Mathematical Foundations: Optimization Primer

Ziping Zhao

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

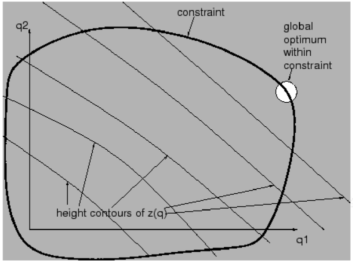
CS182: Introduction to Machine Learning (Fall 2022)
http://cs182.sist.shanghaitech.edu.cn

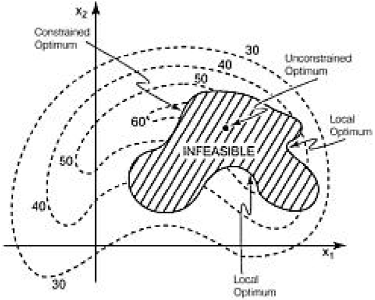
App. C of I2ML

Optimization Problem

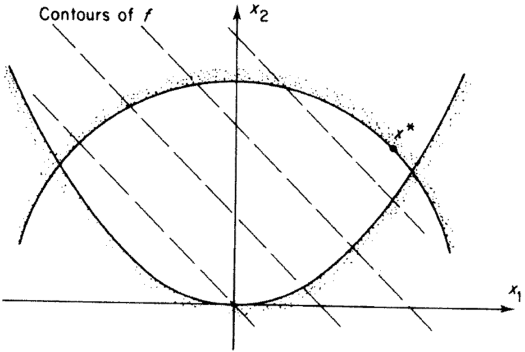
standard form problem

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





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, which can be considered as "costs" of violating the corresponding constraints
- 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) (inequality form)

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

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

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

Lower Bound Property

Property

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

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

  - dual feasible points yield lower bounds on optimal value!

Lagrange Dual Problem

Find the best lower bound on p^*:

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

- (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 A Linear Programming

primal

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

- n variables, m inequality constraints

dual

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

- 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 $k \to \infty$ (for convex optimization problems)

- 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_0 and 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:

$$\begin{aligned} f_i(\mathbf{x}^*) &\leq 0 \quad \text{(primal feasibility)} \\ \lambda_i^* &\geq 0 \quad \text{(dual feasibility)} \\ \lambda_i^* f_i(\mathbf{x}^*) &= 0 \quad \text{(complementary)} \\ \nabla f_0(\mathbf{x}^*) + \sum_i \lambda_i^* \nabla f_i(\mathbf{x}^*) &= 0 \quad \text{(stationarity)} \end{aligned}$$

Equality Constraints

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

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 $\boldsymbol{\lambda} \ge 0$ 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

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

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

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

KKT Optimality Conditions

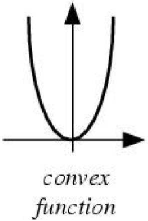
assume f_i, h_i differentiable

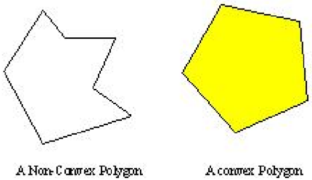
if $\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*$ are optimal, with zero duality gap, then they satisfy KKT conditions

$$
\begin{aligned}
f_i(\mathbf{x}^*) \le 0, \quad h_i(\mathbf{x}^*) &= 0 && \text{(primal feasibility)} \\
\lambda_i^* &\ge 0 && \text{(dual feasibility)} \\
\lambda_i^* f_i(\mathbf{x}^*) &= 0 && \text{(complementary)} \\
\nabla f_0(\mathbf{x}^*) + \sum_i \lambda_i^* \nabla f_i(\mathbf{x}^*) + \sum_i \nu_i^* \nabla h_i(\mathbf{x}^*) &= 0 && \text{(stationarity)}
\end{aligned}
$$

Convex Optimization

Convex optimization (or convex programming): minimize a convex function on a convex set





Convex Sets & Functions

- Convex set: A set $\mathcal{C} \in \mathbb{R}^n$ is said to be convex if the line segment between any two points is in the set:
$$\theta \mathbf{x} + (1-\theta)\mathbf{y} \in \mathcal{C}$$
for all $\mathbf{x}, \mathbf{y} \in \mathcal{C}$, $0 \leq \theta \leq 1$
- in convex optimization, equality constraints are affine
- Convex function: A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is said to be convex if the domain, 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 $\mathbf{x}, \mathbf{y} \in \text{dom } f$, $0 \leq \theta \leq 1$
- f is concave if $-f$ is convex
- if f is a convex function, then $\mathcal{C} = \{\mathbf{x} \mid f(\mathbf{x}) \leq 0\}$ is a convex set

- 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

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

- Pointwise supremum: if $f(x, y)$ is convex in x for each $y \in \mathcal{A}$, then

$$g(x) = \sup_{y \in \mathcal{A}} f(x, y)$$

is convex.

Examples of Convex Optimization Problem - Linear Programming

Linear programming (LP) (or linear program, linear optimization)

- affine objective function, affine constraints

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

- LP generally does not have an analytical solution, but we have efficient methods, such as the simplex method, to find the solution in reasonable time. LP solvers are frequently used in a variety of applications.

Examples of Convex Optimization Problem - Quadratic Programming

Quadratic programming (QP) (or quadratic program, quadratic optimization)

- quadratic objective function, affine constraints

$$
\begin{array}{lll}
\underset{\mathbf{x}}{\text{minimize}} & \frac{1}{2}\mathbf{x}^T\mathbf{G}\mathbf{x} + \mathbf{c}^T\mathbf{x} & \\
\text{subject to} & \mathbf{a}_i^T\mathbf{x} - b_i = 0, & i = 1, \dots, m \\
 & \mathbf{d}_i^T\mathbf{x} - e_i \leq 0, & i = 1, \dots, p
\end{array}
$$

- the QP is convex if **G** is positive semi-definite

Examples of Convex Optimization Problem - Lagrange Dual Problem

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

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

- g is concave (pointwise infimum of affine functions), even when the primal problem is not convex
- (Lagrange) dual problem:

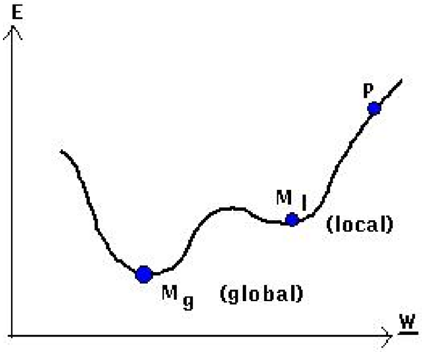
$$\begin{aligned} \underset{\boldsymbol{\lambda},\, \boldsymbol{\nu}}{\text{maximize}} \quad & g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\ \text{subject to} \quad & \boldsymbol{\lambda} \geq \mathbf{0} \end{aligned}$$

- It is a convex optimization problem (maximization of a concave function and affine constraints).

Global Optimality

In convex optimization, every local solution is a global solution

- does not have the problem of local optimum
- If we know a problem is convex, we know that we can solve it optimally. But, solving it may be iterative and rather costly in terms of memory and/or computation.



Local Optimization

- If the problem is not convex, there is no method that guarantees us to find the globally optimal solution in reasonable time.
- Non-convex optimization is NP-hard.
- The usual approach in such a case is local optimization, where we look for a locally optimal solution, which is known to be best in a local region, but it is not guaranteed to be best among all feasible points.
- Typically, we start at some initial value of the parameters and iteratively update the variables based on an algorithm until it reaches some stopping criterion (optimality condition or stationarity condition).

Gradient Descent Algorithm

- If the objective f_0 is differentiable, we can use the gradient information (first-order derivatives) to help us in finding the direction to update the parameters $\mathbf{x}$.
- In a minimization problem, with gradient descent, at iteration t, we update $\mathbf{x}$ in the negative direction of the gradient:

$$\mathbf{x}^{(t)} = \mathbf{x}^{(t-1)} - \eta^{(t)} \nabla f(\mathbf{x}^{(t-1)})$$

  where $\eta^{(t)}$ is called the step size at iteration t, which defines how far to go in the negative gradient direction.
- We stop when we get to a minimum, where the gradient is zero.
  - Numerically, we can set $\left\|\nabla f(\mathbf{x}^{(t)})\right\| \leq \epsilon$.
- Starting from a randomly chosen $\mathbf{x}^{(0)}$, we converge to the nearest local minimum.
- In second-order methods, we also use the second derivatives, and they allow faster convergence because they also use the curvature information.

Numerical (Computational) Optimization

- enumeration method
- direct method
- iterative method

- the efficiency of an iterative algorithm
  - the number of iterations required, i.e, convergence speed
    - global convergence
    - local convergence rate
    - global convergence rate (worst case complexity)
  - arithmetic operations (flop) per iteration