end{aligned}$$

QP is convex iff $\boldsymbol{Q} \succeq \boldsymbol{O}$.

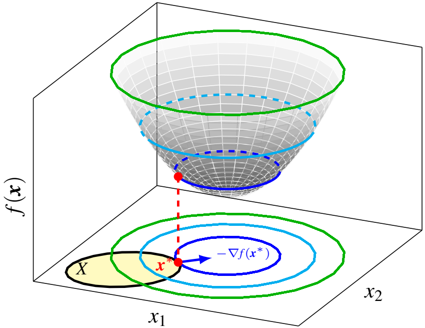


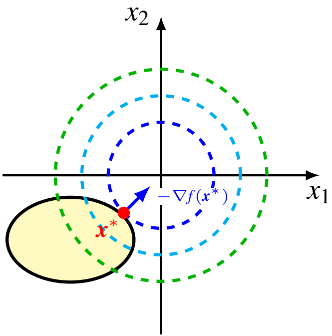


Quadratically Constrained Quadratic Program (QCQP)

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{c}^T\boldsymbol{x} \\ \text{s.t.} \quad & \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}_i\boldsymbol{x} + \boldsymbol{c}_i^T\boldsymbol{x} + \boldsymbol{d}_i \le 0, \quad i = 1, 2, \ldots, m \\ & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \end{aligned}$$

QCQP is convex if $\boldsymbol{Q} \succeq \boldsymbol{O}$ and $\boldsymbol{Q}_i \succeq \boldsymbol{O}$, $\forall i$





Example: Linear Least Squares Regression

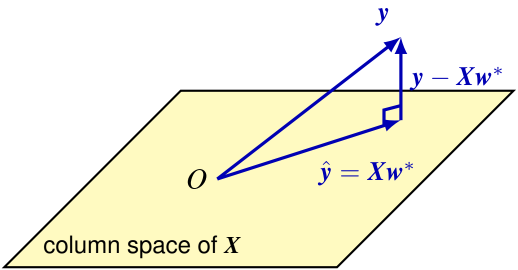
Given $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, find $\boldsymbol{w} \in \mathbb{R}^p$ s.t.

$$\min_{\boldsymbol{w}} \ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2$$

- convex QP with objective

$$f(\boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{y}^T \boldsymbol{X}\boldsymbol{w} + \boldsymbol{y}^T \boldsymbol{y}$$

Geometrically, we are looking for the orthogonal projection $\hat{\boldsymbol{y}}$ of $\boldsymbol{y}$ onto the column space of $\boldsymbol{X}$.



Example: Linear Least Squares Regression (cont'd)

By the first-order optimality condition, $\boldsymbol{w}^*$ is optimal iff

$$\nabla f(\boldsymbol{w}^*) = \boldsymbol{0}$$

i.e. $\boldsymbol{w}^*$ is a solution of the normal equation,

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y}$$

Case I. $\boldsymbol{X}$ has full column rank, i.e. $\operatorname{rank}\boldsymbol{X} = p$

- $\boldsymbol{X}^T\boldsymbol{X} \succ \boldsymbol{O}$
- unique solution

$$\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Note. In this case, the objective $f(\boldsymbol{w})$ is strictly convex and coercive.

$$f(\boldsymbol{w}) \geq \lambda_{\min}(\boldsymbol{X}^T\boldsymbol{X})\|\boldsymbol{w}\|^2 - 2\|\boldsymbol{y}^T\boldsymbol{X}\| \cdot \|\boldsymbol{w}\| + \|\boldsymbol{y}\|^2$$

Example: Linear Least Squares Regression (cont'd)

Example. Solve

$$\min_{\boldsymbol{w}} \ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2$$

with

$$\boldsymbol{X} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} 3 \\ 2 \\ 2 \end{pmatrix}.$$

Solution. The normal equation is

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y}$$

with

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}, \quad \boldsymbol{X}^T\boldsymbol{y} = (6, 2)^T$$

Since $\boldsymbol{X}$ has full column rank,

$$\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = (1.5, 2)^T$$

Example: Linear Least Squares Regression (cont'd)

Case II. $\operatorname{rank} \boldsymbol{X} = r < p$. WLOG assume the first r columns are linearly independent, i.e.

$$\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$$

where $\boldsymbol{X}_1 \in \mathbb{R}^{n \times r}$ and $\operatorname{rank} \boldsymbol{X}_1 = r$.

Claim. There is a solution $\boldsymbol{w}^*$ with the last $p - r$ components being 0.

- $\boldsymbol{X}$ and $\boldsymbol{X}_1$ have the same column space
- If $\boldsymbol{w}_1^*$ solves

$$\min_{\boldsymbol{w}_1 \in \mathbb{R}^r} \|\boldsymbol{y} - \boldsymbol{X}_1 \boldsymbol{w}_1\|$$

  then $\boldsymbol{w}^* = \begin{pmatrix} \boldsymbol{w}_1^* \\ \boldsymbol{0} \end{pmatrix}$ solves $\min_{\boldsymbol{w} \in \mathbb{R}^p} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|$
- $\boldsymbol{w}_1^* = (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \boldsymbol{y}$

Question. Is the solution unique in this case?

A. $\operatorname{rank} \boldsymbol{X} < p \implies \exists \boldsymbol{w}_0$ s.t. $\boldsymbol{X}\boldsymbol{w}_0 = \boldsymbol{0} \implies \boldsymbol{w}^* + \boldsymbol{w}_0$ is also a solution.

Example: Linear Least Squares Regression (cont'd)

Example Solve $\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2$ with

$$\boldsymbol{X} = \begin{pmatrix} 2 & 0 & 2 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} 3 \\ 2 \\ 2 \end{pmatrix}.$$

Solution. Note $\operatorname{rank} \boldsymbol{X} = 2 < 3$.

- Let

$$\boldsymbol{X}_1 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$$

- By the previous example,

$$\boldsymbol{w}_1^* = (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \boldsymbol{y} = (1.5, 2)^T$$

  is a solution to $\min_{\boldsymbol{w}_1 \in \mathbb{R}^2} \|\boldsymbol{y} - \boldsymbol{X}_1 \boldsymbol{w}_1\|^2$.
- $\boldsymbol{w}^* = (1.5, 2, 0)^T$ is a solution to $\min_{\boldsymbol{w} \in \mathbb{R}^3} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2$.

Example: Linear Least Squares Regression (cont'd)

Example (cont'd). The normal equation to the original problem is

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y}$$

where

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} 4 & 0 & 4 \\ 0 & 1 & -1 \\ 4 & -1 & 5 \end{pmatrix}, \quad \boldsymbol{X}^T\boldsymbol{y} = (6, 2, 4)^T$$

- Note $\boldsymbol{X}^T\boldsymbol{X}$ is not invertible, so we cannot use the formula[1] $\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$
- The solution $\boldsymbol{w}^* = (1.5, 2, 0)^T$ satisfies the normal equation.
- The normal equation has infinitely many solutions given by

$$\boldsymbol{w} = (1.5, 2, 0)^T + \alpha(-1, 1, 1)^T, \quad \alpha \in \mathbb{R}.$$

All of them are solutions to the least squares problem.

[1] This formula still applies if we use the so-called pseudo inverse of $\boldsymbol{X}^T\boldsymbol{X}$.

General Unconstrained QP

Minimize quadratic function with $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ s.t. $\boldsymbol{Q} \succeq \boldsymbol{O}$,

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} + c$$

By first-order condition, solution satisfies

$$\nabla f(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b} = \boldsymbol{0}$$

Case I. $\boldsymbol{Q} \succ \boldsymbol{O}$. There is a unique solution $\boldsymbol{x}^* = -\boldsymbol{Q}^{-1}\boldsymbol{b}$.

Example. $n = 2$, $\boldsymbol{Q} =$ diag$\{1, 1\}$, $\boldsymbol{b} = (1, 0)^T$, $c = 0$.

$$f(\boldsymbol{x}) = \frac{1}{2}(x_1, x_2)\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix} + (1, 0)\begin{pmatrix}x_1\\ x_2\end{pmatrix} = \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2 + x_1$$

The first-order condition becomes

$$\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix} + \begin{pmatrix}1\\ 0\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix}$$

which yields the unique optimal solution $\boldsymbol{x}^* = (-1, 0)$.

General Unconstrained QP (cont'd)

Case II. $\det \boldsymbol{Q} = 0$ and $\boldsymbol{b} \notin$ column space of $\boldsymbol{Q}$. There is no solution, and $f^* = -\infty$.

Example. $n = 2$, $\boldsymbol{Q} =$ diag$\{0, 1\}$, $\boldsymbol{b} = (1, 0)^T$, $c = 0$.

$$f(\boldsymbol{x}) = \frac{1}{2}(x_1, x_2)\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + (1, 0)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{2}x_2^2 + x_1$$

The first-order condition becomes

$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

which has no solution.

It is easy to see that $f(\boldsymbol{x}) = \frac{1}{2}x_2^2 + x_1$ is unbounded below, so $f^* = -\infty$.

General Unconstrained QP (cont'd)

Case III. $\det \boldsymbol{Q} = 0$ and $\boldsymbol{b} \in$ column space of $\boldsymbol{Q}$. There are infinitely many solutions.

Example. $n = 2$, $\boldsymbol{Q} = \text{diag}\{1, 0\}$, $\boldsymbol{b} = (1, 0)^T$, $c = 0$.

$$f(\boldsymbol{x}) = \frac{1}{2}(x_1, x_2) \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + (1, 0) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{2}x_1^2 + x_1$$

The first-order condition becomes

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

which has infinitely many solutions of the form $\boldsymbol{x} = (-1, x_2)$ for any $x_2 \in \mathbb{R}^2$, as f is actually independent of x_2.

General Unconstrained QP (cont'd)

For the general case ($\boldsymbol{Q}$ is non-diagonal),

- Diagonalize $\boldsymbol{Q}$ by an orthogonal matrix $\boldsymbol{U}$, so

$$\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$$

  where $\boldsymbol{\Lambda}$ is diagonal.

- Let $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{y}$ and $\tilde{\boldsymbol{b}} = \boldsymbol{U}^T\boldsymbol{b}$. Then

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{y}^T\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\boldsymbol{y} + \boldsymbol{b}^T\boldsymbol{U}\boldsymbol{y} + c = \frac{1}{2}\boldsymbol{y}^T\boldsymbol{\Lambda}\boldsymbol{y} + \tilde{\boldsymbol{b}}^T\boldsymbol{y} + c \triangleq g(\boldsymbol{y})$$

  In the expanded form,

$$g(\boldsymbol{y}) = \sum_{i=1}^{n}\left(\frac{1}{2}\lambda_i y_i^2 + \tilde{b}_i y_i\right) + c$$

- Minimizing $f(\boldsymbol{x})$ is equivalent to minimizing $g(\boldsymbol{y})$. We can minimize each term $\frac{1}{2}\lambda_i y_i^2 + \tilde{b}_i y_i$ independently.

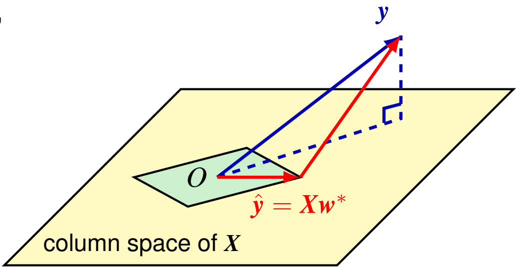
Exercise. Convince yourself the previous three cases apply to the non-diagonal case.

Example: Lasso

Lasso (Least Absolute Shrinkage and Selection Operator)

Given $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $t > 0$,

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 \\ \text{s.t.} \quad & \|\boldsymbol{w}\|_1 \le t \end{aligned}$$

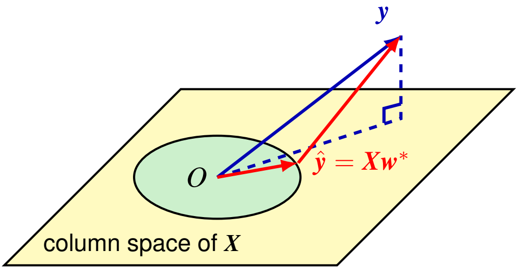


- convex problem? yes
- QP? no, but can be converted to QP
- optimal solution exists? yes
  - compact feasible set
- optimal solution unique?
  - yes if $n \ge p$ and $\boldsymbol{X}$ has full column rank ($\boldsymbol{X}^T\boldsymbol{X} \succ \boldsymbol{O}$, strictly convex)
  - no in general, e.g. $p > n$ and t is large enough for unconstrained optima to be feasible

Example: Ridge Regression

Given $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $t > 0$,

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 \\ \text{s. t.} \quad & \|\boldsymbol{w}\|_2^2 \le t \end{aligned}$$



- convex problem? yes
- QCQP? yes
- optimal solution exists? yes
  - compact feasible set
- optimal solution unique?
  - yes if $n \ge p$ and $\boldsymbol{X}$ has full column rank ($\boldsymbol{X}^T\boldsymbol{X} \succ \boldsymbol{O}$, strictly convex)
  - no in general

Example: SVM

Linearly separable case

$$
\begin{aligned}
\min_{\boldsymbol{w},b} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 \\
\text{s. t.} \quad & y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1, \quad i = 1, 2, \ldots, m
\end{aligned}
$$

Soft margin SVM

$$
\begin{aligned}
\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i \\
\text{s. t.} \quad & y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i = 1, 2, \ldots, m \\
& \boldsymbol{\xi} \geq \mathbf{0}
\end{aligned}
$$

Equivalent unconstrained form

$$
\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{n}(1 - y_i b - y_i\boldsymbol{w}^T\boldsymbol{x}_i)^+
$$

Geometric Program

A monomial is a function $f : \mathbb{R}^n_{++} = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x} > \boldsymbol{0}\} \to \mathbb{R}$ of the form

$$f(\boldsymbol{x}) = \gamma x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}$$

for $\gamma > 0$, $a_1, \ldots, a_n \in \mathbb{R}$. A posynomial is a sum of monomials,

$$f(\boldsymbol{x}) = \sum_{k=1}^{p} \gamma_k x_1^{a_{k1}} x_2^{a_{k2}} \cdots x_n^{a_{kn}}$$

A geometric program (GP) is an optimization problem of the form

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s. t.} \quad & g_i(\boldsymbol{x}) \leq 1, \quad i = 1, \ldots, m \\ & h_j(\boldsymbol{x}) = 1, \quad j = 1, \ldots, r \end{aligned}$$

where $f, g_i, i = 1, \ldots, m$ are posynomials and $h_j, j = 1, \ldots, r$ are monomials. The constraint $\boldsymbol{x} > \boldsymbol{0}$ is implicit.

Geometric Program (cont'd)

GP is nonconvex (why?)

$$
\begin{aligned}
\min_{\boldsymbol{x}} \quad & \sum_{k=1}^{p_0} \gamma_{0k} x_1^{a_{0k1}} x_2^{a_{0k2}} \cdots x_n^{a_{0kn}} \\
\text{s.t.} \quad & \sum_{k=1}^{p_i} \gamma_{ik} x_1^{a_{ik1}} x_2^{a_{ik2}} \cdots x_n^{a_{ikn}} \leq 1, \quad i = 1, \ldots, m \\
& \eta_j x_1^{c_{j1}} x_2^{c_{j2}} \cdots x_n^{c_{jn}} = 1, \quad j = 1, \ldots, r
\end{aligned}
$$

nonconvex
f(x)=x_1x_2^2

By $y_i = \log x_i$, $b_{ik} = \log \gamma_{ik}$, $d_j = \log \eta_j$, GP can be formulated as

$$
\begin{aligned}
\min_{\boldsymbol{y}} \quad & \log\left(\sum_{k=1}^{p_0} e^{\boldsymbol{a}_{0k}^T \boldsymbol{y} + b_{0k}}\right) \\
\text{s.t.} \quad & \log\left(\sum_{k=1}^{p_i} e^{\boldsymbol{a}_{ik}^T \boldsymbol{y} + b_{ik}}\right) \leq 0, \quad i = 1, \ldots, m \\
& \boldsymbol{c}_j^T \boldsymbol{y} + d_j = 0, \quad j = 1, \ldots, r
\end{aligned}
$$

This is convex by the convexity of log-sum-exp (soft max) functions