Mathematics, Optimization, Statistics, and Machine Learning

Sunghee Yun

January 20, 2020

Contents

Part I

Mathematics

Chapter 1

Calculus

Chapter 2

Convex analysis

Chapter 3

Linear Algebra

Part II

Optimization

Chapter 4

Convex Optimization

4.1 Mathematical optimization problem

A mathematical optimization problem can be expressed as

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0 \text{ for } i = 1, \ldots, m \\ & h_i(x) = 0 \text{ for } i = 1, \ldots, p \end{array} \tag{4.1}$$

where $x \in \mathbf{R}^n$ is the optimization variable, $f_0 : \mathbf{R}^n \to \mathbf{R}$ is the objective function, $f_i : \mathbf{R}^n \to \mathbf{R}$ for $i = 1, \ldots, n$ are the inequality constraint functions, and $h_i : \mathbf{R}^n \to \mathbf{R}$ for $i = 1, \ldots, p$ are the equality constraint functions.

The conditions, $f_i(x) \leq 0$ for $i = 1, \ldots, m$, are called inequality constraints and the conditions, $h_i(x) = 0$ for $i = 1, \ldots, p$ are called equation constraints.

Note that this formulation covers pretty much every single-objective optimization problem. For example, consider the following optimization problem.

$$\begin{array}{ll} \text{maximize} & f(x_1, x_2) \\ \text{subject to} & x_1 \geq x_2 \\ & x_1 + x_2 = 2 \end{array} \tag{4.2}$$

This problem can be cast into an equivalent problem as follows.

$$\begin{array}{ll} \text{minimize} & -f(x_1, x_2) \\ \text{subject to} & -x_1 + x_2 \leq 0 \\ & x_1 + x_2 - 2 = 0 \end{array} \tag{4.3}$$

The feasible set for (4.1) is defined by the set of $x \in \mathbf{R}^n$ which satisfies all the contraints. Also, the optimal value for (4.1) is the infimum of $f_0(x)$ while x is in the feasible set. When the infimum is achievable, we define the optimal solution set as the set of all feasible x achieving the infimum value. These are defined in mathematically rigorous terms below.

- The feasible set for (4.1) is defined by

$$\mathcal{F} = \{x \in \mathcal{D} \mid f_i(x) \leq 0 \text{ for } i = 0, \ldots, m,\ h_j(x) = 0 \text{ for } j = 1, \ldots, p\} \subseteq \mathbf{R}^n \tag{4.4}$$

  where

$$\mathcal{D} = \left(\bigcap_{0 \leq i \leq m} \mathbf{dom}\, f_i \right) \cap \left(\bigcap_{1 \leq i \leq p} \mathbf{dom}\, h_i \right). \tag{4.5}$$

- The optimal value for (4.1) is defined by

$$p^* = \inf_{x \in \mathcal{F}} f_0(x) \tag{4.6}$$

  We use the conventions that $p^* = -\infty$ if $f_0(x)$ is unbounded below for $x \in \mathcal{F}$ and that $p^* = \infty$ if $\mathcal{F} = \emptyset$.

- The optimal solution set for (4.1) is defined by

$$\mathcal{X}^* = \{x \in \mathcal{F} \mid f_0(x) = p^*\}. \tag{4.7}$$

4.2 Convex optimization problem

A mathematical optimization problem is called a convex optimization problem if the objective function and all the inequality constraint functions are convex functions and all the equality constraint functions are affine functions.

Hence, a convex optimization problem can be expressed as

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0 \text{ for } i = 1, \ldots, m \\ & Ax = b \end{array} \tag{4.8}$$

where $x \in \mathbf{R}^n$ is the optimization variable, $f_i : \mathbf{R}^n \to \mathbf{R}$ for $i = 0, \ldots, n$ are convex functions, $h_i : \mathbf{R}^n \to \mathbf{R}$ for $i = 1, \ldots, p$ are the equality constraint functions, $A \in \mathbf{R}^{p \times n}$, and $b \in \mathbf{R}^p$.

A function, $f : \mathbf{R}^n \to \mathbf{R}$, is called a convex function if $\mathbf{dom}\, f \subseteq \mathbf{R}^n$ is a convex set and for all $x, y \in \mathbf{dom}\, f$ and all $0 \leq \lambda \leq 1$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \tag{4.9}$$

where $\mathbf{dom}\, f \subseteq \mathbf{R}^n$ denotes the domain of f.

A convex optimization enjoys a number of nice theoretical and practical properties.

- A local minimum of a convex optimization problem is a global minimum, *i.e.*, if for some $R > 0$ and $x_0 \in \mathcal{F}$, $\|x - x_0\| < R$ and $x \in \mathcal{F}$ imply $f_0(x_0) \leq f_0(x)$, then $f_0(x_0) \leq f_0(x)$ for all $x \in \mathcal{F}$.

  Proof: Assume that $x_0 \in \mathcal{F}$ is a local minimum, *i.e.*, for some $R > 0$, $\|x - x_0\| < R$ and $x \in \mathcal{F}$ imply $f_0(x_0) \leq f_0(x)$.

  Now assume that x_0 is not a global minimum, *i.e.*, there exists $y \in \mathcal{F}$ such that $y \neq x_0$ and $f_0(y) < f_0(x_0)$. Then for $z = \lambda y + (1-\lambda)x_0$ with $\lambda = \min\{R/\|y - x_0\|, 1\}/2$, the convexity of f_0 implies

  $$f_0(z) \leq \lambda f_0(y) + (1-\lambda) f_0(x_0) \tag{4.10}$$

  since $0 < \lambda \leq 1/2 < 1$. Furthermore

  $$\|z - x_0\| = \lambda \|y - x_0\| \leq R/2, \tag{4.11}$$

  hence $f_0(z) \geq f_0(x_0)$, which together with (4.10) implies

  $$f_0(x_0) \leq f_0(z) \leq \lambda f_0(y) + (1-\lambda) f_0(x_0) < \lambda f_0(x_0) + (1-\lambda) f_0(x_0) = f_0(x_0), \tag{4.12}$$

  which is a contradiction. Therefore there is no $y \in \mathcal{F}$ such that $y \neq x_0$ and $f_0(y) < f_0(x_0)$. Therefore x_0 is a global minimum.

- For a unconstrained problem, *i.e.*, the problem (4.8) with $m = p = 0$, with differentiable objective function, $x \in \mathbf{dom}\, f_0$ is an optimal solution if and only if $\nabla f_0(x) = 0 \in \mathbf{R}^n$.

Proof: The Taylor theorem implies that for any $x, y \in \textbf{dom}\, f_0$,

$$f_0(y) = f(x) + \nabla f_0(x)^T(y-x) + \frac{1}{2}(y-x)^T \nabla^2 f_0(z)(y-x) \tag{4.13}$$

for some z on the line segment having x and y as its end points, *i.e.*, $z = \alpha x + (1-\alpha)y$ for some $0 \le \alpha \le 1$. Since $\nabla^2 f(x) \succeq 0$ for any $z \in \textbf{dom}\, f_0$, we have

$$f_0(y) \ge f_0(x) + \nabla f_0(x)^T(y-x) \tag{4.14}$$

Thus, if for some $x_0 \in \mathbf{R}^n$, $\nabla f_0(x_0) = 0$, for any $x \in \textbf{dom}\, f_0$,

$$f_0(x) \ge f_0(x_0) + \nabla f_0(x_0)^T(x - x_0) = f_0(x_0), \tag{4.15}$$

hence x_0 is an optimal solution. Now assume that x_0 is an optimal solution, but $\nabla f_0(x_0) \neq 0$. Then for any $k > 0$, if we let $x = x_0$ and $y = x_0 - k\nabla f_0(x_0)$, (4.13) becomes

$$\begin{aligned} f_0(y) = f(x_0) + \nabla f_0(x_0)^T(-k\nabla f_0(x_0)) + \frac{k^2}{2}\nabla f_0(x_0)^T \nabla^2 f_0(z) \nabla f_0(x_0) \\ = \quad f(x_0) - k\|\nabla f_0(x_0)\|^2 + \frac{k^2}{2}\nabla f_0(x_0)^T \nabla^2 f_0(z) \nabla f_0(x_0) \end{aligned}$$

for all $y = x_0 - k\nabla f_0(x_0) \in \textbf{dom}\, f_0$.

Since for $k < 2\|\nabla f_0(x_0)\|^2 / \nabla f_0(x_0)^T \nabla^2 f_0(z) \nabla f_0(x_0)$, $-k\|\nabla f_0(x_0)\|^2 + \frac{k^2}{2}\nabla f_0(x_0)^T \nabla^2 f_0(z) \nabla f_0(x_0) < 0$, thus $f_0(y) < f(x_0)$, hence the constradiction. Therefore, if x_0 is an optimal solution for the unconstrained problem, $\nabla f_0(x_0) = 0$.

4.3 Duality

4.3.1 The Lagrange dual problem

4.3.1.1 Examples

4.3.1.1.1 Standard form LP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b \\ & x \succeq 0 \end{array} \tag{4.16}$$

The Lagrange dual problem is

$$\begin{array}{ll} \text{maximize} & -b^T \nu \\ \text{subject to} & A^T \nu + c \ge 0 \end{array} \tag{4.17}$$

4.3.1.1.2 Inequality form LP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \preceq b \end{array} \tag{4.18}$$

The Lagrange dual problem is

$$\begin{array}{ll} \text{maximize} & -b^T \lambda \\ \text{subject to} & A^T \lambda + c = 0 \\ & \lambda \succeq 0 \end{array} \tag{4.19}$$

4.3.1.1.3 Least-squares solution of linear equations

$$\begin{array}{ll} \text{minimize} & (1/2) x^T x \\ \text{subject to} & Ax = b \end{array} \tag{4.20}$$

The Lagrange dual problem is

$$\text{maximize} \quad -(1/2)\nu^T A A^T \nu - b^T \nu \tag{4.21}$$

4.3.1.1.4 Entropy maximization

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n x_i \log x_i \\ \text{subject to} & Ax = b \\ & \mathbf{1}^T x = 1 \end{array} \tag{4.22}$$

with domain $\mathcal{D} = \mathbf{R}^n_+$

The Lagrange dual problem is

$$\begin{array}{ll} \text{maximize} & -b^T \lambda - \log\left(\sum_{i=1}^n \exp(-a_i^T \lambda)\right) \\ \text{subject to} & \lambda \succeq 0 \end{array} \tag{4.23}$$

4.3.2 Interpretations

4.3.2.1 Max-min characterization of weak and strong duality

We first note that for any $f : X \times Y \to \mathbf{R}$, we have

$$\sup_{y \in Y} \inf_{x \in X} f(x, y) \le \inf_{x \in X} \sup_{y \in Y} f(x, y). \tag{4.24}$$

This inequality is called *max-min inequality.*

We can prove this as follows. Let $g : Y \to \mathbf{R}$ be a function defined by $g(y) = \inf_{x \in X} f(x, y)$ and let $h : X \to \mathbf{R}$ be a function defined by $h(x) = \sup_{y \in Y} f(x, y)$. Then we have that for any $x \in X$ and $y \in Y$

$$g(y) = \inf_{x \in X} f(x, y) \le f(x, y), \tag{4.25}$$

which implies that for any $x \in X$

$$\sup_{y \in Y} g(y) \leq \sup_{y \in Y} f(x, y) = h(x). \tag{4.26}$$

This again implies that

$$\sup_{y \in Y} g(y) \leq \inf_{x \in X} h(x), \tag{4.27}$$

hence the proof.

4.3.2.2 Saddle-point interpretation

Suppose $f : X \times Y \to \mathbf{R}$. We refer a point $(\tilde{x}, \tilde{y}) \in X \times Y$ a *saddle-point* for f (and X and Y) if

$$f(\tilde{x}, y) \leq f(\tilde{x}, \tilde{y}) \leq f(x, \tilde{y}) \tag{4.28}$$

for all $x \in X$ and $y \in Y$.

Now if x^* and λ^* are primal and dual optimal points for a problem in which strong duality obtains, the form a saddle-point for the Lagrangian. Conversely, if (x, λ) is a saddle-point of the Lagrangian, then x is primal optimal, λ is dual optimal, and the optimal duality gap is zero.

To prove these, assume that $x^* \in \mathcal{D}$ and $(\lambda^*, \nu^*) \in \mathbf{R}_+^m \times \mathbf{R}^p$ are primal and dual optimal points for a problem in which strong duality obtains. Then for any $x \in \mathcal{D}$ and $(\lambda, \nu) \in \mathbf{R}_+^m \times \mathbf{R}^p$, we have

$$L(x^*, \lambda, \nu) = f_0(x^*) + \sum_{i=1}^m \lambda_i f_i(x^*) + \sum_{i=1}^p \nu_i h_i(x^*) \leq f_0(x^*) = g(\lambda^*, \nu^*) \leq L(x, \lambda^*, \nu^*) \tag{4.29}$$

where the left inequality comes from the fact that $\lambda_i f_i(x^*) \leq 0$ for all $i = 1, \ldots, m$ and $h_i(x^*) = 0$ for all $i = 1, \ldots, p$ and the right inequality comes from the definition of (Lagrange) dual function. Now from the complementary slackness we know that $\lambda_i f_i(x^*) = 0$ for all $i = 1, \ldots, m$. Therefore

$$L(x^*, \lambda^*, \nu^*) = f_0(x^*), \tag{4.30}$$

thus we have

$$L(x^*, \lambda, \nu) \leq L(x^*, \lambda^*, \nu^*) \leq L(x, \lambda^*, \nu^*), \tag{4.31}$$

hence the proof.

Now suppose that $\tilde{x} \in \mathcal{D}$ and $(\tilde{\lambda}, \tilde{\nu}) \in \mathbf{R}_+^m \times \mathbf{R}^p$ are the saddle-point of the Lagrangian, *i.e.*, for all $x \in \mathcal{D}$ and $(\lambda, \nu) \in \mathbf{R}_+^m \times \mathbf{R}^p$,

$$L(\tilde{x}, \lambda, \nu) \leq L(\tilde{x}, \tilde{\lambda}, \tilde{\nu}) \leq L(x, \tilde{\lambda}, \tilde{\nu}). \tag{4.32}$$

First we show that $\tilde{x}$ is a feasible point. The left inequality says that for all $(\lambda, \nu) \in \mathbf{R}_+^m \times \mathbf{R}^p$,

$$L(\tilde{x}, \lambda, \nu) = f_0(\tilde{x}) + \sum_{i=1}^m \lambda_i f_i(\tilde{x}) + \sum_{i=1}^p \nu_i h_i(\tilde{x}) \leq L(\tilde{x}, \tilde{\lambda}, \tilde{\nu}) \tag{4.33}$$

If $f_i(\tilde{x}) > 0$ for some $i \in \{1, \ldots, m\}$ or $h_i(\tilde{x}) \neq 0$ for some $i \in \{1, \ldots, p\}$, $L(\tilde{x}, \lambda, \nu)$ is unbounded above and the above inequality cannot hold. Therefore $f_i(\tilde{x}) \leq 0$ for all $i \in \{1, \ldots, m\}$ and $h_i(\tilde{x}) = 0$

for all $i \in \{1, \ldots, p\}$, *i.e.*, $\tilde{x}$ is primal feasible. Since the inequality must hold when $\lambda = 0$ and $\nu = 0$, we have

$$f(\tilde{x}) \leq L(\tilde{x}, \tilde{\lambda}, \tilde{\nu}). \tag{4.34}$$

The right inequality of (4.32) implies that

$$L(\tilde{x}, \tilde{\lambda}, \tilde{\nu}) \leq g(\tilde{\lambda}, \tilde{\nu}) = \inf_{x \in \mathcal{D}} L(x, \tilde{\lambda}, \tilde{\nu}), \tag{4.35}$$

which implies that $f_0(\tilde{x}) \leq g(\tilde{\lambda}, \tilde{\nu})$. Since $g(\lambda, \nu)$ is an underestimator of $f_0(x)$ for any feasible $x \in \mathcal{D}$ and $(\tilde{\lambda}, \tilde{\nu}) \in \mathbf{R}_+^m \times \mathbf{R}^p$, *i.e.*, $g(\tilde{\lambda}, \tilde{\nu}) \leq f_0(\tilde{x})$, thus $g(\tilde{\lambda}, \tilde{\nu}) = f_0(\tilde{x})$. Therefore $\tilde{x}$ is an optimal solution for the primal problem and $(\tilde{\lambda}, \tilde{\nu})$ is an optimal solution for the dual problem, hence the proof.

4.4 Convex optimization problems

4.4.1 Equality constrained problem

Consider the following equality constrained problem:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b \end{array} \tag{4.36}$$

where $x \in \mathbf{R}^n$ is the optimization variable, $A \in \mathbf{R}^{m \times n}$, and $b \in \mathbf{R}^m$. The Lagrangian is

$$L(x, \nu) = f(x) + \nu^T (Ax - b) \tag{4.37}$$

and the Lagrange dual function is

$$g(\nu) = \inf_{x \in \mathbf{R}^n} L(x, \nu) = -\sup_{x \in \mathbf{R}^n} (-\nu^T Ax - f(x)) - b^T \nu = -f^*(-A^T \nu) - b^T \nu \tag{4.38}$$

The KKT optimality conditions are

$$\text{primal feasibility:} \quad Ax = b \tag{4.39}$$

$$\text{gradient of Lagrangian vanishes:} \quad \nabla f(x) + A^T \nu = 0 \tag{4.40}$$

4.4.1.1 Equality constrained problem examples

Consider the following equality constraint quadratic problem:

$$\begin{array}{ll} \text{minimize} & x^T P x + q^T x \\ \text{subject to} & Ax = b \end{array} \tag{4.41}$$

where $x \in \mathbf{R}^n$ is the optimization variable, $P \in \mathcal{S}_{++}^n$, $q \in \mathbf{R}^n$, $A \in \mathbf{R}^{m \times n}$, and $b \in \mathbf{R}^m$.

The Lagrangian is

$$L(x, \nu) = x^T P x + q^T x + \nu^T (Ax - b). \tag{4.42}$$

The gradient of the Lagrangian with respect to x is

$$\nabla_x L(x,\nu) = 2Px + q + A^T\nu = 0, \tag{4.43}$$

hence

$$\operatorname*{argmin}_x L(x,\nu) = -\frac{1}{2}P^{-1}(q + A^T\nu) \tag{4.44}$$

The KKT conditions are

$$\text{primal feasibility:} \qquad Ax = b \tag{4.45}$$

$$\text{gradient of Lagrangian vanishes:} \qquad 2Px + q + A^T\nu = 0 \tag{4.46}$$

which are equivalent to

$$\begin{bmatrix} 2P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ \nu \end{bmatrix} = \begin{bmatrix} -q \\ b \end{bmatrix}. \tag{4.47}$$

The conjugate of the objective function is

$$f^*(y) = \sup_x (y^T x - x^T P x - q^T x). \tag{4.48}$$

Since the gradient of $y^T x - x^T P x - q^T x$ is $y - q - 2Px$,

$$\operatorname*{argsup}_x (y^T x - x^T P x - q^T x) = \frac{1}{2}P^{-1}(y - q), \tag{4.49}$$

thus

$$\begin{aligned} f^*(y) &= -\frac{1}{4}(y-q)^T P^{-1}(y-q) + \frac{1}{2}(y-q)^T P^{-1}(y-q) = \frac{1}{4}(y-q)^T P^{-1}(y-q) \\ &= \frac{1}{4}\left(y^T P^{-1} y - 2q^T P^{-1} y + q^T P^{-1} q\right) \end{aligned}$$

4.5 Unconstrained minimization

4.5.1 Gradient descent method

4.5.1.1 Examples

4.5.1.1.1 A quadratic problem in $\mathbf{R}^2$ We consider the quadratic objective function on $\mathbf{R}^2$

$$f(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2) \tag{4.50}$$

where $\gamma > 0$.

We apply the gradient descent method with exact line search. The gradient of f is

$$\nabla f(x) = \begin{bmatrix} x_1 \\ \gamma x_2 \end{bmatrix} \tag{4.51}$$

Let $\tilde{f} : \mathbf{R}_+ \to \mathbf{R}$ defined by $\tilde{f}(t) = f(x - t\nabla f(x))$. Now

$$\tilde{f}(t) = f\left(\begin{bmatrix} (1-t)x_1 \\ (1-\gamma t)x_2 \end{bmatrix}\right) = \frac{1}{2}\left((1-t)^2 x_1^2 + \gamma(1-\gamma t)^2 x_2^2\right) \tag{4.52}$$

and

$$\frac{d}{dt}\tilde{f}(t) = -(1-t)x_1^2 - \gamma^2(1-\gamma t)x_2^2 = 0 \tag{4.53}$$

implies

$$t = \frac{x_1^2 + \gamma^2 x_2^2}{x_1^2 + \gamma^3 x_2^2} \tag{4.54}$$

minimizes $\tilde{f}(t)$. Since

$$1 - t = \frac{\gamma^2(\gamma-1)x_2^2}{x_1^2 + \gamma^3 x_2^2} \tag{4.55}$$

and

$$1 - \gamma t = \frac{(1-\gamma)x_1^2}{x_1^2 + \gamma^3 x_2^2} \tag{4.56}$$

Thus the exact line search yields

$$x^+ = x - t\nabla f(x) = \begin{bmatrix} (1-t)x_1 \\ (1-\gamma t)x_2 \end{bmatrix} = \frac{(1-\gamma)x_1 x_2}{x_1^2 + \gamma^3 x_2^2}\begin{bmatrix} -\gamma^2 x_2 \\ x_1 \end{bmatrix}. \tag{4.57}$$

If $x = \alpha[\gamma\ 1]^T$, then

$$x^+ = \frac{\alpha^3(1-\gamma)\gamma}{\alpha^2\gamma^2(1+\gamma)}\begin{bmatrix} -\gamma^2 \\ \gamma \end{bmatrix} = \alpha\frac{1-\gamma}{1+\gamma}\begin{bmatrix} -\gamma \\ 1 \end{bmatrix}. \tag{4.58}$$

If $x = \alpha[-\gamma\ 1]^T$, then

$$x^+ = -\frac{\alpha^3(1-\gamma)\gamma}{\alpha^2\gamma^2(1+\gamma)}\begin{bmatrix} -\gamma^2 \\ -\gamma \end{bmatrix} = \alpha\frac{1-\gamma}{1+\gamma}\begin{bmatrix} \gamma \\ 1 \end{bmatrix}. \tag{4.59}$$

Therefore if $x^{(0)} = [\gamma\ 1]^T$, then

$$x^{(k)} = \left(\frac{1-\gamma}{1+\gamma}\right)^k \begin{bmatrix} (-1)^k\gamma \\ 1 \end{bmatrix} = \left(\frac{\gamma-1}{\gamma+1}\right)^k \begin{bmatrix} \gamma \\ (-1)^k \end{bmatrix}. \tag{4.60}$$

Chapter 5

Portfolio optimization

Part III

Statistics

Part IV

Machine Learning

Chapter 10

Machine Learning Basics

Chapter 11

Optimization for Machine Learning

Chapter 12

Bayesian Network

Chapter 13

Collaborative Filtering

Chapter 14

Time Series Anomaly Detection

Chapter 15

Reinforcement Learning