Mathematical Foundations: Optimization Primer

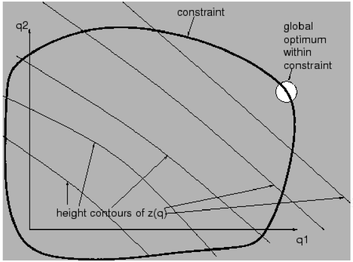
Prof. Ziping Zhao

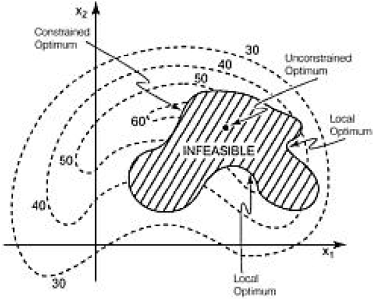
School of Information Science and Technology
ShanghaiTech University, Shanghai, China

CS182: Introduction to Machine Learning (Fall 2021)
http://cs182.sist.shanghaitech.edu.cn

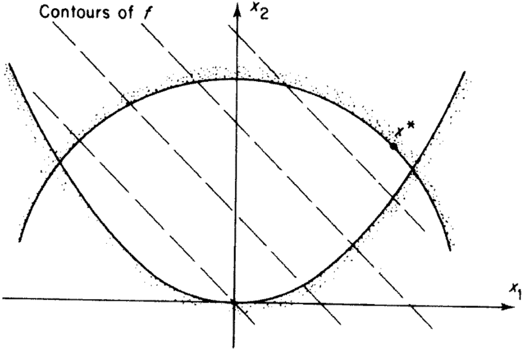
Optimization Problem

$$\begin{array}{lll} \underset{\mathbf{x}}{\text{minimize}} & f_0(\mathbf{x}) & \text{(objective function)} \\ \text{subject to} & f_i(\mathbf{x}) \leq 0, \ \ i = 1, \ldots, m & \text{(inequality constraints)} \\ & h_i(\mathbf{x}) = 0, \ \ i = 1, \ldots, p & \text{(equality constraints)} \end{array}$$





Active Constraint



A constraint is active at $\mathbf{x}$

- $\mathbf{x}$ is on the boundary of its feasible region ($f_i(\mathbf{x}) = 0$)

$\mathcal{A}^*$: set of active constraints at the solution. The remaining constraints can be ignored and the problem can be treated as an equality constraint problem with constraints $\mathcal{A}^*$.

Lagrangian

standard form problem (without equality constraints)

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & f_0(\mathbf{x}) \\ \text{subject to} & f_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, m \end{array}$$

- primal problem
- optimal value p^*

(assume $\mathbf{x} \in \mathbb{R}^n$) Lagrangian $\mathcal{L} : \mathbb{R}^{n+m} \to \mathbb{R}$

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f_0(\mathbf{x}) + \lambda_1 f_1(\mathbf{x}) + \ldots + \lambda_m f_m(\mathbf{x}) = f_0(\mathbf{x}) + \sum_i \lambda_i f_i(\mathbf{x})$$

- λ_i: Lagrange multipliers or dual variables
- objective is augmented with weighted sum of constraint functions

Lagrange Dual Function

(Lagrange) dual function $g : \mathbb{R}^m \to \mathbb{R} \cup \{-\infty\}$

$$g(\boldsymbol{\lambda}) = \inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \inf_{\mathbf{x}} (f_0(\mathbf{x}) + \lambda_1 f_1(\mathbf{x}) + \cdots + \lambda_m f_m(\mathbf{x}))$$

- minimum of augmented cost as function of weights
- can be $-\infty$ for some $\boldsymbol{\lambda}$

Example: linear programming (LP)

$$\operatorname*{minimize}_{\mathbf{x}} \quad \mathbf{c}^T\mathbf{x} \quad \text{subject to} \quad \mathbf{a}_i^T\mathbf{x} - b_i \leq 0, \quad i = 1, \ldots, m$$

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{c}^T\mathbf{x} + \sum_{i=1}^{m} \lambda_i(\mathbf{a}_i^T\mathbf{x} - b_i) = -\mathbf{b}^T\boldsymbol{\lambda} + (\mathbf{A}\boldsymbol{\lambda} + \mathbf{c})^T\mathbf{x}$$

$$g(\boldsymbol{\lambda}) = \begin{cases} -\mathbf{b}^T\boldsymbol{\lambda} & \text{if } \mathbf{A}\boldsymbol{\lambda} + \mathbf{c} = \mathbf{0} \\ -\infty & \text{otherwise} \end{cases}$$

Lower Bound Property

Property

If $\boldsymbol{\lambda} \geq 0$ and $\mathbf{x}$ is primal feasible, then $g(\boldsymbol{\lambda}) \leq f_0(\mathbf{x})$

Proof.

if $f_i(\mathbf{x}) \leq 0$ and $\lambda_i \geq 0$ for $i = 1, \ldots, m$,

$$\begin{aligned} f_0(\mathbf{x}) &\geq f_0(\mathbf{x}) + \sum_i \lambda_i f_i(\mathbf{x}) \\ &\geq \inf_{\mathbf{z}} \left(f_0(\mathbf{z}) + \sum_i \lambda_i f_i(\mathbf{z}) \right) \\ &= g(\boldsymbol{\lambda}) \end{aligned}$$

$\square$

- $f_0(\mathbf{x}) - g(\boldsymbol{\lambda}) \geq 0$: duality gap of (primal feasible) $\mathbf{x}$ and $\boldsymbol{\lambda} \geq 0$
- $\boldsymbol{\lambda} \in \mathbb{R}^m$ is dual feasible if $\boldsymbol{\lambda} \geq 0$ and $g(\boldsymbol{\lambda}) > -\infty$
- minimize $f_0(\mathbf{x}) - g(\boldsymbol{\lambda}) \geq 0$ over primal feasible $\mathbf{x}$

$$\text{for any } \boldsymbol{\lambda} \geq 0, g(\boldsymbol{\lambda}) \leq p^*$$

  - dual feasible points yield lower bounds on optimal value!

Lagrange Dual Problem

Find the best lower bound on p^*:

$$\begin{array}{ll} \underset{\boldsymbol{\lambda}}{\text{maximize}} & g(\boldsymbol{\lambda}) \\ \text{subject to} & \boldsymbol{\lambda} \geq 0 \end{array}$$

- (Lagrange) dual problem (associated with the primal problem)
- optimal value: d^*
- we always have $d^* \leq p^*$ (weak duality)
- $p^* - d^*$: optimal duality gap
- for convex problems, we (usually) have strong duality (i.e., zero duality gap):

$$d^* = p^*$$

Dual of Linear Program

primal

$$\begin{aligned} \text{minimize}_{\mathbf{x}} \quad & \mathbf{c}^T \mathbf{x} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} \leq \mathbf{b} \end{aligned}$$

- n variables, m inequality constraints

dual

$$\begin{aligned} \text{maximize}_{\boldsymbol{\lambda}} \quad & -\mathbf{b}^T \boldsymbol{\lambda} \\ \text{subject to} \quad & \mathbf{A}^T \boldsymbol{\lambda} + \mathbf{c} = 0 \\ & \boldsymbol{\lambda} \geq 0 \end{aligned}$$

- dual of LP is also an LP
- m variables, n equality constraints, m nonnegativity constraints

Duality in Algorithms

many algorithms produce at iteration k

- a primal feasible $\mathbf{x}^{(k)}$
- and a dual feasible $\boldsymbol{\lambda}^{(k)}$

with $f_0(\mathbf{x}^{(k)}) - g(\boldsymbol{\lambda}^{(k)}) \to 0$ as $k \to 0$

- hence at iteration k we know $p^* \in [g(\boldsymbol{\lambda}^{(k)}), f_0(\mathbf{x}^{(k)})]$
- useful for stopping criteria

Complementary Slackness

suppose $\mathbf{x}^*, \boldsymbol{\lambda}^*$ are primal, dual optimal with zero duality gap

$$\begin{aligned} f_0(\mathbf{x}^*) &= g(\boldsymbol{\lambda}^*) \\ &= \inf_{\mathbf{x}} \left(f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i^* f_i(\mathbf{x}) \right) \\ &\leq f_0(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* f_i(\mathbf{x}^*) \end{aligned}$$

hence we have $\sum_{i=1}^{m} \lambda_i^* f_i(\mathbf{x}^*) = 0$, and so

complementary slackness condition

$$\lambda_i^* f_i(\mathbf{x}^*) = 0, \quad i = 1, \ldots, m$$

- ith constraint inactive at optimum $\Rightarrow \lambda_i^* = 0$
- $\lambda_i^* > 0$ at optimum $\Rightarrow$ ith constraint active at optimum

KKT Optimality Conditions

suppose

- f_i are differentiable
- $\mathbf{x}^*, \boldsymbol{\lambda}^*$ are (primal, dual) optimal, with zero duality gap

by complementary slackness we have (from previous slide)

$$f_0(\mathbf{x}^*) + \sum_i \lambda_i^* f_i(\mathbf{x}^*) = \inf_{\mathbf{x}} \left(f_0(\mathbf{x}) + \sum_i \lambda_i^* f_i(\mathbf{x}) \right)$$

- i.e., $\mathbf{x}^*$ minimizes $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*)$ ($\therefore \nabla f_0(\mathbf{x}^*) + \sum_i \lambda_i^* \nabla f_i(\mathbf{x}^*) = 0$)

Karush-Kuhn-Tucker (KKT) optimality conditions:

$$\begin{aligned} f_i(\mathbf{x}^*) &\leq 0 \\ \lambda_i^* &\geq 0 \\ \lambda_i^* f_i(\mathbf{x}^*) &= 0 \\ \nabla f_0(\mathbf{x}^*) + \sum_i \lambda_i^* \nabla f_i(\mathbf{x}^*) &= 0 \end{aligned}$$

Equality Constraints

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & f_0(\mathbf{x}) \\ \text{subject to} \quad & f_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, m \\ & h_i(\mathbf{x}) = 0, \quad i = 1, \ldots, p \end{aligned}$$

define Lagrangian $\mathcal{L} : \mathbb{R}^{n+m+p} \to \mathbb{R}$ as

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x})$$

dual function: $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$

- $(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is dual feasible if $\boldsymbol{\lambda} \geq 0$ and $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) > -\infty$
- No sign condition on $\boldsymbol{\nu}$

lower bound property: if $\mathbf{x}$ is primal feasible and $(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is dual feasible, then $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f_0(\mathbf{x})$, hence

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^*$$

dual problem: find best lower bound

$$\begin{array}{ll} \underset{\boldsymbol{\lambda},\, \boldsymbol{\nu}}{\text{maximize}} & g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\ \text{subject to} & \boldsymbol{\lambda} \geq 0 \end{array}$$

- note: $\boldsymbol{\nu}$ unconstrained

weak duality: $d^* \leq p^*$ always
strong duality: if primal is convex then (usually) $d^* = p^*$

KKT Optimality Conditions

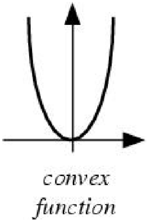
assume f_i, h_i differentiable

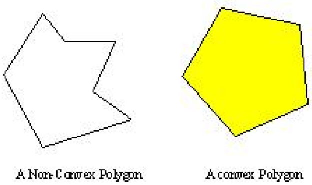
if $\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*$ are optimal, with zero duality gap, then they satisfy KKT conditions

$$
\begin{aligned}
f_i(\mathbf{x}^*) \leq 0, \quad h_i(\mathbf{x}^*) &= 0 \\
\lambda_i^* &\geq 0 \\
\lambda_i^* f_i(\mathbf{x}^*) &= 0 \\
\nabla f_0(\mathbf{x}^*) + \sum_i \lambda_i^* \nabla f_i(\mathbf{x}^*) + \sum_i \nu_i^* \nabla h_i(\mathbf{x}^*) &= 0
\end{aligned}
$$

Convex Programming

minimize a convex function on a convex set





Covex Functions & Sets

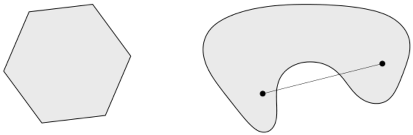
- Convex function: A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be convex if the domain, $\operatorname{dom} f$, is convex and

$$f(\theta \mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y})$$

  for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$, $0 \leq \theta \leq 1$
- Convex set: A set $C \in \mathbb{R}^n$ is said to be convex if the line segment between any two points is in the set:

$$\theta \mathbf{x} + (1-\theta)\mathbf{y} \in C$$

  for all $\mathbf{x}, \mathbf{y} \in C$, $0 \leq \theta \leq 1$



Key Properties of Convex Functions

- α-sublevel set: sublevel sets of a convex function f are convex (converse is false)

$$C_\alpha = \{\mathbf{x} \in \text{dom } f \mid f(\mathbf{x}) \leq \alpha\}$$

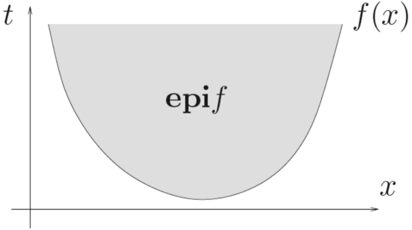
- Epigraph: a function f is convex if and only if its epigraph

$$\text{epi } f = \{(\mathbf{x}, t) \in \mathbb{R}^{n+1} \mid \mathbf{x} \in \text{dom } f, f(\mathbf{x}) \leq t\}$$

  is a convex set

- Relation between convexity in sets and convexity in functions:

$$f \text{ is convex} \iff \text{epi } f \text{ is convex}$$



- First-order condition: a differentiable f with convex domain is convex if and only if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) \quad \forall \mathbf{x}, \mathbf{y} \in \text{dom } f$$

- Second-order condition: a twice differentiable f with convex domain is convex if and only if

$$\nabla^2 f(\mathbf{x}) = \mathbf{H}(\mathbf{x}) \succeq \mathbf{0} \quad \forall \mathbf{x} \in \text{dom } f$$

- Jensen's inequality: if f is convex, and X is a random variable supported on dom f, then

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

Examples

Linear programming

- linear objective function, linear constraints

$$\underset{\mathbf{x}}{\text{minimize}} \quad \mathbf{c}^T\mathbf{x} \quad \text{subject to} \quad \mathbf{a}_i^T\mathbf{x} - b_i \leq 0, \quad i = 1, \ldots, m$$

Quadratic programming (QP)

- quadratic objective function, linear constraints

$$\underset{\mathbf{x}}{\text{minimize}} \quad \frac{1}{2}\mathbf{x}^T\mathbf{G}\mathbf{x} + \mathbf{g}^T\mathbf{x}$$

$$\text{subject to} \quad \begin{cases} \mathbf{a}_i^T\mathbf{x} - b_i = 0 \\ \mathbf{a}_i^T\mathbf{x} - b_i \leq 0 \end{cases}$$

- **G**: (positive semi-definite) matrix; **g**: vector

Global Optimality

Every local solution is a global solution

- does not have the problem of local optimum