Guanghui Lan, with materials adapted mainly
from Ben-Tal and Nemiroski's lectures for
ISyE 6663 and a few other sources

Lectures for Convex and Nonlinear Optimization

November 22, 2022

H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology, Atlanta, GA

Contents

1 **Motivating Applications** 1
1.1 Regression and Classification 1
1.2 Reinforcement Learning 8
1.3 Radiation Therapy Treatment Planning 11
1.4 General Formulation 15

2 **Convex Sets** 17
2.1 Definition and Examples 17
2.2 "Calculus" of Convex Sets 25
2.3 Topological Properties of Convex Sets 26
2.4 Caratheodory's Theorem 30
2.5 Radon's Theorem 32
2.6 Helley's Theorem 33
2.7 Separation Theorem 36
2.8 Supporting Planes 41
2.9 Extreme Points* 42
2.10 Exercises 44

3 **Convex Functions** 51
3.1 Definition and examples 51
3.2 Jensen's Inequality 52
3.3 Extended Real 53
3.4 Convexity-preserving Operations 53
3.5 Detection of Convexity 55
3.6 Lipschitz continuity of convex functions* 57
3.7 Minima and Maxima of Convex Functions 59
3.8 Optimality Condition over Polyhedron 62
3.9 Subgradients 64
3.10 Exercises 67

4 Duality and Optimality Conditions 77
4.1 Convex Programming 77
4.2 Convex Theorem on Alternative 78
4.3 Lagrange Duality 80
4.4 Saddle Points 81
4.5 Saddle Point Form of Optimality Conditions 83
4.6 Karush-Kuhn-Tucker Optimality Condition 84
4.7 Convex-Cancave Saddle Points* 85
4.8 First-order Optimality Conditions 88
4.9 Second-order Condition for Unconstrained Problems 91
4.10 Second-order Necessary Condition for Constrained Problems 92
4.11 Second-order Sufficient Condition for Constrained Problems 96
4.12 Applications: Eigenvalue Decomposition and S-Lemma 99
4.13 Exercises 101

5 Optimization Methods 111
5.1 Introduction 111
5.2 Rate of Convergence 113
5.3 Gradient Descent Method 116
5.4 Conjugate Gradient Method 123
5.5 Projected gradient method 127
5.6 Accelerated Gradient Descent Method 130
5.7 Lower Complexity Bound 132
5.8 Conditional gradient method 135
5.9 Ellipsoid methods for convex optimization 138
5.10 Newton's Method 141
5.11 Quasi-Newton (Variable Metric) method 143
5.12 Cubic Regularization* 144
5.13 Methods for function constraints 146
5.14 Excercises 147
References 153

Chapter 1
Motivating Applications

Nonlinear optimization has played an important role in a few different areas, both as a modeling apparatus and a solution method. In this chapter, we introduce some well-known nonlinear optimization models in order to motivate our later discussion about optimization theory and algorithms.

1.1 Regression and Classification

To motivate our discussion, let us start with a simple example. Julie needs to decide whether she should go to the restaurant "Bamboo Garden" for lunch or not. She went to ask for her friends Judy and Jim, who had been to this restaurant. Both of them gave a rating of 3 in the scale between 1 to 5 for the service in this restaurant. Given these ratings, it is a bit difficult for Julie to decide if she should pay a visit to "Bamboo Garden". Fortunately, she has kept a table of Judy and Jim's ratings for some other restaurants, as well as her own ratings in the past, as shown in Table 1.1.

Table 1.1: Historical ratings for the restaurants.

Restaurant	Judy's rating	Jim's rating	Julie's ratings?
Goodfellas	1	5	2.5
Hakkasan	4.5	4	5
···	···	···	···
Bamboo Garden	3	3	?

To fix notation, let us use $u^{(i)}$ to denote the "input" variables (ratings of Judy and Jim in this example), also called input features, and $v^{(i)}$ to denote the "output" or target variable (rating of Julie's) to predict. A pair $(u^{(i)}, v^{(i)})$ is called a training example, and the dataset — a list of N training examples $\{(u^{(i)}, v^{(i)}\}$, $i = 1, \ldots, N$, is called a training set. We will also use U denote the space of input values, and V the space of output values. In this example, $U = \mathbb{R}^2$ and $V = \mathbb{R}$. Specifically, $u_1^{(1)}$ and

$u_2^{(1)}$ are the Judy and Jim's ratings for Goodfellas, respectively, and $v^{(1)}$ represents Julie's rating for Goodfellas.

Our goal is, given a training set, to learn a function $h : U \to V$ so that $h(u)$ is a "good" predictor for the corresponding value of v. This function h is usually called a hypothesis or decision function. Machine learning tasks of these types are called *supervised learning*. When the output v is continuous, we call the learning task *regression*. Otherwise, if v takes values on a discrete set of values, the learning task is called *classification*. Regression and classification are the two main tasks in supervised learning.

Linear Regression

One simple idea is to approximate v by a linear function of u:

$$h(u) \equiv h_\theta(u) = \theta_0 + \theta_1 u_1 + \ldots + \theta_n u_n.$$

In our example, n simply equals 2. For notational convenience, we introduce the convention of $u_0 = 1$ so that

$$h(u) = \textstyle\sum_{i=0}^{n} \theta_i u_i = \theta^T u,$$

where $\theta = (\theta_0; \ldots; \theta_n)$ and $u = (u_0; \ldots; u_n)$. In order to find the parameters $\theta \in \mathbb{R}^{n+1}$, we formulate an optimization problem of

$$\min_\theta \left\{ f(\theta) := \textstyle\sum_{i=1}^{N} (h_\theta(u^{(i)}) - v^{(i)})^2 \right\}, \tag{1.1.1}$$

which gives rise to the ordinary least square regression model.

To derive a solution of θ for (1.1.1), let

$$U = \begin{bmatrix} {u^{(1)}}^T \\ {u^{(2)}}^T \\ \vdots \\ {u^{(N)}}^T \end{bmatrix}.$$

U is sometimes called the design matrix and it consists of all the input variables. Then, $f(\theta)$ can be written as:

$$\begin{aligned} f(\theta) &= \textstyle\sum_{i=1}^{N} ({u^{(i)}}^T \theta - v^{(i)})^2 \\ &= (U\theta - v)^T (U\theta - v) \\ &= \theta^T U^T U \theta - 2\theta^T U^T v - v^T v. \end{aligned}$$

Taking the derivative of $f(\theta)$ and setting it to zero, we obtain the normal equation

$$U^T U\theta - U^T v = 0.$$

Thus the minimizer of (1.1.1) is given by

$$\theta^* = (U^T U)^{-1} U^T v.$$

The ordinary least square regression is among very few machine learning models that has an explicit solution. Note, however, that to compute θ^*, one needs to compute the inverse of an $(n+1)\times(n+1)$ matrix $(U^T U)$. If the dimension of n is big, to compute the inverse of a large matrix can still be computationally expensive.

The formulation of the optimization problem in (1.1.1) follows a rather intuitive approach. In the sequel, we provide some statistical reasoning about this formulation. Let us denote

$$\varepsilon^{(i)} = v^{(i)} - \theta^T u^{(i)}, i = 1,\dots,N. \tag{1.1.2}$$

In other words, $\varepsilon^{(i)}$ denotes the error associated with approximating $v^{(i)}$ by $\theta^T u^{(i)}$. Moreover, assume that $\varepsilon^{(i)}$, $i = 1,\dots,N$, are i.i.d. (independently and identically distributed) according to a Gaussian (or Normal) distribution with mean 0 and variance σ^2. Then, the density of $\varepsilon^{(i)}$ is then given by

$$p(\varepsilon^{(i)}) = \tfrac{1}{\sqrt{2\pi}\sigma}\exp\left(-\tfrac{(\varepsilon^{(i)})^2}{2\sigma^2}\right).$$

Using (1.1.2) in the above equation, we have

$$p(v^{(i)}|u^{(i)};\theta) = \tfrac{1}{\sqrt{2\pi}\sigma}\exp\left(-\tfrac{(v^{(i)}-\theta^T u^{(i)})^2}{2\sigma^2}\right). \tag{1.1.3}$$

Here, $p(v^{(i)}|u^{(i)};\theta)$ denotes the distribution of the output $v^{(i)}$ given input $u^{(i)}$ and parameterized by θ.

Given the input variables $u^{(i)}$ and output $v^{(i)}$, $i = 1,\dots,N$, the likelihood function with respect to (w.r.t.) the parameters θ is defined as

$$L(\theta) := \prod_{i=1}^{N} p(v^{(i)}|u^{(i)};\theta) = \prod_{i=1}^{N} \tfrac{1}{\sqrt{2\pi}\sigma}\exp\left(-\tfrac{(v^{(i)}-\theta^T u^{(i)})^2}{2\sigma^2}\right).$$

The principle of *maximum likelihood* tells us that we should choose θ to maximize the likelihood $L(\theta)$, or equivalently, the *log likelihood*

$$\begin{aligned} l(\theta) &:= \log L(\theta) \\ &= \textstyle\sum_{i=1}^N \log\left[\tfrac{1}{\sqrt{2\pi}\sigma}\exp\left(-\tfrac{(v^{(i)}-\theta^T u^{(i)})^2}{2\sigma^2}\right)\right] \\ &= N\log\tfrac{1}{\sqrt{2\pi}\sigma} - \tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^N (v^{(i)} - \theta^T u^{(i)})^2. \end{aligned}$$

This is exactly the ordinary least square regression problem, i.e., to minimize $\sum_{i=1}^N (v^{(i)} - \theta^T u^{(i)})^2$ w.r.t. θ. The above reasoning tells us that under certain probabilistic assumptions, the ordinary least square regression is the same as maximum likelihood estimation. It should be noted, however, that the probabilistic assumptions are by no means necessary for least-squares to be a rational procedure for regression.

Logistic Regression

Let us come back to the previous example. Suppose that Julie only cares about whether she will like the restaurant "Bamboo Garden" or not, rather her own ratings. Moreover, she only recorded some historical data indicating whether she likes or dislikes some restaurants, as shown in Table 1.2. These records are also visualized in Figure 1.1, where each restaurant is represented by a green "O" or a red "X", corresponding to whether Julie liked or disliked the restaurant, respectively. The question is: with the rating of 3 from both of her friends, will Julie like Bamboo Garden? Can she use the past data to come up with a reasonable decision?

Table 1.2: Historical ratings for the restaurants.

Restaurant	Judy's rating	Jim's rating	Julie likes?
Goodfellas	1	5	No
Hakkasan	4.5	4	Yes
…	…	…	…
Bamboo Garden	3	3	?

Similar to the regression model, the input values are still denoted by $U = ({u^{(1)}}^T; \ldots; {u^{(N)}}^T)$, i.e., the ratings given by Judy and Jim. But the output values are now binary, i.e., $v^{(i)} \in \{0,1\}$, $i = 1, \ldots, N$. Here $v^{(i)} = 1$ means that Julie likes the i-th restaurant and $v^{(i)} = 0$ means that she dislikes the restaurant. Julie's goal is to come up with a decision function $h(u)$ to approximate these binary variables v. This type of machine learning task is called *binary classification*.

Julie's decision function can be as simple as a weighted linear combination of her friends' ratings:

$$h_\theta(u) = \theta_0 + \theta_1\, u_1 + \ldots + \theta_n u_n \tag{1.1.4}$$

with $n = 2$. One obvious problem with the decision function in (1.1.4) is that its values can be arbitrarily large or small. On the other hand, Julie wishes its values to fall between 0 and 1 because those represent the range of v. A simple way to force h to fall within 0 and 1 is to map the linear decision function $\theta^T u$ through another function called the sigmoid (or logistic) function

$$g(z) = \tfrac{1}{1+\exp(-z)} \tag{1.1.5}$$

and define the decision function as

$$h_\theta(u) = g(\theta^T u) = \tfrac{1}{1+\exp(-\theta^T u)}. \tag{1.1.6}$$

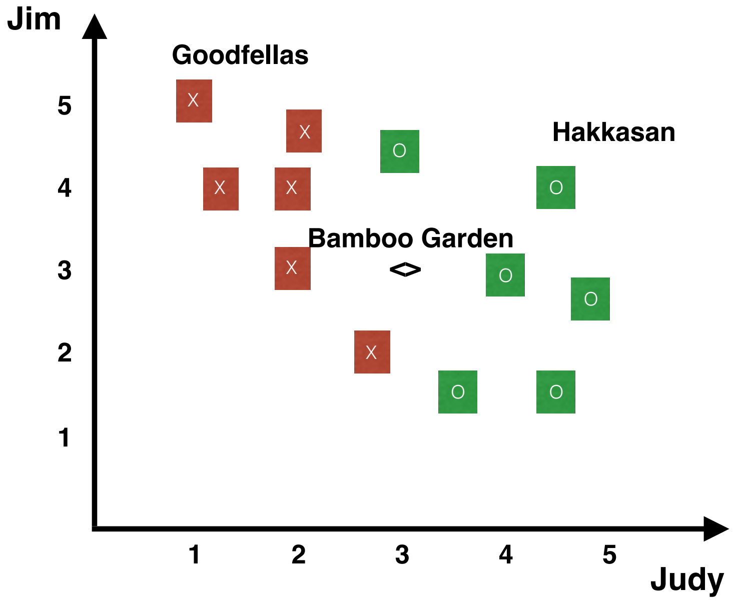


Fig. 1.1: Visualizing ratings of the restaurants

Note that the range of the sigmoid function is given by $(0,1)$, as shown in Figure 1.2.

Now the question is how to determine the parameters θ for the decision function in (1.1.6). We have seen the derivation of the ordinary least square regression model as the consequence of maximum likelihood estimation under certain probabilistic assumptions. We will follow a similar approach for the classification problem.

We assume that $v^{(i)}$, $i = 1,\ldots,N$, are independent Bernoulli random variables with success probability (or mean) of $h_\theta(u^{(i)})$. Thus their probability mass functions are given by

$$p(v^{(i)}|u^{(i)};\theta) = [h_\theta(u^{(i)})]^{v^{(i)}}\,[1-h_\theta(u^{(i)})]^{1-v^{(i)}}, v^{(i)} \in \{0,1\},$$

and the associated likelihood function $L(\theta)$ is defined as

$$L(\theta) = \prod_{i=1}^{N}\left\{[h_\theta(u^{(i)})]^{v^{(i)}}\,[1-h_\theta(u^{(i)})]^{1-v^{(i)}}\right\}.$$

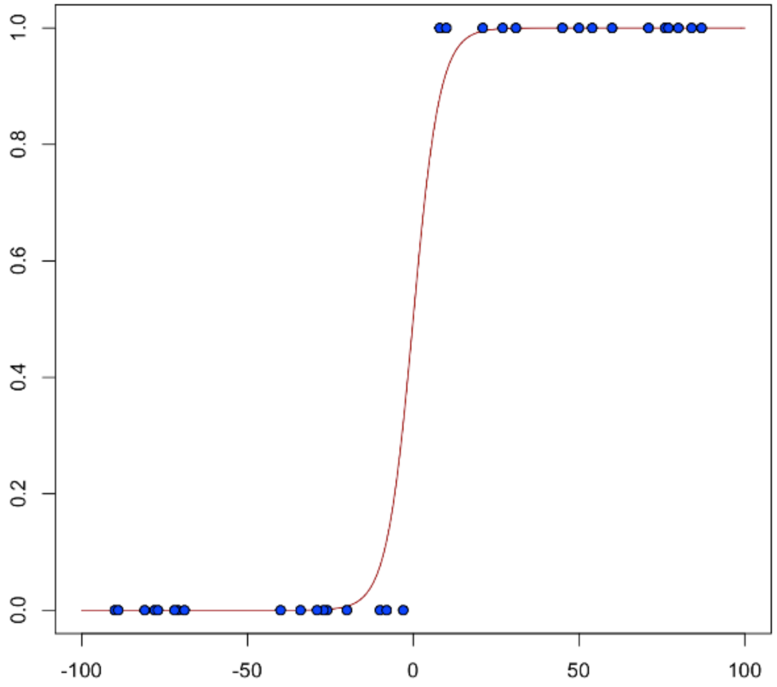


Fig. 1.2: The Sigmoid (logistic) function

In view of the principle of maximum likelihood, we intend to maximize $L(\theta)$, or equivalently, the log likelihood

$$\begin{aligned} l(\theta) &= \textstyle\sum_{i=1}^{N} \log \left\{ [h_\theta(u^{(i)})]^{v^{(i)}} [1 - h_\theta(u^{(i)})]^{1-v^{(i)}} \right\} \\ &= \textstyle\sum_{i=1}^{N} \left\{ v^{(i)} \log h_\theta(u^{(i)}) + [1 - v^{(i)}] \log[1 - h_\theta(u^{(i)})] \right\}. \end{aligned}$$

Accordingly, we formulate an optimization problem of

$$\max_{\theta} \textstyle\sum_{i=1}^{N} \left\{ -\log[1 + \exp(-\theta^T u^{(i)})] - [1 - v^{(i)}]\theta^T u^{(i)} \right\}. \tag{1.1.7}$$

Even though this model is used for binary classification, it is often called logistic regression for historical reasons.

Unlike linear regression, (1.1.7) does not have an explicit solution. Instead, we need to develop some numerical procedures to find its approximate solutions. These procedures are called optimization algorithms, a subject to be studied later in our lectures.

Suppose that Julie can solve the above problem and find at least one of its optimal solutions θ^*. She then obtains a decision function $h_{\theta^*}(u)$ which can be used to predict whether she likes a new restaurant (say "Bamboo Garden") or not. More specifically, recall that the example corresponding to "Bamboo Garden" is $u = (1,3,3)$ (recall $u_1 = 1$). If $h_{\theta^*}((1,3,3)) > 0.5$, then Julie thinks she will like the restaurant, otherwise she will not. The values of u's that cause $h_{\theta^*}(u)$ to be 0.5 is called the "decision boundary" as shown in Figure 1.3. The black line is the "decision boundary." Any point lying above the decision boundary represents a restaurant that Julie likes, while any point lying below the decision boundary is a restaurant that she does not like. With this decision boundary, it seems that Bamboo Garden is slightly on the positive side, which means she may like this restaurant.

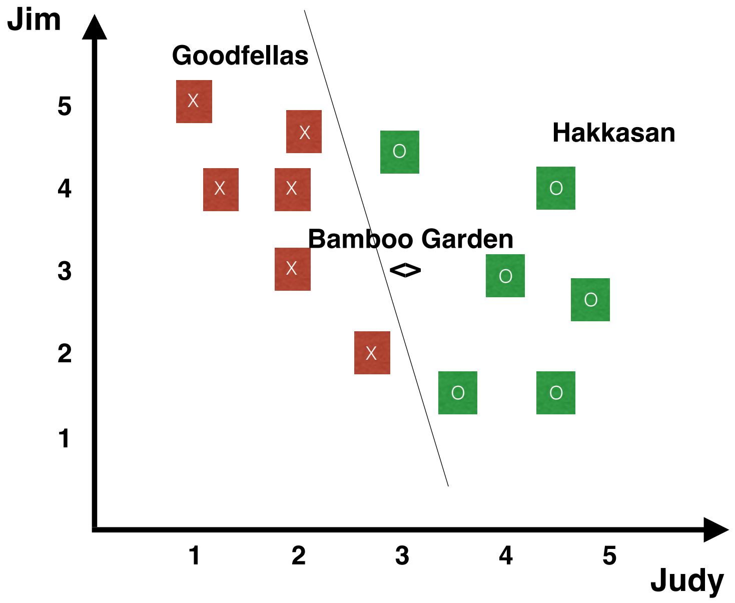
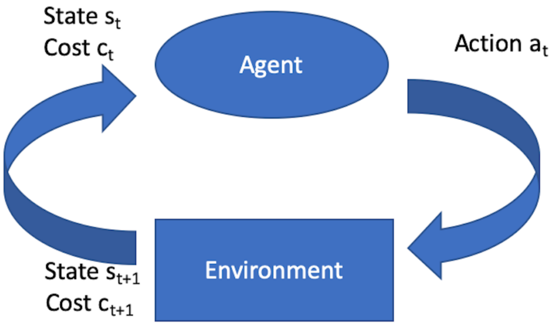


Fig. 1.3: Decision boundary

1.2 Reinforcement Learning

Stochastic dynamic programming provides a general framework to model the interactions between the agents and their environment, and to improve the agents' decisions through these interactions. More specifically, the status of the environment is described by either discrete or continuous state variables, while the agent's behavior is described by actions. Upon the agent's action, the system's state gets updated, and the agent receives some reward (or pays some cost). The goal of stochastic dynamic programming is to find the optimal policy which specifies the agent's best action at a given state.



Consider a search-and-rescue mission in a sophisticated environment, described as follows. Due to an environmental hazard (e.g., sand storm, severe thunderstorm, etc.), a few friendly units lost contact with the base of operation. A scouting unit needs to be dispatched to search and rescue these lost friendly units. However, due to the sophisticated environment there exist a few difficulties. First, since the friendly units have lost contact, their locations are unknown. Second, due to the sophisticated environment, the scouting unit may not be able to navigate properly according to its designated direction. Third, certain hazardous factors would inflict damage on the scouting unit, and the locations of such hazardous factors are also unknown. Fourth, the scouting unit might be ambushed by an adversary. In this case, the scouting unit would prefer to randomized moves, since deterministic moves are more predictable, and hence more vulnerable to an adversary attack. The main question is: How should the scouting unit plan its scouting trajectory for the rescue mission?

In the above search-and-rescue mission, in order to design an execution plan, we need to take several factors into consideration. From the perspective of cost analysis, we may associate a cost to each action that the scouting unit takes. Specifically, when designing the plan, the scouting unit needs to consider regular costs for maintaining its navigation, the potential loss when encountering a hazardous factor, and the reward for finding the lost friendly unit. In order to be less predictable, the cost analysis should also take into consideration the preference of a randomized strategy over deterministic ones. It is important to point out that time is critical for this mission; not only that the first several hours/days are well-known to be golden in rescue missions, but also the scouting unit needs to focus more on its imminent risks and rewards.

As a prototypical example, we can model the aforementioned search-and-rescue mission as a mathematical optimization problem using the terminology in stochastic dynamic programming. Our prototypical model is a simple and idealized one known as "GridWorld" in the literature. It can be extended with more complex factors to better model realistic search-and-rescue missions.

We describe a simple example of the search-and-rescue mission in Figure 1.4. The environment described in Figure 1.4 consists of seven locations on the grid, in which the base of operation is at gridpoint 1 (denoted as *). The friendly unit to be rescued is dispersed to two locations (gridpoints indexed by 3 and 7), which are our goals of the mission (denoted as G). There is a location on gridpoint 2 that is a hazardous factor (denoted as H). The remaining gridpoints 4, 5, and 6 are regular ones. At any gridpoint, the scouting unit could move and scout four directions (moving left/right/up/down by one gridpoint). Using terminology of stochastic dynamic programming, we call the scouting unit the *agent*, the gridpoints that the agent can reside on the states, and say that that an agent could perform four *actions* at any state (left/right/up/down). Upon choosing an action, due to possible navigation error, the agent then moves along the chosen direction with probability p, or a randomly chosen direction (among the remaining three directions) with probability $(1-p)/3$. After taking an action and moving to an updated gridpoint, the agent pays a cost associated with the updated gridpoint and the process repeats itself at the next timestep. The costs of goal, hazardous factor, and regular gridpoints are denoted by c_G, c_H, and c_R respectively, with $c_G < c_R < c_H$. When either at a goal or a hazardous factor gridpoint, the agent would transit to the base of operator with probabilities p_G and p_H respectively. Such probabilities model the possible respective scenarios that the friendly unit is impaired and needs the scouting unit's transport back to the base, or that the scouting unit is heavily damaged by the hazardous factor and need to return to the base and re-deploy. The goal of our model is to compute a *policy* for the agent to decide its strategies of choosing actions performed at different states in order to minimize its cumulative cost throughout the time horizon of actions. The cumulative cost is discounted by a factor to balance the objective of minimizing the cost and the urgency of reducing the imminent costs.

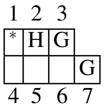


Fig. 1.4: A simple and idealized 2D GridWorld description of the search-and-rescue mission

Mathematically, we may model the aforementioned search-and-rescue mission through a finite Markov decision process (MDP), one of the most widely used stochastic dynamic programming models. The finite Markov decision process is abstracted by a quintuple $M = (\mathscr{S}, \mathscr{A}, P, c, \gamma)$, where $\mathscr{S}$ is the finite state space, $\mathscr{A}$ is the finite action space, $P : \mathscr{S} \times \mathscr{S} \times \mathscr{A} \to \mathbb{R}$ is the transition model (also known as transition probability/kernel in the literature), $c : \mathscr{S} \times \mathscr{A} \to \mathbb{R}$ is the cost function, and $\gamma \in [0,1]$ is the discount factor. A policy $\pi : \mathscr{A} \times \mathscr{S} \to \mathbb{R}$ determines the probability $\pi(a|s)$ of selecting a particular action $a \in \mathscr{A}$ at a given state $s \in \mathscr{S}$.

For a given policy π, we measure its performance by the *state-value function* $V^\pi : \mathscr{S} \to \mathbb{R}$ defined as

$$\begin{aligned} V^\pi(s) := \mathbb{E}\,\big[\textstyle\sum_{t=0}^{\infty} \gamma^t [c(s_t, a_t) + h^\pi(s_t)] \\ \mid s_0 = s, a_t \sim \pi(\cdot|s_t), s_{t+1} \sim P(\cdot|s_t, a_t)\big], \end{aligned} \tag{1.2.1}$$

where γ is the discount factor and h^π is a regularizer that describes the difference between the desirable policy and a reference policy. The above state-value function definition reflects the following interpretation of the associated value of policy π at state s: Initialized at state $s_0 = s$ and timestep $t = 0$, the agent (e.g., the scouting unit) chooses an action (e.g., a move and scout direction) $a_0 \sim \pi(\cdot|s_0)$ randomly based on the probabilities described by policy π. After action a_0 is performed, the agent transitions to state $s_1 \sim P(\cdot|s_0, a_0)$ following the transition model that incorporates the navigation error probability p and the return-to-base probabilities p_G and p_H. The process repeats then at state s_1 and continues on infinitely. The total value is the discounted sum of all costs $c(s_t, a_t)$'s (e.g., the scouting unit's costs by c_G, c_H, and c_R described above) and regularization values $h^\pi(s_t)$'s (e.g., the Kullback–Leibler (KL) divergence to promote randomized policies) over the infinite horizon $t = 0, 1, \cdots$, with discount factor γ. Here γ captures the balance between cumulative cost minimization and urgency of reducing imminent costs. Our main objective is to solve the policy optimization problem, namely, to find an optimal policy $\pi^* : \mathscr{S} \times \mathscr{A} \to \mathbb{R}$ such that the associated expected discounted cumulative cost is minimized:

$$V^{\pi^*}(s) \le V^\pi(s), \forall \pi(\cdot|s) \in \Delta_{|\mathscr{A}|}, \tag{1.2.2}$$

for any state $s \in \mathscr{S}$. Here $\Delta_{|\mathscr{A}|}$ denotes the standard simplex constraint for describing probabilities. Regarding our simple example described in Figure 1.4, there are in total $|\mathscr{S}| = 7$ states and $|\mathscr{A}| = 4$ actions, and hence we need to minimize $|\mathscr{S}| = 7$ state-value objective functions $V^\pi(s)$ in (1.2.2) (one for each state), each with respect to $|\mathscr{S}| \cdot |\mathscr{A}| = 28$ decision variables (one for probability of each actions in each state) in policy π.

Note that although there are $|\mathscr{S}|$ state-value objective functions in (1.2.2), we may reformulate the problem to a single objective by taking the weighted sum of V^π over s (with arbitrary weights $\rho_s > 0$ and $\sum_{s \in S} \rho_s = 1$):

$$\min_\pi f(\pi) := \mathbb{E}_{s\sim\rho}[V^\pi(s)] \ \text{ s.t. } \pi(\cdot|s) \in \Delta_{|\mathscr{A}|}, \forall s \in \mathscr{S}. \tag{1.2.3}$$

The above problem is a nonlinear nonconvex optimization. For example, in the unregularized case when $h^\pi(s) \equiv 0$ in the definition (1.2.1) of state-value function $V^\pi(s)$, $f(\pi)$ in the above problem is a weighted sum of $V^\pi(s)$'s, and it can be verified that for any $s \in \mathscr{S}$ the state-value function $V^{\pi}(s)$ defined in (1.2.1) is possibly nonconvex as a quotient of two polynomial functions.

It should be noted that the nonlinear optimization problem (1.2.3) is defined based on the probability transition model P in (1.2.1). However, the transition probability P may not be accessible when we compute for the optimal policy π^* and we may need to rely on samples obtained via the transition model. In some of the stochastic dynamic programming literature, the term MDP refers to the case when the transition model P is provided in advance; the case when we have no access to exact information in regards to P is categorized as reinforcement learning (RL). While both belong to the stochastic dynamic programming realm, the a-priori knowledge of P can be viewed as the main difference between MDP and RL.

1.3 Radiation Therapy Treatment Planning

In this section, we turn our attention to the intensity modulated radiation therapy (IMRT) problem arising from healthcare engineering. According to CDC, in 2017, the latest year for which incidence data are available in the United States, about 1.7 millions new cases of cancer were reported, and around $600,000$ people died of cancer. Cancer is the second leading cause of death, exceeded only by heart disease. One of every four deaths in the US is due to cancer.

Among many different types of treatment for cancer, radiation therapy can benefit more than half of these patients. It helps to cure cancer, prevent it from returning, and stop or slow its growth. Technology advancements in cancer treatment in general, and radiation therapy in particular, are critical to save patients' lives and improve their life quality. Radiation therapy applies high doses of radiation to kill cancer cells and shrink tumors. In particular, Intensity modulated radiation therapy (IMRT) is one type of external beam radiation therapy (see Figure 1.5). During the treatment, the patient will be irradiated by a linear accelerator from several different angles. Th target structures of patient are discretized into small voxels. We expect that the voxels of tumor receive high doses, while little or no doses will be applied to those in healthy organs.

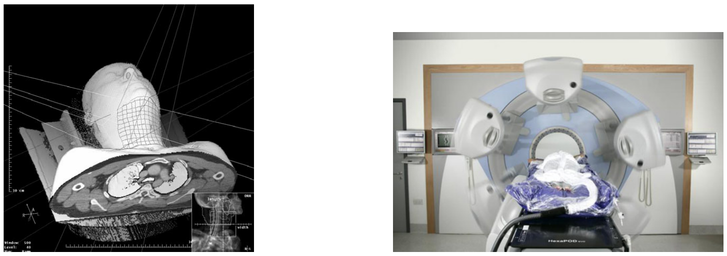


Fig. 1.5: Intensity modulated radiation therapy

The process flow of IMRT can be briefly described as follows. Images, such as magnetic resonance images (MRI) or computer tomography (CT), will be taken before the treatment starts. A treatment usually consists of a few (e.g., five) sessions. Before each session begins, treatment planners need to make two types of decisions. First, dose fractionation determines how to allocate doses across different sessions. Second, dose localization determines how to apply the right amount of does to the target structures. Inter-session images will be used observe the effectiveness of the previous treatment sessions, and provide guidance for the subsequent ones.

Optimization are quite useful throughout these steps. Linear regression with total variation regularization has been widely used for MRI or CT image reconstruction. The dose fractionation problem can be modeled as a MDP and hence falls into the reinforcement learning framework described in the previous section. Here we focus on the dose localization problem.

Aperture is an important concept for dose localization in IMRT. Recall that the beam will applied to the patient from different angles, as shown in Figure 1.5. A beam in each angle is decomposed into a rectangular grid of *beamlets*. A beamlet (i,j) is effective if it is not blocked by either the left, l_i, and right, r_i, leaves. These leaves in the IMRT equipment are zoomed-in in the left picture below. An aperture is defined as the collection of effective beamlets, as shown in Figure 1.6 on the right. The motion of the leaves controls the set of effective beamlets and thus the shape of the aperture. A linear combination of effective beamlets from different aperture will determine the doses applied to the patient at different voxels, as we will see a little bit later.

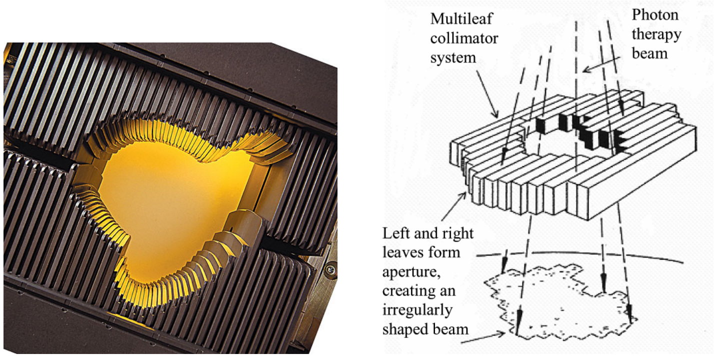

Multileaf collimator system
Photon therapy beam
Left and right leaves form aperture, creating an irregularly shaped beam

Fig. 1.6: Aperture

We need to consider two types of decision variables to properly define and use an aperture. The $x^{k,a}$ variables describe the shape of the k-th aperture in angle a. These are binary matrix variables. Specifically, the (i,j)-th entry of $x^{k,a} = 1$ if beamlet (i,j) is effective, that is, it falls within the left and right leaves of row i, o.w., it equals 0. Figure 1.7 shows two examples of aperture, where “gray" means 0 and “white" means 1.

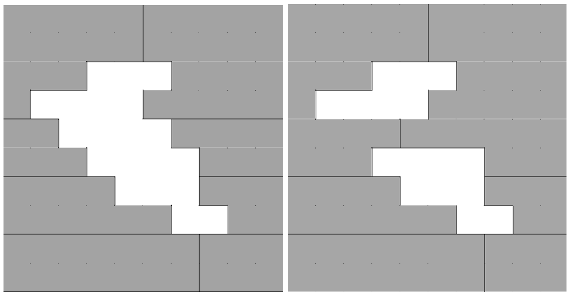


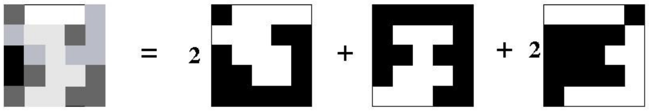
Fig. 1.7: Different example of apertures

Let K_a be the set of allowed apertures in beam angle a. It is easy to see that for an $m \times n$ grid, the number of possible apertures in a beam angle is $\mathcal{O}(n^m)$, which increases exponentially w.r.t. m. Once after an aperture is defined, we use the $y^{k,a}$ variable to represent the influence rate for aperture (k,a). It will determine the dose intensity and radiation time from the k-th aperture of the a-th angle.

Dose in Gray(Gy) absorbed by voxel v is given by the summations of doses received from each aperture. More specifically, the dose received at each voxel is a convex combination of doses received from each aperture, with weights given by the y variables:

$$z_v = \sum_{a \in \mathscr{A}} \sum_{k \in \mathscr{K}_a} (y^{k,a} R \sum_{i=1}^{m} \sum_{j=1}^{n} D_{(i,j)v} x_{ij}^{k,a}).$$

Here $D_{(i,j)v}$ denotes how much does are received from beamlet (i,j) at unit intensity. Figure 1.8 provides a simple illustrative example.



$$\begin{bmatrix} 1 & 4 & 4 & 4 & 2 \\ 2 & 3 & 3 & 1 & 2 \\ 0 & 2 & 3 & 2 & 2 \\ 0 & 1 & 3 & 3 & 1 \\ 1 & 3 & 3 & 1 & 1 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Fig. 1.8: Example of received dose with $y = (0.4, 0.2, 0.4, 0, \ldots, 0)$, $R = 5$, $D_{(i,j)v} = 1$.

Since the dose received at each voxel has some prescribed lower and upper threshold values. We can define objective function f by penalizing the weighted sum of doses falling outside the pre-specified lower and upper bounds. In particular, we have

$$f(\mathbf{z}) := \sum_{v \in \mathscr{V}} \left\{ \underline{w}_v \left[\underline{T}_v - z_v\right]_+^2 + \overline{w}_v \left[z_v - \overline{T}_v\right]_+^2 \right\},$$

where $[\cdot]_+$ denotes $\max\{0,\cdot\}$, and $\underline{T}_v$ and $\overline{T}_v$ are pre-specified lower and upper dose thresholds.

Putting the objective function and constraints together, we have a basic formulation for the dose localization problem. In this formulation, we intend to minimize a convex objective function over a simplex constraint.

$$\begin{aligned} \min \quad & f(\mathbf{z}) := \tfrac{1}{N_v}\textstyle\sum_{v\in\mathscr{V}} \left\{\underline{w}_v \, [\underline{T}_v - z_v]_+^2 + \overline{w}_v \, [z_v - \overline{T}_v]_+^2\right\} \\ \text{s.t.} \quad & z_v = \textstyle\sum_{a\in\mathscr{A}}\sum_{k\in\mathscr{K}_a} R\hat{D}_v^{k,a} y^{k,a}, \\ & \textstyle\sum_{a\in\mathscr{A}}\sum_{k\in K_a} y^{k,a} = 1, \\ & y^{k,a} \geq 0. \end{aligned} \tag{1.3.1}$$

Here $\hat{D}_v^{k,a} := \sum_{i=1}^{m}\sum_{j=1}^{n} D_{(i,j)v}\, x_{ij}^{k,a}$.

Problem (1.3.1) appears to be a relatively easy problem. The challenge, however, comes from its high dimensionality. Since the number of apertures in each angle increases exponentially with m. For a problem with 180 angles and 10×10 grids, the dimension is 180×45^{10}. It is impossible to compute the full gradient of ∇f. This fact excludes any algorithms that require the computation of full gradient information at each iteration.

High dimensionality is not the only challenge we have. In order to make the formulation more practical, we need to consider two additional types of constraints. Firstly, we need to introduce risk averse constraints that help to to generate treatment plans satisfying certain clinical criteria. These criteria are usually specified as value at risk (VaR) constraints. For example, “PTV56:V56$\geq$ 95%" means that the percentage of voxels in structure PTV56 that receive at least 56 Gy dose should be $\geq 95\%$. Similarly we need to avoid overdoses. For instance, “PTV68: V74.8$\leq$ 10%" says that the percentage of voxels in structure PTV68 that receive more than 74.8 Gy dose should be $\leq 10\%$. In our formulation, we suggest to use Conditional Value at Risk (CVaR) as a convex approximation for value at risk constraints.

Secondly, in practice we prefer to have a small number of angles in order to avoid frequently adjusting the patient’s position, which will reduce the treatment time. We suggest to incorporate a group sparsity constraint to handle this issue. Viewing the y variables from each angle as a group, we add the requirement that the summation of the group l_∞ norm should be smaller than a threshold value. Intuitively, this constraint will encourage the selection of apertures in those angles K_a that have already contained some nonzero $y^{k,a}$, $k \in K_a$ (see Figure 1.9).

Putting all these pieces together, we now have a complete problem formulation.

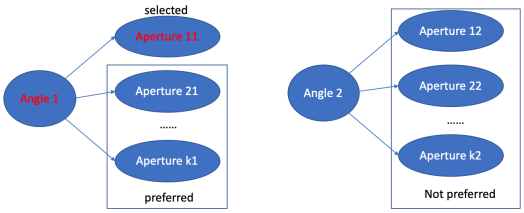


Fig. 1.9: Group sparsity.

$$\begin{array}{ll}
\min & f(\mathbf{z}) := \frac{1}{N_v}\sum_{v\in\mathscr{V}} \underline{w}_v \, [\underline{T}_v - z_v]_+^2 + \overline{w}_v \, [z_v - \overline{T}_v]_+^2 \\
\text{s.t.} & z_v = \sum_{a\in\mathscr{A}}\sum_{k\in\mathscr{K}_a} R\hat{D}_v^{k,a} y^{k,a}, \\
& -\tau_i + \frac{1}{p_i N_i}\sum_{v\in S_i}[\tau_i - z_v]_+ \le -b_i, \forall i \in \mathrm{UD}, \\
& \tau_i + \frac{1}{p_i N_i}\sum_{v\in S_i}[z_v - \tau_i]_+ \le b_i, \forall i \in \mathrm{OD}, \\
& \sum_{a\in\mathscr{A}} \max_{k\in\mathscr{K}_a} y^{k,a} \le \Phi, \\
& \sum_{a\in\mathscr{A}}\sum_{k\in K_a} y^{k,a} = 1, \\
& y^{k,a} \ge 0, \tau_i \in [\underline{\tau}_i, \overline{\tau}_i].
\end{array}$$

Here OD and UD denote the set of overdose and underdose clinical criteria. In addition to high problem dimensionality, the above problem formulation has complicated function constraints, which further complicated its solutions.

1.4 General Formulation

A typical *Mathematical Programming* problem is given in the form of

$$\begin{array}{ll}
\text{minimize} & \\
\qquad f(x) & [\,\text{objective}\,] \\
\text{subject to} & \\
h_i(x) = 0,\, i = 1,\dots,m & \begin{bmatrix}\text{equality} \\ \text{constraints}\end{bmatrix} \\
g_j(x) \le 0,\, j = 1,\dots,k & \begin{bmatrix}\text{inequality} \\ \text{constraints}\end{bmatrix} \\
\quad x \in X & [\,\text{domain}\,]
\end{array} \tag{1.4.1}$$

In (1.4.1), a *solution* $x \in \mathbb{R}^n$ represents a candidate decision, and the *constraints* express restrictions on the meaningful decisions. These restrictions can be bounds on the resources, the definition of a probability vector (see Section 1.2, or risk averse requirement (see Section 1.3). The *objective* to be minimized represents the losses (minus profit) associated with a decision.

To solve problem (1.4.1) means to find its *optimal solution* x^*, that is, a *feasible* solution satisfying all the constraints

$$h_i(x^*) = 0 \forall i;\ g_j(x^*) \leq 0 \forall j;\ x^* \in X,$$

such that its objective is smaller than or equal to that at any other feasible solutions, i.e.,

$$f(x^*) \leq f(x)$$

for any x satisfying

$$h_i(x) = 0 \forall i;\ g_j(x) \leq 0 \forall j;\ x \in X.$$

In *combinatorial* (or *discrete*) optimization, the domain X is a discrete set, such as the set of all integral or $0/1$ vectors. In contrast to this, in *continuous* optimization, X is a "continuum" set like the entire $\mathbb{R}^n$,a *box* $\{x : a \leq x \leq b\}$, or *simplex*, etc., and the objective and the constraints are (at least) continuous on X. In *linear programming*, $X = \mathbb{R}^n$ and the objective and the constraints are linear functions of x. In contrast to this, for *nonlinear continuous optimization*, (some of) the objectives and the constraints are nonlinear.

The goals of our course is to present

a) basic theory of continuous optimization, with emphasis on *existence* and *uniqueness* of optimal solutions and their *characterization* (i.e., necessary and/or sufficient optimality conditions);
b) basic algorithms for building approximate optimal solutions to continuous optimization problems.

The mathematical foundation of optimization theory is given by *Convex Analysis*, which is a specific combination of of real analysis and geometry unified by and focusing on investigating convexity-related notions. We will start with this fundamental topic in continuous optimization in next chapter.

Chapter 2
Convex Sets

2.1 Definition and Examples

We begin with the definition of the notion of a convex set.

Definition 2.1. A set $X \subseteq \mathbb{R}^n$ is said to be convex if it contains all of its segments, that is

$$\lambda x + (1-\lambda)y \in X, \quad \forall (x,y,\lambda) \in X \times X \times [0,1]. \tag{2.1.1}$$

Note that the point $\lambda x + (1-\lambda)y$ is called a convex combination of x and y. Figure 2.1 show the examples of a convex set (left) and a nonconvex set (right).

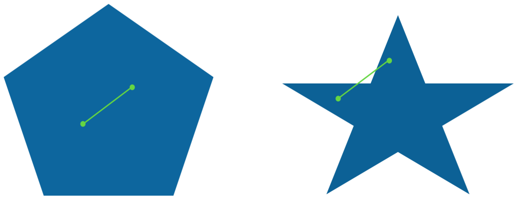


Fig. 2.1: Convex vs. nonconvex sets

We can see some immediate examples of convex sets.

a) An n-dimensional Euclidean space, $\mathbb{R}^n$. Given $x, y \in \mathbb{R}^n$, we must have $\lambda x + (1-\lambda)y \in \mathbb{R}^n$.
b) The empty set $\emptyset$. This set is defined as being convex by convention, since we can not identify any two points in $\emptyset$ violating (2.1.1).

We now give a few more nontrivial examples of convex sets that are widely used in optimization.

Linear and Affine Subspaces

Recall that a *linear subspace* L is a nonempty subset in $\mathbb{R}^n$ such that for any $x,y \in L$ and $\lambda \in \mathbb{R}$, we have we have $x+y \in L$ and $\lambda x \in L$. Geometrically, L is a special plane that passes through the origin.

A *linear subspace* can be constructed by the so-called *inner representation*. Given a set of vectors X, the linear span of X, denoted by $\text{Lin}(X)$, is comprised of all the linear combinations of X defined as $\sum_{i=1}^{k} \lambda_i x_i$, where $\lambda_i \in \mathbb{R}$, $i = 1,\ldots,k$ and k may depend on x. Let $d \leq n$ be the maximum number of linearly independent vectors in X, and $\{x_1,\ldots,x_d\}$ be a set of basis vectors, $\text{Lin}(X)$ can be written succinctly as the linear span of of basis vectors $\{x_1,\ldots,x_d\} \subseteq X$ and $d \leq n$ is also called the dimension of L. It is known that $\text{Lin}(X)$ is the smallest subspace containing X.

A more convenient way to to check the convexity of L is through its outer representation. More specifically, L can be written as the solution set to a homogeneous linear system, i.e., $\{x \in \mathbb{R}^n | \langle a_i, x\rangle = 0, i = 1,\ldots,r\}$. Here $r = n-d$ and $\{a_i, i = 1,\ldots,r\}$ is an arbitrary basis of the orthogonal complement of L in $\mathbb{R}^n$, denoted by $L^{\perp}$. Using this representation, the convexity of L can be easily verified by definition.

Given a linear subspace L and a vector $a \in \mathbb{R}^n$, an *affine space* M is defined as $M := a + L$. Geometrically, it represents a plane obtained by shifting L by a fixed vector a. The linear subspace L used in this decomposition is uniquely determined by M, given by $L = M - M = \{x - y : x,y \in M\}$. The shifting vector a is not uniquely defined and can be chosen arbitrarily from M. The affine dimension of M is defined as the dimension of L. Using the outer presentation of the linear subspace L, we can derive an outer representation of $M = \{y : \langle a_i, y\rangle = \langle a_i, a\rangle\}$. Therefore, M must be the solution set of a non-homogeneous linear system. The convexity of M follows easily from this representation and the definition in (2.1.1).

Polyhedron

An affine space of dimension $n-1$ in $\mathbb{R}^n$ is called a *hyperplane*. It can be written as the solution set of one linear equation $\{x \in \mathbb{R}^n : \langle a_1, x\rangle = b_1\}$. A hyperplane separates $\mathbb{R}^n$ into two halfspaces, denoted by $\{x \in \mathbb{R}^n : \langle a_1, x\rangle \geq b_1\}$ and $\{x \in \mathbb{R}^n : \langle a_1, x\rangle \leq b_1\}$.

A polyhedron is defined as the intersection of a finite or infinite number of halfspaces. Equivalently, it is the solution set of an arbitrary (finite or infinite) system of linear inequalities given by $P = \{x \in \mathbb{R}^n : \langle a_i, x\rangle \leq b_i, i \in I\}$.

The convexity of polyhedron directly follows from (2.1.1). Indeed, let $x,y \in P$, then for any $i \in I$, we have $\langle a_i, x\rangle \leq b_i$ and $\langle a_i, y\rangle \leq b_i$. For any $\lambda \in [0,1]$, we have

$$\langle a_i, \lambda x + (1-\lambda)y\rangle = \lambda\langle a_i, x\rangle + (1-\lambda)\langle a_i, y\rangle \leq \lambda b_i + (1-\lambda)b_i = b_i.$$

Using the same argument, we can show that the set $P_0 = \{x \in \mathbb{R}^n : \langle a_i, x\rangle < b_i, i \in I\}$ is convex. While P is a closed set (as it contains the limit point of any convergent sequences), P_0 is not a closed set. We will see later that every closed and convex set $X \subseteq \mathbb{R}^n$ is the solution set of an appropriate countable system of non-strict linear inequalities, i.e., $X = \{x \in \mathbb{R}^n : \langle a_i, x\rangle \leq b_i, i = 1, 2, \ldots\}$.

Polyhedral Representations

By definition, a *polyhedral* set $X \subset \mathbb{R}^n$ is a set which can be represented as

$$X = \{x : Ax \leq b\},$$

that is, as the solution set of a finite system of nonstrict linear inequalities. A *polyhedral representation* of a set $X \subset \mathbb{R}^n$ is a representation of X of the form:

$$X = \{x : \exists w : Px + Qw \leq r\}.$$

In other words, a representation of X is the *a projection* onto the space of x-variables of a polyhedral set $X^+ = \{[x;w] : Px + Qw \leq r\}$ in the space of x, w-variables.

Below we state a few examples of polyhedral representations.

a) The set $X = \{x \in \mathbb{R}^n : \sum_i |x_i| \leq 1\}$ admits the polyhedral representation

$$X = \left\{ x \in \mathbb{R}^n : \exists w \in \mathbb{R}^n : \begin{array}{c} -w_i \leq x_i \leq w_i, \\ 1 \leq i \leq n, \\ \sum_i w_i \leq 1 \end{array} \right\}.$$

b) The set

$$X = \big\{x \in \mathbb{R}^6 : \max[x_1, x_2, x_3] + 2\max[x_4, x_5, x_6] \\ \leq x_1 - x_6 + 5\big\}$$

admits the polyhedral representation

$$X = \left\{ x \in \mathbb{R}^6 : \exists w \in \mathbb{R}^2 : \begin{array}{l} x_1 \leq w_1, x_2 \leq w_1, x_3 \leq w_1 \\ x_4 \leq w_2, x_5 \leq w_2, x_6 \leq w_2 \\ w_1 + 2w_2 \leq x_1 - x_6 + 5 \end{array} \right\}.$$

A natural question we may have is whether a polyhedrally represented set is polyhedral. More specifically, let X be given by a polyhedral representation:

$$X = \{x \in \mathbb{R}^n : \exists w : Px + Qw \leq r\},$$

i.e., as the projection of the solution set

$$Y = \{[x;w] : Px + Qw \leq r\} \tag{2.1.2}$$

of a finite system of linear inequalities in variables x, w onto the space of x-variables. Is it true that X is polyhedral, i.e., X is a solution set of finite system of linear inequalities in variables x only?

This question will be answered positively below through the development of the so-called Fourier-Motzkin elimination procedure.

Elimination step: eliminating a *single* slack variable. Given set (2.1.2), assume that $w = [w_1;...;w_m]$ is nonempty, and let Y^+ be the projection of Y on the space of variables $x, w_1, ..., w_{m-1}$:

$$Y^+ = \{[x; w_1; ...; w_{m-1}] : \exists w_m : Px + Qw \leq r\}.$$

Let us show that Y^+ is polyhedral. Indeed, let us split the linear inequalities $p_i^T x + q_i^T w \leq r$, $1 \leq i \leq I$, defining Y into three groups:

a) The coefficient at w_m is 0.
b) The coefficient at w_m is > 0.
c) The coefficient at w_m is < 0.

Then

$$\begin{array}{rl} Y = \big\{x \in \mathbb{R}^n : \exists w = [w_1; ...; w_m] : & \\ a_i^T x + b_i^T [w_1; ...; w_{m-1}] \leq c_i, & i \text{ is in group a)} \\ w_m \leq a_i^T x + b_i^T [w_1; ...; w_{m-1}] + c_i, & i \text{ is in group b)} \\ w_m \geq a_i^T x + b_i^T [w_1; ...; w_{m-1}] + c_i, & i \text{ is in group c)}\big\} \end{array}$$

which implies that

$$\begin{array}{l} Y^+ = \big\{[x; w_1; ...; w_{m-1}] : \\ a_i^T x + b_i^T [w_1; ...; w_{m-1}] \leq c_i, \ i \text{ is in group a)} \\ a_\mu^T x + b_\mu^T [w_1; ...; w_{m-1}] + c_\mu \geq a_\nu^T x + b_\nu^T [w_1; ...; w_{m-1}] + c_\nu \\ \qquad \text{whenever } \mu \text{ is in group b) and } \nu \text{ is in group c)}\big\} \end{array}$$

and thus Y^+ is polyhedral. Now that the projection

$$Y^+ = \{[x; w_1; ...; w_{m-1}] : \exists w_m : [x; w_1; ...; w_m] \in Y\}$$

of the polyhedral set $Y = \{[x, w] : Px + Qw \leq r\}$ is polyhedral, iterating the process, we conclude that the set $X = \{x : \exists w : [x, w] \in Y\}$ is polyhedral.

Therefore, we arrive at the following conclusion.

Theorem 2.1. *Every polyhedrally representable set is polyhedral.*

Now let us consider an linear optimization problem of

$$\text{Opt} = \max_x \left\{c^T x : Ax \leq b\right\} \tag{2.1.3}$$

Observe that the set of values of the objective at feasible solutions can be represented as

$$\begin{aligned} T &= \{\tau \in \mathbb{R} : \exists x : Ax \leq b, c^T x - \tau = 0\} \\ &= \{\tau \in \mathbb{R} : \exists x : Ax \leq b, c^T x \leq \tau, c^T x \geq \tau\}, \end{aligned}$$

which means that T *is polyhedrally representable*. By Theorem 2.1, T is polyhedral, i.e., T can be represented by a finite system of linear inequalities *in variable τ only*.

It immediately follows that *if T is nonempty and is bounded from above, T has the largest element.* Thus, we have proved the following important consequence of Theorem 2.1.

Corollary 2.1. *A feasible and bounded linear optimization problem admits an optimal solution and thus is solvable.*

Fourier-Motzkin Elimination Scheme also suggests a finite algorithm for solving an LO program using the following procedure. First, we apply the scheme to get a representation of T by a finite system S of linear inequalities in variable τ. Second, we analyze S to find out whether the solution set is nonempty and bounded from above, and when it is the case, to find out the optimal value $\mathrm{Opt} \in T$ of the program. Third, use the Fourier-Motzkin elimination scheme in the backward fashion to find x such that $Ax \leq b$ and $c^T x = \mathrm{Opt}$, thus recovering an optimal solution to the problem of interest. Unfortunately, the resulting algorithm is completely impractical, since the number of inequalities we should handle at a step usually rapidly grows with the number of steps and can become astronomically large when eliminating just tens of variables.

Unit Balls

A real-valued function $\|x\|$ on $\mathbb{R}^n$ is a norm if

$$\begin{aligned} \|x\| \geq 0, \forall x \in \mathbb{R}^n; \|x\| = 0 \text{ iff } x = 0 \\ \|\lambda x\| = |\lambda| \|x\|, \forall x \in \mathbb{R}^n \text{ and } \lambda \in \mathbb{R} \\ \|x+y\| \leq \|x\| + \|y\|. \end{aligned}$$

To show that a ball defined by an arbitrary norm, $\{x \in \mathbb{R}^n | \|x\| \leq 1\}$ (e.g., the l_2 norm $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$ or l_1 norm $\|x\|_1 = \sum_{i=1}^n |x_i|$ balls) is convex, it suffices to apply the Triangular inequality and the positive homogeneity associated with a norm. Suppose that $\|x\| \leq 1, \|y\| \leq 1$ and $\lambda \in [0,1]$. Then

$$\|\lambda x + (1-\lambda)y\| \leq \|\lambda x\| + \|(1-\lambda)y\| = \lambda \|x\| + (1-\lambda)\|y\| \leq 1.$$

Notice that except some popular norms listed above, there is a more general characterization of the unit ball of a norm. A set $V \in \mathbb{R}^n$ is the unit ball of a norm iff V is (a) convex and symmetric w.r.t. 0, i.e., $V = -V$, (b) bounded and closed, and (c) contains a neighbourhood of the origin.

Ellipsoid

Let Q be an $n \times n$ positive definite and symmetric matrix ($Q \succ 0$), the center $a \in \mathbb{R}^n$ and radius $r > 0$ be given. An ellipsoid in $\mathbb{R}^n$ is defined as

$$X := \{x : (x-a)^T Q(x-a) \leq r^2\}.$$

To show X is convex, we can write

$$(x-a)^T Q(x-a) = [(x-a)^T Q^{1/2}][Q^{1/2}(x-a)] = \|Q^{1/2}(x-a)\|_2^2 = \|x-a\|_Q^2.$$

Hence, X is a $\|\cdot\|_Q$-ball and is therefore a convex set.

ε-neighbourhood of Convex Set

Let M be a nonempty convex set in $\mathbb{R}^n$, $\|\cdot\|$ be a norm and $\varepsilon \geq 0$. Then the set

$$X := \{x : \mathrm{dist}_{\|\cdot\|}(x, M) \equiv \inf_{y \in M} \|x-y\| \leq \varepsilon\}$$

is convex.

To show this statement, first observe that $x \in X$ if and only for every $\varepsilon' > \varepsilon$ there exists $y \in M$ such that $\|x-y\| \leq \varepsilon'$. For any $x, y \in X$, $\lambda \in [0,1]$, and any $\varepsilon' > \varepsilon$, there exists $u, v \in M$, such that $\|x-u\| \leq \varepsilon'$ and $\|y-v\| \leq \varepsilon'$. Setting $w = \lambda u + (1-\lambda)v$, we conclude

$$\begin{aligned} \|\lambda x + (1-\lambda)y - w\| &= \|\lambda(x-u) + (1-\lambda)(y-v)\| \\ &\leq \lambda\|x-u\| + (1-\lambda)\|y-v\| \leq \varepsilon', \end{aligned}$$

which implies that X is convex.

Convex Combinations and Convex Hulls

Let $x_1, \ldots, x_m \in \mathbb{R}^n$ be given. $y = \sum_{i=1}^m \lambda_i x_i$ is call their convex combination if $\lambda_i \geq 0$ and $\sum_{i=1}^m = 1$.

We claim that a set $X \in \mathbb{R}^n$ is convex iff it is closed w.r.t. convex combinations, i.e.,

$$X_i \in X, \lambda_i \geq 0, \textstyle\sum_{i=1}^m \lambda_i = 1 \Rightarrow \sum_{i=1}^m \lambda_i x_i \in X. \tag{2.1.4}$$

To show the sufficient condition is straightforward. One just need to fix $m = 2$ and use the definition of convexity.

We need to prove the necessarily condition by using induction. (2.1.4) holds obviously for $m = 1$. Suppose it is true when the number of terms is given by m.

Then

$$\begin{aligned}\textstyle\sum_{i=1}^{m+1}\lambda_i x_i &= \lambda_{m+1}x_{m+1} + \textstyle\sum_{i=1}^{m}\lambda_i x_i \\ &= \lambda_{m+1}x_{m+1} + \tfrac{1-\lambda_{m+1}}{\sum_{i=1}^{m}\lambda_i}\textstyle\sum_{i=1}^{m}\lambda_i x_i \\ &\in X,\end{aligned}$$

where the second identity follows from $1-\lambda_{m+1} = \sum_{i=1}^{m}\lambda_i$, and the last inclusion follows from the fact that $\sum_{i=1}^{m}\lambda_i x_i / \sum_{i=1}^{m}\lambda_i$ by our induction hypothesis.

It can be easily checked by the definition of convex sets that the intersection $\cap_{\alpha\in A}X_\alpha$ of an arbitrary family of convex set $\{X_\alpha\}_{\alpha\in A}$ in $\mathbb{R}^n$ is convex. Now let $X \subset \mathbb{R}^n$ be an arbitrary set. Then among convex sets containing X (which do exist, e.g. $\mathbb{R}^n$), there exists the smallest one, namely, the intersection of all convex sets containing X. We use the *convex hull* $\mathrm{Conv}(X)$ to denote this smallest convex set containing X.

The following simple result shows the inner construction of a convex hull.

Proposition 2.1. *Let a subset $X \subseteq \mathbb{R}^n$ (not necessarily convex) be given, and let $\hat{X}$ denote the set of all convex combination of points in X. Then* $\mathrm{Conv}(X) = \hat{X}$.

Proof. First note that every convex set which contains X must contain any convex combination of point from X. Therefore, $\hat{X} \subseteq \mathrm{Conv}(X)$.

To show the opposite site, we only need to show that the set $\hat{X} \supseteq X$ is convex since $\mathrm{Conv}(X)$ is the smallest one containing X. This is immediate. Let $x, y \in \hat{X}$. Then $x = \sum_{i=1}^{m}u_i x_i$ and $y = \sum_{i=1}^{n}v_i y_i$, with $x_i \in X, i = 1,\ldots,m$, $y_i \in X, i = 1,\ldots,n$, $\sum_{i=1}^{m}u_i = 1$, $u_i \geq 0$, $\sum_{i=1}^{m}v_i = 1$, $v_i \geq 0$. For some $\lambda \in [0,1]$,

$$\lambda x + (1-\lambda)y = \textstyle\sum_{i=1}^{m}\lambda u_i x_i + \sum_{i=1}^{n}(1-\lambda)v_i y_i \in \hat{X}.$$

■

Simplex

We have introduce the affine subspace from the geometric point of view together with its outer representation using non-homogenous linear systems. We now review an inner representation of an affine subspace, which will play an important role later in the theory of convex sets.

For an nonempty set $Y \subseteq \mathbb{R}^n$, the affine hull $\mathrm{Aff}(Y)$ is defined as the set comprised of all *affine combinations* of elements of Y defined as $y = \sum_{i=0}^{k}\lambda_i y_i$. Here $y_i \in Y$, k may depend on y, $\lambda_i \in \mathbb{R}$ and $\sum_i \lambda_i = 1$. To see that this is an affine subspace, fix a point $y_0 \in Y$, we have

$$y = y_0 - (1-\lambda_0)y_0 + \textstyle\sum_{i=1}^{k}\lambda_i y_i = y_0 + \sum_{i=1}^{k}\lambda_i(y_i - y_0).$$

Since the last term $\sum_{i=1}^{k}\lambda_i(y_i - y_0)$ denotes the linear subspace spanned by $Y - y_0$, $\mathrm{Aff}(Y)$ must be an affine subspace by definition. In fact, since $\mathrm{Ln}(Y - y_0)$ is the smallest linear subspace containing $Y - y_0$, $\mathrm{Aff}(Y)$ is also the smallest affine subspace containing Y. Let $\{y_1 - y_0, \dots, y_d - y_0\}$ denotes the basis vectors in $\mathrm{Ln}(Y - y_0)$, with d being its dimension. It then follows that the generic element of Y can be more compactly written as

$$y = y_0 + \textstyle\sum_{i=1}^{d}(y_i - y_0).$$

We say that $y_0, \dots, y_d$ are affinely independent if $y_1 - y_0, \dots, y_d - y_0$ are linearly independent.

An m-dimensional simplex Δ with vertices $x_0,...,x_m$ is defined as the convex hull of $m+1$ affinely independent points $x_0,...,x_m$:

$$\Delta = \Delta(x_0,...,x_m) = \mathrm{Conv}(\{x_0,...,x_m\}).$$

A few example are given as follows. a) 2-dimensional simplex is given by 3 points not belonging to a line and is the triangle with vertices at these points; b) Let $e_1,...,e_n$ be the standard basic orths in $\mathbb{R}^n$. These n points are affinely independent, and the corresponding $(n-1)$-dimensional simplex is the *standard simplex* $\Delta_n = \{x \in \mathbb{R}^n : x \geq 0, \sum\limits_i x_i = 1\}$; and c) adding to $e_1,...,e_n$ the vector $e_0 = 0$, we get $n+1$ affine independent points. The corresponding n-dimensional simplex is $\Delta_n^+ = \{x \in \mathbb{R}^n : x \geq 0, \sum\limits_i x_i \leq 1\}$. Note that Simplex with vertices $x_0,...,x_m$ is convex since it is the convex hull of a set), and every point from the simplex is a convex combination of the vertices with the coefficients uniquely defined by the point.

Cone

A subset K of $\mathbb{R}^n$ is conic if $K \neq \emptyset$ and $tx \in K$ for any $x \in K$ and $t \geq 0$. A convex conic set is called a cone.

A few examples of cone are given below. a) *Nonnegative orthant*

$$\mathbb{R}_+^n = \{x \in \mathbb{R}^n : x \geq 0\};$$

b) *Lorentz cone*

$$\mathbb{L}^n = \{x \in \mathbb{R}^n : x_n \geq \sqrt{x_1^2 + ... + x_{n-1}^2}\};$$

c) *Semidefinite cone* $\mathbb{S}_+^n$. This cone “lives” in the space $\mathbb{S}^n$ of $n \times n$ symmetric matrices and is comprised of all positive semidefinite symmetric $n \times n$ matrices;
d) *The solution set* $\{x : a_\alpha^T x \leq 0 \forall \alpha \in \mathscr{A}\}$ of an arbitrary (finite or infinite) *homogeneous* system of *nonstrict* linear inequalities is a *closed* cone. In particular, so is a *polyhedral cone* $\{x : Ax \leq 0\}$. It is worth noting that every *closed* cone in $\mathbb{R}^n$ is the solution set of a countable system of nonstrict *homogeneous* linear inequalities.

Below we provide a different characterization of a cone.

Proposition 2.2. *A nonempty subset $K \subseteq \mathbb{R}^n$ is a cone iff*

a) K is conic: $x \in K, t \geq 0 \Rightarrow tx \in K$;
b) K is closed w.r.t. addition: $x, y \in K \Rightarrow x + y \in K$.

Proof. $\Rightarrow$: Let K be convex and $x, y \in K$, Then $\frac{1}{2}(x+y) \in K$ by convexity, and since K is conic, we also have $x + y \in K$. Thus, a convex conic set is closed w.r.t. addition.

$\Leftarrow$: Let K be conic and closed w.r.t. addition. In this case, a convex combination $\lambda x + (1-\lambda)y$ of vectors x, y from K is the sum of the vectors λx and $(1-\lambda)y$ and thus belongs to K, since K is closed w.r.t. addition. Thus, a conic set which is closed w.r.t. addition is convex ■

Cones form an extremely important class of convex sets with properties "parallel" to those of general convex sets. For example,

- Intersection of an arbitrary family of cones again is a cone. As a result, *for every nonempty set X, among the cones containing X there exists the smallest cone* $\mathrm{Cone}(X)$, *called the conic hull of X.*
- A nonempty set is a cone iff it is closed w.r.t. taking *conic* combinations of its elements (i.e., linear combinations with nonnegative coefficients).
- The conic hull of a nonempty set X is exactly the set of all conic combinations of elements of X.

2.2 "Calculus" of Convex Sets

The following operations preserve convexity of sets.

1. **Intersection:** If $X_\alpha \subset \mathbb{R}^n$, $\alpha \in \mathscr{A}$, are convex sets, so is $\bigcap\limits_{\alpha \in \mathscr{A}} X_\alpha$.
2. **Direct product:** If $X_\ell \subset \mathbb{R}^{n_\ell}$ are convex sets, $\ell = 1, ..., L$, so is the set
$$\begin{aligned} X &= X_1 \times ... \times X_L \\ &\equiv \{x = (x^1, ..., x^L) : x^\ell \in X_\ell, 1 \leq \ell \leq L\} \\ &\subset \mathbb{R}^{n_1 + ... + n_L}. \end{aligned}$$
3. **Taking weighted sums:** Let $X_1, ..., X_L$ be nonempty convex subsets in $\mathbb{R}^n$ and $\lambda_1, ..., \lambda_L$ be reals. Then the set
$$\begin{aligned} &\lambda_1 X_1 + ... + \lambda_L X_L \\ \equiv &\{x = \lambda_1 x_1 + ... + \lambda_L x_\ell : x_\ell \in X_\ell, 1 \leq \ell \leq L\} \end{aligned}$$
is convex.
4. **Affine image:** Let $X \subset \mathbb{R}^n$ be convex and $x \mapsto \mathscr{A}(x) = Ax + b$ be an affine mapping from $\mathbb{R}^n$ to $\mathbb{R}^k$. Then the image of X under the mapping – the set
$$\mathscr{A}(X) = \{y = Ax + b : x \in X\}$$

is convex.

5. **Inverse affine image:** Let $X \subset \mathbb{R}^n$ be convex and $y \mapsto \mathscr{A}(y) = Ay + b$ be an affine mapping from $\mathbb{R}^k$ to $\mathbb{R}^n$. Then the inverse image of X under the mapping – the set
$$\mathscr{A}^{-1}(X) = \{y : Ay + b \in X\}$$
is convex.

All these statements can be easily checked by using the definition of convexity.

2.3 Topological Properties of Convex Sets

Recall that a set $X \subset \mathbb{R}^n$ is called *closed*, if X contains the limits of all converging sequences of its points:

$$x_i \in X \ \& \ x_i \to x, i \to \infty \Rightarrow x \in X.$$

It is *open*, if it contains, along with every of its points x, a ball of a positive radius centered at x:
$$x \in X \Rightarrow \exists r > 0 : \{y : \|y - x\|_2 \leq r\} \subset X.$$

For example, the solution set of an arbitrary system of *nonstrict* linear inequalities $\{x : a_\alpha^T x \leq b_\alpha\}$ is closed, while the solution set of a *finite* system of *strict* linear inequalities $\{x : Ax < b\}$ is open.

Also there are a few important facts about open and closed set.

A. X is closed iff $\mathbb{R}^n \backslash X$ is open.
B. The intersection of an arbitrary family of closed sets and the union of a finite family of closed sets are closed.
B$'$.The union of an arbitrary family of open sets and the intersection of a finite family of open sets are open.

Observe that B' is equivalent to B in view of the identity that $\mathbb{R}^n \backslash (A \cap B) = (\mathbb{R}^n \backslash A) \cup (\mathbb{R}^n \backslash B)$. From **B** it follows that *the intersection of all closed sets containing a given set X is closed.* The smallest closed set containing X, called *the closure* $\mathrm{cl}X$ of X is exactly the set of limits of all converging sequences of points of X:

$$\mathrm{cl}X = \{x : \exists x_i \in X : x = \lim_{i \to \infty} x_i\}.$$

From **B$'$** it follows that *the union of all open sets contained in a given set X is open.* The largest open set contained in X, called *the interior* $\mathrm{int}X$ of X is exactly the set of all *interior* points of X – points x belonging to X along with balls of positive radii centered at the points:

$$\mathrm{int}X = \{x : \exists r > 0 : \{y : \|y - x\|_2 \leq r\} \subset X\}.$$

Let $X \subset \mathbb{R}^n$. Then $\mathrm{int}X \subset X \subset \mathrm{cl}X$. The "difference" $\partial X = \mathrm{cl}X \backslash \mathrm{int}X$ is called the *boundary* of X; boundary always is closed (as the intersection of the closed sets $\mathrm{cl}X$ and the complement of $\mathrm{int}X$).

In general, the discrepancy between $\mathrm{int}X$ and $\mathrm{cl}X$ can be pretty large. For example, let $X \subset \mathbb{R}^1$ be the set of irrational numbers in $[0,1]$. Then $\mathrm{int}X = \emptyset$, $\mathrm{cl}X = [0,1]$, so that $\mathrm{int}X$ and $\mathrm{cl}X$ differ dramatically. Fortunately, a *convex* set is perfectly well approximated by its closure (and by interior, if the latter is nonempty) as shown in the following proposition.

Proposition 2.3. *Let $X \subset \mathbb{R}^n$ be a nonempty convex set. Then*

a) Both $\mathrm{int}X$ *and* $\mathrm{cl}X$ *are convex.*
b) If $\mathrm{int}X$ *is nonempty, then* $\mathrm{int}X$ *is dense in* $\mathrm{cl}X$. *Moreover,*

$$x \in \mathrm{int}X,\, y \in \mathrm{cl}X \Rightarrow \lambda x + (1-\lambda)y \in \mathrm{int}X \;\forall \lambda \in (0,1]. \tag{2.3.1}$$

Proof. We first show part a). To prove that $\mathrm{int}X$ is convex, note that for every two points $x, y \in \mathrm{int}X$ there exists a common $r > 0$ such that the balls B_x, B_y of radius r centered at x and y belong to X. Since X is convex, for every $\lambda \in [0,1]$, X contains the set $\lambda B_x + (1-\lambda)B_y$, which clearly is nothing but the ball of the radius r centered at $\lambda x + (1-\lambda)y$. Thus, $\lambda x + (1-\lambda)y \in \mathrm{int}X$ for all $\lambda \in [0,1]$.
Similarly, to prove that $\mathrm{cl}X$ is convex, assume that $x, y \in \mathrm{cl}X$, so that $x = \lim_{i\to\infty} x_i$ and $y = \lim\limits_{i\to\infty} y_i$ for appropriately chosen $x_i, y_i \in X$. Then for $\lambda \in [0,1]$ we have

$$\lambda x + (1-\lambda)y = \lim_{i\to\infty} \underbrace{[\lambda x_i + (1-\lambda)y_i]}_{\in X},$$

so that $\lambda x + (1-\lambda)y \in \mathrm{cl}X$ for all $\lambda \in [0,1]$.

To prove part b), it suffices to prove (2.3.1). Indeed, let $\bar{x} \in \mathrm{int}X$ (the latter set is nonempty). Every point $x \in \mathrm{cl}X$ is the limit of the sequence $x_i = \frac{1}{i}\bar{x} + \left(1 - \frac{1}{i}\right)x$. Given (2.3.1), all points x_i belong to $\mathrm{int}X$, thus $\mathrm{int}X$ is dense in $\mathrm{cl}X$.

Now to show (2.3.1), Let $x \in \mathrm{int}X$, $y \in \mathrm{cl}X$, $\lambda \in (0,1]$. Let us prove that $\lambda x + (1-\lambda)y \in \mathrm{int}X$. Since $x \in \mathrm{int}X$, there exists $r > 0$ such that the ball B of radius r centered at x belongs to X. Since $y \in \mathrm{cl}X$, there exists a sequence $y_i \in X$ such that $y = \lim_{i\to\infty} y_i$. Now let

$$\begin{aligned} B^i &= \lambda B + (1-\lambda)y_i \\ &= \{\underbrace{z = [\lambda x + (1-\lambda)y_i]}_{z_i} + \lambda h : \|h\|_2 \le r\} \\ &\equiv \{z = z_i + \delta : \|\delta\|_2 \le r' = \lambda r\}. \end{aligned}$$

Since $B \subset X$, $y_i \in X$ and X is convex, the sets B^i (which are balls of radius $r' > 0$ centered at z_i) are contained in X. Since $z_i \to z = \lambda x + (1-\lambda)y$ as $i \to \infty$, all these balls, starting with certain number, contain the ball B' of radius $r'/2$ centered at z. Thus, $B' \subset X$, i.e., $z \in \mathrm{int} X$. ■

Let X be a convex set. It may happen that $\mathrm{int} X = \emptyset$ (e.g., X is a segment in 3D). In this case, interior definitely does not approximate X and $\mathrm{cl} X$. So how should we do with this? A natural way to overcome this difficulty is to pass to *relative* interior, which is nothing but the interior of X *taken w.r.t. the affine hull* $\mathrm{aff}(X)$ *of* X rather than to $\mathbb{R}^n$. This affine hull, geometrically, is just certain $\mathbb{R}^m$ with $m \le n$; replacing, if necessary, $\mathbb{R}^n$ with this $\mathbb{R}^m$, we arrive at the situation where $\mathrm{int} X$ is nonempty.

To implement the outlined idea, we need to the following definition.

Definition 2.2 (relative interior and relative boundary). Let X be a nonempty convex set and M be the affine hull of X. The *relative interior* $\mathrm{rint} X$ is the set of all point $x \in X$ such that a ball *in* M of a positive radius, centered at x, is contained in X:

$$\mathrm{rint} X = \{x : \exists r > 0 s.t. \{y \in \mathrm{aff}(X), \|y - x\|_2 \le r\} \subset X\}.$$

The *relative boundary* of X is, by definition, $\mathrm{cl} X \backslash \mathrm{rint} X$.

We have the following important result for nonempty convex sets.

Proposition 2.4. *Let $X \subset \mathbb{R}^n$ be a nonempty convex set. Then* $\mathrm{rint} X \neq \emptyset$.

Proof. By Linear Algebra, whenever $x \in \mathbb{R}^n$ is nonempty, one can find in X an affine basis for the affine hull $\mathrm{aff}(X)$ of X. In other words, there exists $x_0, x_1, ..., x_m \in X$ so that every $x \in \mathrm{aff}(X)$ admits a representation

$$x = \sum_{i=0}^{m} \lambda_i x_i, \sum_i \lambda_i = 1$$

and the coefficients in this representation are uniquely defined by x.

When $x_i \in X$, $i = 0, 1, ..., m$, form an affine basis in $\mathrm{aff}(X)$, the system of linear equations

$$\begin{aligned} \sum_{i=0}^{m} \lambda_i x_i &= x \\ \sum_{i=0}^{m} \lambda_i &= 1 \end{aligned}$$

in variables λ has a *unique* solution whenever $x \in \mathrm{aff}(X)$. Since this solution is unique, it, again by Linear Algebra, depends continuously on $x \in \mathrm{aff}(X)$. In particular, when $x = \bar{x} = \frac{1}{m+1} \sum_{i=0}^{m} x_i$, the solution is positive; by continuity, it remains positive when $x \in \mathrm{aff}(X)$ is close enough to $\bar{x}$. Therefore,

$$\begin{aligned} &\exists r > 0 : x \in \mathrm{aff}(X), \|x - \bar{x}\|_2 \le r \Rightarrow \\ &x = \sum_{i=0}^{m} \lambda_i(x) x_i \text{ with } \textstyle\sum_i \lambda_i(x) = 1 \text{ and } \lambda_i(x) > 0 \end{aligned}$$

It follows that when X is convex, $\bar{x} \in \operatorname{rint} X$. ■

In view of the above result, by replacing, if necessary, the original "universe" $\mathbb{R}^n$ with a smaller *geometrically similar* universe (i.e. $\operatorname{aff}(X)$ or geometrically, $\mathbb{R}^m$ with certain $m \leq n$), we can reduce investigating an *arbitrary* nonempty convex set X to the case where this set has a nonempty interior (which is nothing but the *relative* interior of X). In particular, our results for the "full-dimensional" case in Proposition 2.3 imply that for a nonempty convex set X, both $\operatorname{rint} X$ and $\operatorname{cl} X$ are convex sets such that

$$\emptyset \neq \operatorname{rint} X \subset X \subset \operatorname{cl} X \subset \operatorname{aff}(X)$$

and $\operatorname{rint} X$ is dense in $\operatorname{cl} X$. Moreover, whenever $x \in \operatorname{rint} X$, $y \in \operatorname{cl} X$ and $\lambda \in (0,1]$, one has

$$\lambda x + (1-\lambda) y \in \operatorname{rint} X.$$

Now, we discuss how to construct the $\operatorname{cl} X$. Let X be convex and $\bar{x} \in \operatorname{rint} X$. As we know,

$$\lambda \in [0,1], y \in \operatorname{cl} X \Rightarrow y_\lambda = \lambda \bar{x} + (1-\lambda) y \in X.$$

It follows that *in order to pass from* X *to its closure* $\operatorname{cl} X$, *it suffices to pass to "radial closure"* as follows. For every direction $0 \neq d \in \operatorname{aff}(X) - \bar{x}$, let $T_d = \{t \geq 0 : \bar{x} + td \in X\}$. Note that T_d is a convex subset of $\mathbb{R}_+$ which contains all small enough positive t's. A few cases may happen.

- If T_d is unbounded or is a bounded segment: $T_d = \{t : 0 \leq t \leq t(d) < \infty\}$, the intersection of $\operatorname{cl} X$ with the ray $\{\bar{x} + td : t \geq 0\}$ is exactly the same as the intersection of X with the same ray.
- If T_d is a bounded half-segment: $T_d = \{t : 0 \leq t < t(d) < \infty\}$, the intersection of $\operatorname{cl} X$ with the ray $\{\bar{x} + td : t \geq 0\}$ is larger than the intersection of X with the same ray by exactly one point, namely, $\bar{x} + t(d)d$. Adding to X these "missing points" for all d, we arrive at $\operatorname{cl} X$.

2.4 Caratheodory's Theorem

Let M be affine subspace in $\mathbb{R}^n$, so that $M = a + L$ for a linear subspace L. The *linear* dimension of L is called the *affine* dimension $\dim M$ of M.

Examples: The affine dimension of a singleton is 0. The affine dimension of $\mathbb{R}^n$ is n. The affine dimension of an affine subspace $M = \{x : Ax = b\}$ is $n - \mathrm{rank}\,(A)$. For a nonempty set $X \subset \mathbb{R}^n$, the *affine dimension* $\dim X$ *of* X is exactly the affine dimension of the affine hull $\mathrm{aff}\,(X)$ of X.

Theorem 2.2. *Let* $\emptyset \neq X \subset \mathbb{R}^n$. *Then every point* $x \in \mathrm{conv}\,(X)$ *is a convex combination of* at most $\dim(X) + 1$ *points of* X.

Proof. We will go through the following few steps.

1^0. We should prove that if x is a convex combination of finitely many points $x_1, ..., x_k$ of X, then x is a convex combination of at most $m+1$ of these points, where $m = \dim(X)$. Replacing, if necessary, $\mathbb{R}^n$ with $\mathrm{aff}\,(X)$, it suffices to consider the case of $m = n$.

2^0. Consider a representation of x as a convex combination of $x_1, ..., x_k$ *with minimum possible number of nonzero coefficients*; it suffices to prove that this number is $\leq n+1$. Let, on the contrary, the "minimum representation" of x

$$x = \sum_{i=1}^{p} \lambda_i x_i, [\lambda_i \geq 0, \sum_i \lambda_i = 1]$$

has $p > n+1$ terms.

3^0. Consider the homogeneous system of linear equations in p variables δ_i

$$\begin{cases} (a) \ \sum\limits_{i=1}^{p} \delta_i x_i = 0 \ [n \text{ linear equations}] \\ (b) \quad \sum\limits_{i} \delta_i = 0 \ [\text{single linear equation}] \end{cases}$$

Since $p > n+1$, this system has a nontrivial solution δ (i.e. $\delta \neq 0$). Observe that for every $t \geq 0$ one has

$$x = \sum_{i=1}^{p} \underbrace{[\lambda_i + t\delta_i]}_{\lambda_i(t)} x_i \ \& \ \sum_i \lambda_i(t) = 1.$$

Now consider how to represent x as a convex combination of a smaller number of x_i's by varying t.

- When $t = 0$, all coefficients $\lambda_i(t)$ are nonnegative.
- When $t \to \infty$, some of the coefficients $\lambda_i(t)$ go to $-\infty$ (indeed, otherwise we would have $\delta_i \geq 0$ for all i, which is impossible since $\sum\limits_i \delta_i = 0$ and not all δ_i are zeros).

It follows that the quantity

$$t_* = \max\{t : t \geq 0 \ \& \ \lambda_i(t) \geq 0 \forall i\}$$

is well defined; when $t = t_*$, all coefficients in the representation

$$x = \sum_{i=1}^{p} \lambda_i(t_*) x_i$$

are nonnegative, sum of them equals to 1, *and at least one of the coefficients $\lambda_i(t_*)$ vanishes*. This contradicts the assumption of minimality of the original representation of x as a convex combination of x_i. ■

We also have a conic version of the Caratheodory Theorem. We will leave its proof as an exercise.

Theorem 2.3. *Let $\emptyset \neq X \subset \mathbb{R}^n$. Then every vector $x \in \operatorname{cone}(X)$ is a conic combination of at most n vectors from X.*

It is worth noting that the bounds given by Caratheodory Theorems (usual and conic version) are sharp as shown in the following examples.

- For a simplex Δ with $m+1$ vertices $v_0, ..., v_m$ one has $\dim \Delta = m$, and it takes *all* the vertices to represent the barycenter $\frac{1}{m+1} \sum\limits_{i=0}^{m} v_i$ as a convex combination of the vertices.
- The conic hull of n standard basic orths in $\mathbb{R}^n$ is exactly the nonnegative orthant $\mathbb{R}^n_+$, and it takes *all* these vectors to get, as their conic combination, the n-dimensional vector of ones.

As an illustrative example of the above results, consider a supermarkets sell 99 different herbal teas; every one of them is certain blend of 26 herbs A,...,Z. Let x_i denote the fraction of each herb for a herbal tea. These herbal teas sit within an affine space of dimension 25, i.e., $\{x \in \mathbb{R}^{26} : \sum_{i=1}^{26} x_i = 1\}$.

In spite of such a variety of marketed blends, John is not satisfied with any one of them; the only herbal tea he likes is their mixture, in the proportion

$$1 : 2 : 3 : ... : 98 : 99$$

Once it occurred to John that in order to prepare his favorite tea, there is no necessity to buy all 99 marketed blends; a smaller number of them will do. With some arithmetics, John found a combination of 66 marketed blends which still allows to prepare his tea. Do you believe John's result can be improved?

2.5 Radon's Theorem

Theorem 2.4. *Let $x_1,...,x_m$ be $m \geq n+2$ vectors in $\mathbb{R}^n$. One can split these vectors into two nonempty and non-overlapping groups A and B such that*

$$\operatorname{conv}(A) \cap \operatorname{conv}(B) \neq \emptyset.$$

Proof. Consider the homogeneous system of linear equations in m variables δ_i:

$$\begin{cases} \sum\limits_{i=1}^{m} \delta_i x_i = 0 \text{ [}n\text{ linear equations]} \\ \sum\limits_{i=1}^{m} \delta_i = 0 \text{ [single linear equation]} \end{cases}$$

Since $m \geq n+2$, the system has a nontrivial solution δ. Setting $I = \{i : \delta_i > 0\}$, $J = \{i : \delta_i \leq 0\}$, we split the index set $\{1,...,m\}$ into two *nonempty* (due to $\delta \neq 0, \sum\limits_i \delta_i = 0$) groups such that

$$\sum_{i \in I} \delta_i x_i = \sum_{j \in J} [-\delta_j] x_j$$
$$\gamma = \sum_{i \in I} \delta_i = \sum_{j \in J} -\delta_j > 0$$

whence

$$\underbrace{\sum_{i \in I} \frac{\delta_i}{\gamma} x_i}_{\in \operatorname{conv}(\{x_i : i \in I\})} = \underbrace{\sum_{j \in J} \frac{-\delta_j}{\gamma} x_j}_{\in \operatorname{conv}(\{x_j : j \in J\})} .$$

■

2.6 Helley's Theorem

The following is a basic version of Helley's Theorem.

Theorem 2.5. *Let $A_1,...,A_M$ be convex sets in $\mathbb{R}^n$. Assume that every $n+1$ sets from the family have a point in common. Then all sets also have point in common.*

Proof. We prove the result based on induction in M. The base $M \leq n+1$ is trivially true. Now assume that for certain $M \geq n+1$ our statement hods true for every M-member family of convex sets, and let us prove that it holds true for $M+1$-member family of convex sets $A_1,...,A_{M+1}$.

- By inductive hypotheses, every one of the $M+1$ sets
$$B_\ell = A_1 \cap A_2 \cap ... \cap A_{\ell-1} \cap A_{\ell+1} \cap ... \cap A_{M+1}$$
is nonempty. Let us choose $x_\ell \in B_\ell$, $\ell = 1,...,M+1$.
- By Radon's Theorem, the collection $x_1,...,x_{M+1}$ can be split in two sub-collections with intersecting convex hulls. W.l.o.g., let the split be $\{x_1,...,x_{J-1}\} \cup \{x_J,...,x_{M+1}\}$, and let
$$z \in \mathrm{conv}\left(\{x_1,...,x_{J-1}\}\right) \bigcap \mathrm{conv}\left(\{x_J,...,x_{M+1}\}\right).$$

We claim that $z \in A_\ell$ for all $\ell \leq M+1$ and hence the result is proved. Indeed, the points $x_J, x_{J+1},...,x_{M+1}$ belong to the convex set A_ℓ for $\ell \leq J-1$, whence

$$z \in \mathrm{conv}\left(\{x_J,...,x_{M+1}\}\right) \subset A_\ell.$$

Moreover, the points $x_1,...,x_{J-1}$ belong to the convex set A_ℓ for $\ell \geq J$, whence

$$z \in \mathrm{conv}\left(\{x_1,...,x_{J-1}\}\right) \subset A_\ell.$$

■

We can refine Helley's theorem as follows.

Theorem 2.6. *Assume that $A_1,...,A_M$ are convex sets in $\mathbb{R}^n$ and that*

- *the union $A_1 \cup A_2 \cup ... \cup A_M$ of the sets belongs to an affine subspace P of affine dimension m.*
- *every $m+1$ sets from the family have a point in common.*

Then all the sets have a point in common.

Proof. We can think of A_j as of sets in P, or, which is the same, as sets in $\mathbb{R}^m$ and apply the Helley Theorem ■

When trying to extend Helley's Theorem from finite to infinite collections of convex sets A_α, $\alpha \in \mathscr{A}$, we meet two immediate obstacles. First, things can go wrong when the sets A_α are not closed. For example, for the collection $\{A_i = (0, 1/i)\}, i \geq 1$ of convex subsets of $\mathbb{R}$, the intersection of sets from every finite subcollection is nonempty, but the intersection of all A_i is empty. Second, things can go wrong when the intersections of sets from finite subcollections can “run to infinity," as is the case for collection $\{A_i = [i,\infty)\}, i \geq 1$ of convex subsets of $\mathbb{R}$. Here again intersection of sets from every finite subcollection is nonempty, but the intersection of all A_i is empty. It turns out that these are the only two obstacles for Helley Theorem to be applicable to infinite collections of convex sets.

Below we show a more general form of Helley’s theorem for an infinite collection of convex sets.

Theorem 2.7. *Let A_α, $\alpha \in \mathscr{A}$, be a family of convex sets in $\mathbb{R}^n$ such that every $n+1$ sets from the family have a point in common. Assume, in addition, that*

- *the sets A_α are closed.*
- *one can find finitely many sets $A_{\alpha_1},\dots,A_{\alpha_M}$ with a bounded intersection.*

Then all sets A_α, $\alpha \in \mathscr{A}$, have a point in common.

Proof. By Helley’s Theorem, every finite collection of the sets A_α has a point in common, and it remains to apply the following standard fact from Analysis:
Let B_α be a family of closed sets in $\mathbb{R}^n$ such that

- *every finite collection of the sets has a nonempty intersection;*
- *in the family, there exists finite collection with bounded intersection.*

Then all sets from the family have a point in common.

The proof of the Standard Fact is based upon the following fundamental property of $\mathbb{R}^n$: Every closed and bounded subset of $\mathbb{R}^n$ is a compact set. Recall two equivalent definitions of a compact set:

- A subset X in a metric space M is called compact, if from every sequence of points of X one can extract a sub-sequence converging to a point from X.
- A subset X in a metric space M is called compact, if from every open covering of X (i.e., from every family of open sets such that every point of X belongs to at least one of them) one can extract a finite sub-covering.

Now let B_α be a family of closed sets in $\mathbb{R}^n$ such that every finite sub-family of the sets has a nonempty intersection and at least one of these intersection, let it be B, is bounded. Let us prove that all sets B_α have a point in common. Assume that it is not the case. Then for every point $x \in B$ there exists a set B_α which does not contain x. Since B_α is closed, it does not intersect an appropriate open ball V_x centered at x. Note that the system $\{V_x : x \in B\}$ forms an open covering of B. By its origin, B is closed (as intersection of closed sets) and bounded and thus is a compact set. Therefore one can find a *finite* collection $V_{x_1},\dots,V_{x_M}$ which covers B. For every $i \leq M$, there exists a set B_{α_i} in the family which does not intersect V_{x_i}; therefore $\bigcap\limits_{i=1}^{M} B_{\alpha_i}$ does not intersect B. Since B itself is the intersection of finitely many sets in B_α, we see that the intersection of these finitely many sets from B_α used to define B and the constructed sets $B_{\alpha_1},\dots,B_{\alpha_M}$ is empty, which is a contradiction (to the assumption that every finite sub-family of the sets has a nonempty intersection). ■

Example 2.1. We are given a function $f(x)$ on a 7,000,000-point set $X \subset \mathbb{R}$. At every 7-point subset of X, this function can be approximated, within accuracy 0.001 at every point, by appropriate polynomial of degree 5. To approximate the function on the entire X, we want to use a spline of degree 5 (a piecewise polynomial function with pieces of degree 5). How many pieces do we need to get accuracy 0.001 at every point?

The answer is: *Just one. Indeed, let A_x, $x \in X$, be the set of coefficients of all polynomials of degree 5 which reproduce $f(x)$ within accuracy 0.001:*

$$A_x = \left\{ p = (p_0, ..., p_5) \in \mathbb{R}^6 : |f(x) - \sum_{i=0}^{5} p_i x^i| \leq 0.001 \right\}.$$

The set A_x is polyhedral and therefore convex, and we know that every $6+1=7$ sets from the family $\{A_x\}_{x \in X}$ have a point in common. By Helley Theorem, all sets A_x, $x \in X$, have a point in common, that is, there exists a single polynomial of degree 5 which approximates f within accuracy 0.001 at every point of X.

Example 2.2. Consider an optimization program

$$c_* = \left\{ c^T x : g_i(x) \leq 0, i = 1, ..., 2022 \right\}$$

with 11 variables $x_1, ..., x_{11}$. Assume that the constraints are convex, that is, every one of the sets

$$X_i = \{x : g_i(x) \leq 0\}, i = 1, ..., 2022$$

is convex. Assume also that the problem is solvable with optimal value 0. Clearly, when dropping one or more constraints, the optimal value can only decrease or remain the same. Is it possible to find a constraint such that dropping it, we preserve the optimal value? Two constraints which can be dropped simultaneously with no effect on the optimal value? Three of them?

In fact, *you can drop as many as $2022 - 11 = 2011$ appropriately chosen constraints without varying the optimal value!*

Assume, on the contrary, that every 11-constraint relaxation of the original problem has negative optimal value. Since there are finitely many such relaxations, there exists $\varepsilon < 0$ such that every problem of the form

$$\min_x \{c^T x : g_{i_1}(x) \leq 0, ..., g_{i_{11}}(x) \leq 0\} \tag{2.6.1}$$

has a feasible solution with the value of the objective $< -\varepsilon$. Since this problem has a feasible solution with the value of the objective equal to 0 (namely, the optimal solution of the original problem) and its feasible set is convex, problem (2.6.1) has a feasible solution x with $c^T x = -\varepsilon$, obtained by taking a convex combination of the two optimal solutions to the original problem and the reduced problem in (2.6.1). In other words, every 11 of the 2022 sets (each defined by one single constraint $g_i(x) \leq 0$ and $c^T x = -\varepsilon$)

$$Y_i = \{x : c^T x = -\varepsilon, g_i(x) \le 0\},\ i = 1, ..., 2022$$

have a point in common. The sets Y_i are convex (as intersections of convex sets X_i and an affine subspace). If $c \neq 0$, then these sets belong to affine subspace of affine dimension 10, and since every 11 of them intersect, all 2022 intersect; a point x from their intersection is a feasible solution of the original problem with $c^T x < 0$, which is impossible. When $c = 0$, the claim is evident: we can drop all 2022 constraints without varying the optimal value!

2.7 Separation Theorem

Separating Linear Form

Recall from linear algebra that every linear form $f(x)$ on $\mathbb{R}^n$ is representable via inner product:

$$f(x) = \langle f, x \rangle$$

for appropriate vector $f \in \mathbb{R}^n$ uniquely defined by the form. Nontrivial (not identically zero) forms correspond to nonzero vectors f. A *level set*

$$M = \{x : \langle f, x \rangle = a\} \tag{2.7.1}$$

of a *nontrivial* linear form on $\mathbb{R}^n$ is affine subspace of affine dimension $n-1$; vice versa, every affine subspace M of affine dimension $n-1$ in $\mathbb{R}^n$ can be represented by (2.7.1) with appropriately chosen $f \neq 0$ and a; f and a are defined by M up to multiplication by a common nonzero factor. $(n-1)$-dimensional affine subspaces in $\mathbb{R}^n$ are called *hyperplanes*.

Level set (2.7.1) of nontrivial linear form splits $\mathbb{R}^n$ into two parts:

$$M_+ = \{x : \langle f, x \rangle \ge a\} \ \text{ and } \ M_- = \{x : \langle f, x \rangle \le a\}$$

called *closed half-spaces* given by (f, a); the hyperplane M is the common boundary of these half-spaces. The interiors M_{++} of M_+ and M_{--} of M_- are given by

$$M_{++} = \{x : \langle f, x \rangle > a\} \ \text{ and } \ M_{--} = \{x : \langle f, x \rangle < a\}$$

and are called *open half-spaces* given by (f, a). We have

$$\mathbb{R}^n = M_- \bigcup M_+ \quad [M_- \bigcap M_+ = M]$$

and

$$\mathbb{R}^n = M_{--} \bigcup M \bigcup M_{++}.$$

Let S and T be two nonempty sets in $\mathbb{R}^n$. We say that a hyperplane M in (2.7.1) *separates* S and T, if $S \subset M_-$, $T \subset M_+$, and $S \cup T \not\subset M$. We say that a nontrivial linear form $\langle f, x\rangle$ separates S and T if, for properly chosen a, the hyperplane (2.7.1) separates S and T. Geometrically, whenever M separates S and T, S does not go above M, T does not go below M, and their union is not contained completely in M.

Below we give a few examples of separating hyperplanes.

Example 2.3. The linear form x_1 on $\mathbb{R}^2$ (with $a = 0$)

1) separates the sets

$$\begin{aligned} S &= \{x \in \mathbb{R}^2 : x_1 \le 0, x_2 \le 0\}, \\ T &= \{x \in \mathbb{R}^2 : x_1 \ge 0, x_2 \ge 0\}. \end{aligned}$$

2) separates the sets

$$\begin{aligned} S &= \{x \in \mathbb{R}^2 : x_1 \le 0, x_2 \le 0\}, \\ T &= \{x \in \mathbb{R}^2 : x_1 + x_2 \ge 0, x_2 \le 0\}. \end{aligned}$$

3) does <u>not</u> separate the sets

$$\begin{aligned} S &= \{x \in \mathbb{R}^2 : x_1 = 0, 1 \le x_2 \le 2\}, \\ T &= \{x \in \mathbb{R}^2 : x_1 = 0, -2 \le x_2 \le -1\}. \end{aligned}$$

4) separates the sets

$$\begin{aligned} S &= \{x \in \mathbb{R}^2 : x_1 = 0, 1 \le x_2 \le 2\}, \\ T &= \{x \in \mathbb{R}^2 : 0 \le x_1 \le 1, -2 \le x_2 \le -1\}. \end{aligned}$$

An important characterization of separating linear form is given as follows. A linear form $\langle f, x\rangle$ separates nonempty sets S and T if and only if

$$\sup_{x \in S} \langle f, x\rangle \le \inf_{y \in T} \langle f, y\rangle \quad \text{and} \quad \inf_{x \in S} \langle f, x\rangle < \sup_{y \in T} \langle f, y\rangle.$$

In this case, the hyperplanes associated with f separating S and T are exactly the hyperplanes

$$\{x : \langle f, x\rangle = a\} \text{ with } \sup_{x \in S} \langle f, x\rangle \le a \le \inf_{y \in T} \langle f, y\rangle.$$

Main Theorem

Separation theorems states that two nonempty convex sets can be separated iff their relative interiors do not intersect. In order to prove this fundamental results in convex analysis, we need to prepare a few technical results.

Lemma 2.1. *Let X be a convex set, $f(x) = \langle f, x\rangle \equiv f^T x$ be a linear form and $a \in$ rint$X$. Then*

$$f^T a = \max_{x \in X} f^T x \Leftrightarrow f(\cdot)\Big|_X = \text{const}.$$

Proof. Shifting X, we may assume $a = 0$. Let, on the contrary to what should be proved, $f^T x$ be non-constant on X, so that there exists $y \in X$ with $f^T y \neq f^T a = 0$. The case of $f^T y > 0$ is impossible, since $f^T a = 0$ is the maximum of $f^T x$ on X. Thus, $f^T y < 0$. The line $\{ty : t \in \mathbb{R}\}$ passing through 0 and through y belongs to $\mathrm{aff}(X)$; since $0 \in \mathrm{rint} X$, all points $z = -\varepsilon y$ on this line belong to X, provided that $\varepsilon > 0$ is small enough. At every point of this type, $f^T z > 0$, which contradicts the fact that $\max\limits_{x \in X} f^T x = f^T a = 0$. ■

The following result is well-known in real analysis. We put it here for the sake of completeness.

Lemma 2.2. *Every nonempty subset S in $\mathbb{R}^n$ is separable, meaning that one can find a sequence $\{x_i\}$ of points from S which is dense in S, i.e., is such that every point $x \in S$ is the limit of an appropriate subsequence of the sequence.*

Proof. Let $r_1, r_2, \ldots$ be the countable set of all rational vectors in $\mathbb{R}^n$. For every positive integer t, let $X_t \subset S$ be the countable set given by the following construction:

> We look, one after another, at the points $r_1, r_2, \ldots$ and for every point r_s check whether there is a point z in S which is at most at the distance $1/t$ away from r_s. If points z with this property exist, we take one of them and add it to X_t and then pass to r_{s+1}, otherwise directly pass to r_{s+1}.

Is is clear that every point $x \in S$ is at the distance at most $2/t$ from certain point of X_t. Indeed, since the rational vectors are dense in $\mathbb{R}^n$, there exists s such that r_s is at the distance $\leq 1/t$ from x. Therefore, when processing r_s, we definitely add to X_t a point z which is at the distance $\leq 1/t$ from r_s and thus is at the distance $\leq 2/t$ from x. By construction, the countable union $\bigcup\limits_{t=1}^{\infty} X_t$ of countable sets $X_t \subset S$ is a countable set in S, and this set is dense in S since every point $x \in S$ is at the distance at most $2/t$ from certain point of X_t. ■

We are now ready to prove the main result in this subsection.

Theorem 2.8. *Two nonempty convex sets S and T can be separated iff their relative interiors do not intersect.*

Proof. $\Rightarrow$: Suppose that $f^T x$ separates S and T so that $\sup\limits_{x \in S} f^T x \leq \inf\limits_{y \in T} f^T y$. Assume, on contrary to what should be proved, that $a \in \mathrm{rint} S \cap \mathrm{rint} T$. Since $a \in T$, we get $f^T a \geq \sup\limits_{x \in S} f^T x$, that is, $f^T a = \max\limits_{x \in S} f^T x$. By Lemma 2.1, $f^T x = f^T a$ for all $x \in S$. Moreover, since $a \in S$, we get $f^T a \leq \inf\limits_{y \in T} f^T y$, that is, $f^T a = \min\limits_{y \in T} f^T y$. By Lemma 2.1, $f^T y = f^T a$ for all $y \in T$. Thus, for any $z \in S \cup T$, we have $f^T z \equiv f^T a$, so that f does not separate S and T, which is a contradiction.

$\Leftarrow$: Assume that S and T are nonempty convex sets such that $\mathrm{rint} S \cap \mathrm{rint} T = \emptyset$. We will prove that S and T can be separated by a few steps.

1^0: Separating a point and a convex hull of a finite set. Let $S = \mathrm{conv}(\{b_1, \ldots, b_m\})$ and $T = \{b\}$ with $b \notin S$, and let us prove that S and T can be separated. Indeed,

$$S = \mathrm{conv}\,(\{b_1,...,b_m\}) = \{x : \exists\lambda \text{ s.t. } \lambda \geq 0, \sum_i \lambda_i = 1, x = \sum_i \lambda_i b_i\}$$

is polyhedrally representable and thus by Theorem 2.1, is polyhedral, i.e.,

$$S = \{x : p_l^T x \leq q_l, l \leq L\}.$$

Since $b \notin S$, for some $\bar{l}$, we have $p_{\bar{l}}^T b > q_{\bar{l}} \geq \sup_{x\in S} p_{\bar{l}}^T x$, which is the desired separation.

2^0: Separating a point and a convex set which does not contain the point. Let S be a nonempty convex set and $T = \{b\}$ with $b \notin S$, and let us prove that S and T can be separated. W.L.O.G., shifting S and T by $-b$ (which clearly does not affect the possibility of separating the sets), we can assume that $T = \{0\} \not\subset S$. Moreover, replacing, if necessary, $\mathbb{R}^n$ with $\mathrm{lin}\,(S)$, we may further assume that $\mathbb{R}^n = \mathrm{lin}\,(S)$. In view of Lemma 2.2, Let $\{x_i \in S\}$ be a sequence which is dense in S. Since S is convex and does not contain 0, we have

$$0 \notin \mathrm{conv}\,(\{x_1,...,x_i\})\ \forall i$$

whence by Step 1,

$$\exists f_i : 0 = f_i^T 0 > \max_{1\leq j\leq i} f_i^T x_j. \tag{2.7.2}$$

By scaling, we may assume that $\|f_i\|_2 = 1$. The sequence $\{f_i\}$ of unit vectors possesses a converging subsequence $\{f_{i_s}\}_{s=1}^{\infty}$; the limit f of this subsequence is, of course, a unit vector. By (2.7.2), for every fixed j and all large enough s we have $f_{i_s}^T x_j < 0$, whence

$$f^T x_j \leq 0\ \forall j. \tag{2.7.3}$$

Since $\{x_j\}$ is dense in S, (2.7.3) implies that $f^T x \leq 0$ for all $x \in S$, whence

$$\sup_{x\in S} f^T x \leq 0 = f^T 0. \tag{2.7.4}$$

In view of the above relation, all we need to prove is to verify that

$$\inf_{x\in S} f^T x < f^T 0 = 0.$$

Assuming the opposite, (2.7.4) would say that $f^T x = 0$ for all $x \in S$, which is impossible, since $\mathrm{lin}\,(S) = \mathbb{R}^n$ and f is nonzero.

3^0: Separating two non-intersecting nonempty convex sets. Suppose that S and T are nonempty convex sets which do not intersect. Let us prove that they can be separated. Let $\widehat{S} = S - T$ and $\widehat{T} = \{0\}$. The set $\widehat{S}$ clearly is convex and does not contain 0 (since $S \cap T = \emptyset$). By Step 2, $\widehat{S}$ and $\{0\} = \widehat{T}$ can be separated: there exists f such that

$$\left\{\begin{array}{l}\overbrace{\sup\limits_{x\in S,y\in T}[f^Tx-f^Ty]}^{\sup\limits_{x\in S}f^Ts-\inf\limits_{y\in T}f^Ty}\leq 0=\inf\limits_{z\in\{0\}}f^Tz\\ \\ \underbrace{\inf\limits_{x\in S,y\in T}[f^Tx-f^Ty]}_{\inf\limits_{x\in S}f^Tx-\sup\limits_{y\in T}f^Ty}<0=\sup\limits_{z\in\{0\}}f^Tz\end{array}\right.$$

whence

$$\sup_{x\in S}f^Tx\leq\inf_{y\in T}f^Ty \text{ and } \inf_{x\in S}f^Tx<\sup_{y\in T}f^Ty.$$

4^0: Completing the proof of Separation Theorem. Finally, suppose that S and T are nonempty convex sets with non-intersecting relative interiors. Let us prove that S and T can be separated. As we know, the sets $S'=\mathrm{rint}\,S$ and $T'=\mathrm{rint}\,T$ are convex and nonempty, and hence we are in the situation when these sets do not intersect. By Step 3, S' and T' can be separated: for properly chosen f, one has

$$\begin{array}{l}\sup\limits_{x\in S'}f^Tx\leq\inf\limits_{y\in T'}f^Ty\\ \inf\limits_{x\in S'}f^Tx<\sup\limits_{y\in T'}f^Ty\end{array}$$

Since S' is dense in S and T' is dense in T, inf's and sup's in the above relation remain the same when replacing S' with S and T' with T. Thus, f separates S and T. ■

Separation of sets S and T by linear form f^Tx is called *strict*, if

$$\sup_{x\in S}f^Tx<\inf_{y\in T}f^Ty$$

Theorem 2.9. *Let S and T be nonempty convex sets. These sets can be strictly separated iff they are at positive distance:*

$$\mathrm{dist}(S,T)=\inf_{x\in S,y\in T}\|x-y\|_2>0.$$

Proof. $\Rightarrow$: Suppose f strictly separate S and T. Let us prove that S,T are at positive distance. Otherwise we could find sequences $x_i\in S$, $y_i\in T$ with $\|x_i-y_i\|_2\to 0$ as $i\to\infty$, whence $f^T(y_i-x_i)\to 0$ as $i\to\infty$. It follows that the sets on the axis

$$\widehat{S}=\{a=f^Tx:x\in S\},\widehat{T}=\{b=f^Ty:y\in T\}$$

are at zero distance, which is a contradiction with

$$\sup_{a\in\widehat{S}}a<\inf_{b\in\widehat{T}}b.$$

$\Leftarrow$: Suppose S and T be nonempty convex sets which are at positive distance 2δ:

$$2\delta = \inf_{x\in S, y\in T} \|x-y\|_2 > 0.$$

Let

$$S^+ = S + \{z : \|z\|_2 \leq \delta\}$$

The sets S^+ and T are convex and do not intersect, and thus can be separated:

$$\sup_{x_+\in S^+} f^T x_+ \leq \inf_{y\in T} f^T y \qquad [f \neq 0]$$

Since

$$\begin{aligned} \sup_{x_+\in S^+} f^T x_+ &= \sup_{x\in S, \|z\|_2\leq\delta} [f^T x + f^T z] \\ &= [\sup_{x\in S} f^T x] + \delta\|f\|_2, \end{aligned}$$

we arrive at

$$\sup_{x\in S} f^T x < \inf_{y\in T} f^T y.$$

■

2.8 Supporting Planes

Let Q be a *closed* convex set in $\mathbb{R}^n$ and $\bar{x}$ be a point from the relative boundary of Q. A hyperplane

$$\Pi = \{x : f^T x = a\} \qquad [a \neq 0]$$

is called *supporting to Q at the point $\bar{x}$*, if the hyperplane separates Q and $\{\bar{x}\}$:

$$\begin{aligned} \sup_{x\in Q} f^T x &\leq f^T \bar{x} \\ \inf_{x\in Q} f^T x &< f^T \bar{x} \end{aligned}$$

Equivalently, $\Pi = \{x : f^T x = a\}$ supports Q at $\bar{x}$ if and only if the linear form $f^T x$ attains its maximum on Q, equal to a, at the point $\bar{x}$ and the form is non-constant on Q.

Proposition 2.5. *Let Q be a convex closed set in $\mathbb{R}^n$ and $\bar{x}$ be a point from the relative boundary of Q. Then*

a) There exist at least one hyperplane Π which supports Q at $\bar{x}$;
b) For every such hyperplane Π, the set $Q \cap \Pi$ has dimension less than the one of Q.

Proof. Existence of supporting plane is given by the Separation Theorem in Theorem 2.8 since

$$\bar{x} \notin \operatorname{rint} Q \Rightarrow (\{\bar{x}\} \equiv \operatorname{rint}\{\bar{x}\}) \cap \operatorname{rint} Q = \emptyset.$$

Furthermore,

$$Q \not\subseteq \Pi \Rightarrow \operatorname{aff}(Q) \not\subseteq \Pi \Rightarrow \operatorname{aff}(\Pi \cap Q) \subsetneqq \operatorname{aff}(Q).$$

If two *distinct* affine subspaces are embedded one into another, then the dimension of the embedded subspace is strictly less than the dimension of the embedding one. ■

2.9 Extreme Points*

We first give a few equivalent definitions of extreme points. Let Q be a convex set in $\mathbb{R}^n$ and $\bar{x}$ be a point of Q. The point is called *extreme*, if it is not a convex combination, with positive weights, of two points of X distinct from $\bar{x}$:

$$\begin{gathered} \bar{x} \in \operatorname{ext}(Q) \\ \Updownarrow \\ \{\bar{x} \in Q\} \;\&\; \left\{ \left\{ \begin{array}{c} u, v \in Q, \lambda \in (0,1) \\ \bar{x} = \lambda u + (1-\lambda)v \end{array} \right\} \Rightarrow u = v = \bar{x} \right\} \end{gathered}$$

Equivalently, a point $\bar{x} \in Q$ is extreme if and only if it is not the midpoint of a nontrivial segment in Q:

$$x \pm h \in Q \Rightarrow h = 0.$$

It is also equivalent to say that a point $\bar{x} \in Q$ is extreme if and only if the set $Q \backslash \{\bar{x}\}$ is convex. Here are a few examples for extreme points.

a) Extreme points of $[x,y]$ are the end points x and y.
b) Extreme points of $\triangle ABC$ are the vertices A, B and C.
c) Extreme points of the ball $\{x : \|x\|_2 \leq 1\}$ are the points $\{x : \|x\|_2 = 1\}$ on the boundary of the ball.

Krein-Milman's theorem below tells us when a convex set possess extreme points. Moreover, it shows that a *closed convex bounded* set Q is given by the convex hull of its extreme points. Before presenting this result, we need to prove two important technical results.

Lemma 2.3. *Let S be a closed convex set and $\Pi = \{x : f^T x = a\}$ be a hyperplane which supports S at certain point. Then*

$$\operatorname{ext}(\Pi \cap S) \subset \operatorname{ext}(S).$$

Proof. Let $\bar{x} \in \operatorname{ext}(\Pi \cap S)$. We should prove that $\bar{x} \in \operatorname{ext}(S)$. Assume, on the contrary, that $\bar{x}$ is a midpoint of a nontrivial segment $[u,v] \subset S$. Then $f^T \bar{x} = a = \max\limits_{x \in S} f^T x$, whence $f^T \bar{x} = \max\limits_{x \in [u,v]} f^T x$. A linear form can attain its maximum on a segment at the midpoint of the segment if and only if the form is constant on the segment. Thus, $a = f^T\bar{x} = f^T u = f^T v$, that is, $[u,v] \subset \Pi \cap S$. This contradicts with the assumption that $\bar{x}$ is an extreme point of $\Pi \cap S$. ■

Here is another important lemma.

Lemma 2.4. *Let S be a closed convex set such that $\{\bar{x}+th : t \geq 0\} \subset S$ for certain $\bar{x}$. Then*

$$\{x+th : t \geq 0\} \subset S\, \forall x \in S.$$

Proof. For every $s > 0$ and $x \in S$ we have

$$x + sh = \lim_{i\to\infty} \underbrace{[(1-s/i)x + (s/i)[\bar{x}+(i/s)h]]}_{\in S}.$$

Hence $x + sh \in S$ by the closeness of S. ■

Note that the set of all directions $h \in \mathbb{R}^n$ such that $\{x+th : t \geq 0\} \subset S$ for some (and then, for all) $x \in S$, is called the *recessive cone* $\mathrm{rec}\,(S)$ of closed convex set S. $\mathrm{rec}\,(S)$ indeed is a cone, and $S + \mathrm{rec}\,(S) = S$.

Theorem 2.10 (Krein-Milman). *Let Q be a closed convex and nonempty set in $\mathbb{R}^n$. Then*

a) Q possess extreme points if and only if Q does not contain lines;
b) If Q is bounded, then Q is the convex hull of its extreme points:

$$Q = \mathrm{conv}\,(\mathrm{ext}\,(Q))$$

so that every point of Q is convex combination of extreme points of Q.

Proof. We prove the results through a few steps.
1^0: If closed convex set Q does not contain lines, then $\mathrm{ext}\,(Q) \neq \emptyset$. In order to build an extreme point of Q, we apply the following *Purification algorithm.*

Initialization: Set $S_0 = Q$ and choose $x_0 \in Q$.
Step t: Given a nonempty closed convex set S_t which does not contain lines and is such that $\mathrm{ext}\,(S_t) \subset \mathrm{ext}\,(Q)$ and $x_t \in S_t$,
1) check whether S_t is a singleton $\{x_t\}$. If it is the case, terminate: $x_t \in \mathrm{ext}\,\{S_t\} \subset \mathrm{ext}\,(Q)$.
2) if S_t is not a singleton, find a point x_{t+1} on the relative boundary of S_t and build a hyperplane Π_t which supports S_t at x_{t+1}. More specifically, to find x_{t+1}, take a direction $h \neq 0$ parallel to $\mathrm{aff}\,(S_t)$. Since S_t does not contain lines, when moving from x_t either in the direction h, or in the direction $-h$, we eventually leave S_t, and thus cross the relative boundary of S_t. The intersection point is the desired x_{t+1}.
3) Set $S_{t+1} = S_t \cap \Pi_t$, replace t with $t+1$ and loop to 1).

By Lemma 2.3, we have

$$\mathrm{ext}\,(S_{t+1}) \subset \mathrm{ext}\,(S_t),$$

so that

$$\mathrm{ext}\,(S_t) \subset \mathrm{ext}\,(Q)\ \forall t.$$

Besides this, $\dim(S_{t+1}) < \dim(S_t)$, so that Purification algorithm does terminate with an extreme point.

2^0: If a closed convex set Q contains lines, it has no extreme points. By Lemma 2.4, If a closed convex set Q contains a line ℓ (both directions of a ray), then the parallel lines, passing through points of Q, also belong to Q. In particular, Q possesses no extreme points.

3^0: If a nonempty closed convex set Q is bounded, then $Q = \mathrm{conv}\,(\mathrm{ext}\,(Q))$. The inclusion $\mathrm{conv}\,(\mathrm{ext}\,(Q)) \subset Q$ is evident. Let us prove the opposite inclusion, i.e., prove that every point of Q is a convex combination of extreme points of Q based on induction on $k = \dim Q$. The base case $k = 0$ (Q is a singleton) is evident. Suppose the result holds for any dimension up to k. Now given $(k+1)$-dimensional closed and bounded convex set Q and a point $x \in Q$, we, as in the Purification algorithm, can represent x as a convex combination of two points x_+ and x_- *from the relative boundary of* Q. Let Π_+ be a hyperplane which supports Q at x_+, and let $Q_+ = \Pi_+ \cap Q$. As we know, Q_+ is a closed convex set such that

$$\dim Q_+ < \dim Q,\ \mathrm{ext}\,(Q_+) \subset \mathrm{ext}\,(Q),\ x_+ \in Q_+.$$

Invoking inductive hypothesis,

$$x_+ \in \mathrm{conv}\,(\mathrm{ext}\,(Q_+)) \subset \mathrm{conv}\,(\mathrm{ext}\,(Q)).$$

Similarly, $x_- \in \mathrm{conv}\,(\mathrm{ext}\,(Q))$. Since $x \in [x_-, x_+]$, we get $x \in \mathrm{conv}\,(\mathrm{ext}\,(Q))$. ■

Suppose that we are given a linear form $g^T x$ which is bounded from above on Q. Then in the Purification algorithm we can easily ensure that $g^T x_{t+1} \geq g^T x_t$. Thus, If Q is a nonempty closed set in $\mathbb{R}^n$ which does not contain lines and $f^T x$ is a linear form which is bounded above on Q, then for every point $x_0 \in Q$ there exists (and can be found by Purification) a point $\bar{x} \in \mathrm{ext}\,(Q)$ such that $g^T \bar{x} \geq g^T x_0$. In particular, if $g^T x$ attains its maximum on Q, then the maximizer can be found among extreme points of Q.

2.10 Exercises

1 Show that a set is convex if and only if its intersection with any line is convex. Show that a set is affine if and only if its intersection with any line is affine.

2 Voronoi description of halfspace. Let a and b be distinct points in $\mathbf{R}^n$. Show that the set of all points that are closer (in Euclidean norm) to a than b, i.e., $\{x \mid \|x-a\|_2 \leq \|x-b\|_2\}$, is a halfspace. Describe it explicitly as an inequality of the form $c^T x \leq d$. Draw a picture.

3 Which of the following sets S are polyhedra? If possible, express S in the form $S = \{x \mid Ax \leq b, Fx = g\}$

(a) $S = \{y_1 a_1 + y_2 a_2 \mid -1 \leq y_1 \leq 1, -1 \leq y_2 \leq 1\}$, where $a_1, a_2 \in \mathbf{R}^n$

(b) $S = \left\{x \in \mathbf{R}^n \mid x \geq 0, \quad \mathbf{1}^T x = 1, \sum_{i=1}^n x_i a_i = b_1, \sum_{i=1}^n x_i a_i^2 = b_2\right\}$, where $a_1, \ldots, a_n \in \mathbf{R}$ and $b_1, b_2 \in \mathbf{R}$.

(c) $S = \left\{x \in \mathbf{R}^n \mid x \geq 0, x^T y \leq 1 \text{ for all } y \text{ with } \|y\|_2 = 1\right\}$.

(d) $S = \left\{x \in \mathbf{R}^n \mid x \geq 0, x^T y \leq 1 \text{ for all } y \text{ with } \sum_{i=1}^n |y_i| = 1\right\}$.

4 Which of the following sets are convex?

(a) A slab, i.e., a set of the form $\left\{x \in \mathbf{R}^n \mid \alpha \leq a^T x \leq \beta\right\}$.

(b) A rectangle, i.e., a set of the form $\{x \in \mathbf{R}^n \mid \alpha_i \leq x_i \leq \beta_i, i = 1, \ldots, n\}$. A rectangle is sometimes called a hyperrectangle when $n > 2$.

(c) A wedge, i.e., $\left\{x \in \mathbf{R}^n \mid a_1^T x \leq b_1, a_2^T x \leq b_2\right\}$.

(d) The set of points closer to a given point than a given set, i.e.,

$$\{x \mid \|x - x_0\|_2 \leq \|x - y\|_2 \text{ for all } y \in S\}$$

where $S \subseteq \mathbf{R}^n$.

(e) The set of points closer to one set than another, i.e.,

$$\{x \mid \mathbf{dist}(x, S) \leq \mathbf{dist}(x, T)\},$$

where $S, T \subseteq \mathbf{R}^n$, and

$$\mathbf{dist}(x, S) = \inf\{\|x - z\|_2 \mid z \in S\}.$$

(f) The set $\{x \mid x + S_2 \subseteq S_1\}$, where $S_1, S_2 \subseteq \mathbf{R}^n$ with S_1 convex.

(g) The set of points whose distance to a does not exceed a fixed fraction θ of the distance to b, i.e., the set $\{x \mid \|x - a\|_2 \leq \theta \|x - b\|_2\}$. You can assume $a \neq b$ and $0 \leq \theta \leq 1$

5 Expanded and restricted sets. Let $S \subseteq \mathbf{R}^n$, and let $\|\cdot\|$ be a norm on $\mathbf{R}^n$.

(a) For $a \geq 0$ we define S_a as $\{x \mid \text{dist}(x, S) \leq a\}$, where $\text{dist}(x, S) = \inf_{y \in S} \|x - y\|$. We refer to S_a as S expanded or extended by a. Show that if S is convex, then S_a is convex.

(b) For $a \geq 0$ we define $S_{-a} = \{x \mid B(x, a) \subseteq S\}$, where $B(x, a)$ is the ball (in the norm $\|\cdot\|$), centered at x, with radius a. We refer to S_{-a} as S shrunk or restricted by a, since S_{-a} consists of all points that are at least a distance a from $\mathbf{R}^n \backslash S$. Show that if S is convex, then S_{-a} is convex.

6 Some sets of probability distributions. Let x be a real-valued random variable with $\mathbf{prob}\,(x = a_i) = p_i, i = 1, \ldots, n$, where $a_1 < a_2 < \cdots < a_n$. Of course $p \in \mathbf{R}^n$ lies in the standard probability simplex $P = \left\{p \mid \mathbf{1}^T p = 1, p \succeq 0\right\}$. Which of the following conditions are convex in p ? (That is, for which of the following conditions is the set of $p \in P$ that satisfy the condition convex?)

(a) $\alpha \leq \mathbf{E} f(x) \leq \beta$, where $\mathbf{E} f(x)$ is the expected value of $f(x)$, i.e., $\mathrm{E}\ f(x) = \sum_{i=1} p_i f(a_i)$ (The function $f : \mathbf{R} \to \mathbf{R}$ is given.)
(b) $\mathbf{prob}(x > \alpha) \leq \beta$.
(c) $\mathbf{E}\left|x^3\right| \leq \alpha \mathbf{E}|x|$.
(d) $\mathbf{E}x^2 \leq \alpha$.
(e) $\mathbf{E}x^2 \geq \alpha$.
(f) $\mathbf{var}(x) \leq \alpha$, where $\mathbf{var}(x) = \mathbf{E}(x - \mathbf{E}x)^2$ is the variance of x.
(g) $\mathbf{var}(x) \geq \alpha$.
(h) $\mathbf{quartile}(x) \geq \alpha$, where $\mathbf{quartile}(x) = \inf\{\beta \mid \mathbf{prob}(x \leq \beta) \geq 0.25\}$.
(i) $\mathbf{quartile}(x) \leq \alpha$.

7 Show that if S_1 and S_2 are convex sets in $\mathbf{R}^{m+n}$, then so is their partial sum

$$S = \{(x, y_1 + y_2) \mid x \in \mathbf{R}^m, y_1, y_2 \in \mathbf{R}^n, (x, y_1) \in S_1, (x, y_2) \in S_2\}$$

8 The set of separating hyperplanes. Suppose that C and D are disjoint subsets of $\mathbf{R}^n$. Consider the set of $(a,b) \in \mathbf{R}^{n+1}$ for which $a^T x \leq b$ for all $x \in C$, and $a^T x \geq b$ for all $x \in D$. Show that this set is a convex cone (which is the singleton $\{0\}$ if there is no hyperplane that separates C and D).

9 Supporting hyperplanes.

(a) Express the closed convex set $\left\{x \in \mathbf{R}_+^2 \mid x_1 x_2 \geq 1\right\}$ as an intersection of halfspaces.
(b) Let $C = \{x \in \mathbf{R}^n \mid \|x\|_\infty \leq 1\}$, the ℓ_∞-norm unit ball in $\mathbf{R}^n$, and let $\hat{x}$ be a point in the boundary of C. Identify the supporting hyperplanes of C at $\hat{x}$ explicitly.

10 Inner and outer polyhedral approximations. Let $C \subseteq \mathbf{R}^n$ be a closed convex set, and suppose that $x_1, \ldots, x_K$ are on the boundary of C. Suppose that for each $i, a_i^T (x - x_i) = 0$ defines a supporting hyperplane for C at x_i, i.e., $C \subseteq \left\{x \mid a_i^T (x - x_i) \leq 0\right\}$. Consider the two polyhedra

$$P_{\text{inner}} = \operatorname{conv}\{x_1, \ldots, x_K\}, \quad P_{\text{outer}} = \left\{x \mid a_i^T (x - x_i) \leq 0, i = 1, \ldots, K\right\}.$$

Show that $P_{\text{inner}} \subseteq C \subseteq P_{\text{outer.}}$ Draw a picture illustrating this.

11 Support function. The support function of a set $C \subseteq \mathbf{R}^n$ is defined as

$$S_C(y) = \sup\left\{y^T x \mid x \in C\right\}.$$

(We allow $S_C(y)$ to take on the value $+\infty$.) Suppose that C and D are closed convex sets in $\mathbf{R}^n$. Show that $C = D$ if and only if their support functions are equal.

12 Converse supporting hyperplane theorem. Suppose the set C is closed, has nonempty interior, and has a supporting hyperplane at every point in its boundary. Show that C is convex.

13 Separation of cones. Let K and $\tilde{K}$ be two convex cones whose interiors are nonempty and disjoint. Show that there is a nonzero y such that $y \in K^*, -y \in \tilde{K}^*$.

14 Mark in the following list the sets which are convex:

a. $\left\{x \in \mathbf{R}^2 : x_1 + i^2 x_2 \le 1, i = 1, \cdots, 10\right\}$
b. $\left\{x \in \mathbf{R}^2 : x_1^2 + 2ix_1x_2 + i^2x_2^2 \le 1, i = 1, \cdots, 10\right\}$
c. $\left\{x \in \mathbf{R}^2 : x_1^2 + ix_1x_2 + i^2x_2^2 \le 1, i = 1, \cdots, 10\right\}$
d. $\left\{x \in \mathbf{R}^2 : x_1^2 + 5x_1x_2 + 4x_2^2 \le 1\right\}$
e. $\left\{x \in \mathbf{R}^{10} : x_1^2 + 2x_2^2 + 3x_3^2 + \cdots + 10x_{10}^2 \le 2004x_1 - 2003x_2 + 2002x_3 - \cdots + 1996x_9 - 1995x_{10}\right\}$
f. $\left\{x \in \mathbf{R}^2 : \exp\{x_1\} \le x_2\right\}$
g. $\left\{x \in \mathbf{R}^2 : \exp\{x_1\} \ge x_2\right\}$
h. $\left\{x \in \mathbf{R}^n : \sum_{i=1}^n x_i^2 = 1\right\}$
i. $\left\{x \in \mathbf{R}^n : \sum_{i=1}^n x_i^2 \le 1\right\}$
j. $\left\{x \in \mathbf{R}^n : \sum_{i=1}^n x_i^2 \ge 1\right\}$
k. $\left\{x \in \mathbf{R}^n : \max_{i=1,\cdots,n} x_i \le 1\right\}$
l. $\left\{x \in \mathbf{R}^n : \max_{i=1,\cdots,n} x_i \ge 1\right\}$
m. $\left\{x \in \mathbf{R}^n : \max_{i=1,\cdots,n} x_i = 1\right\}$
n. $\left\{x \in \mathbf{R}^n : \min_{i=1,\cdots,n} x_i \le 1\right\}$
o. $\left\{x \in \mathbf{R}^n : \min_{i=1,\cdots,n} x_i \ge 1\right\}$
p. $\left\{x \in \mathbf{R}^n : \min_{i=1,\cdots,n} x_i = 1\right\}$

15 Which ones of the following three statements are true?

a. The convex hull of a closed set in $\mathbf{R}^n$ is closed
b. The convex hull of a closed convex set in $\mathbf{R}^n$ is closed
c. The convex hull of a closed and bounded set in $\mathbf{R}^n$ is closed and bounded

For true statements, present proofs; for wrong, give counterexamples.

Hint: Recall that a bounded and closed subset of $\mathbf{R}^n$ is compact and that there exists Caratheodory Theorem.

16 A cake contains 300 g[1)] of raisins (you may think of every one of them as of a 3*D* ball of positive radius). John and Jill are about to divide the cake according to the following rules:

- first, Jill chooses a point a in the cake;
- second, John makes a cut through a, that is, chooses a 2*D* plane Π passing through a and takes the part of the cake on one side of the plane (both Π and

the side are up to John, with the only restriction that the plane should pass through a); all the rest goes to Jill.

a. Prove that it may happen that Jill cannot guarantee herself 76 g of the raisins
b. Prove that Jill always can choose a in a way which guarantees her at least 74 g of the raisins
c. Consider n-dimensional version of the problem, where the raisins are n-dimensional balls, the cake is a domain in $\mathbf{R}^n$, and “a cut" taken by John is defined as the part of the cake contained in the half-space

$$\left\{x \in \mathbf{R}^n : e^T(x-a) \geq 0\right\},$$

where $e \neq 0$ is the vector (“inner normal to the cutting hyperplane") chosen by John. Prove that for every $\varepsilon > 0$, Jill can guarantee to herself at least $\frac{300}{n+1} - \varepsilon g$ of raisins, but in general cannot guarantee to herself $\frac{300}{n+1} + \varepsilon g$.

17 Prove the following Kirchberger's Theorem:
Assume that $X = \{x_1, \cdots, x_k\}$ and $Y = \{y_1, \cdots, y_m\}$ are finite sets in $\mathbf{R}^n$, with $k+m \geq n+2$, and all the points $x_1, \cdots, x_k, y_1, \cdots, y_m$ are distinct. Assume that for any subset $S \subset X \cup Y$ comprised of $n+2$ points the convex hulls of the sets $X \cap S$ and $Y \cap S$ do not intersect. Then the convex hulls of X and Y also do not intersect.
Hint: Assume, on contrary, that the convex hulls of X and Y intersect, so that

$$\sum_{i=1}^{k} \lambda_i x_i = \sum_{j=1}^{m} \mu_j y_j$$

for certain nonnegative $\lambda_i, \sum_i \lambda_i = 1$, and certain nonnegative $\mu_j, \sum_j \mu_j = 1$, and look at the expression of this type with the minimum possible total number of nonzero coefficients λ_i, μ_j.

18 Show that:

(a) The intersection $\cap_{i \in I} C_i$ of a collection $\{C_i \mid i \in I\}$ of cones is a cone.
(b) The Cartesian product $C_1 \times C_2$ of two cones C_1 and C_2 is a cone.

19 Let C_1 and C_2 be two nonempty convex sets such that $C_1 \subset C_2$.

(a) Give an example showing that $\mathrm{ri}\,(C_1)$ need not be a subset of $\mathrm{ri}\,(C_2)$.
(b) Assuming that the sets C_1 and C_2 have the same affine hull, show that $\mathrm{ri}\,(C_1) \subset \mathrm{ri}\,(C_2)$.
(c) Assuming that the set $\mathrm{ri}\,(C_1) \cap \mathrm{ri}\,(C_2)$ is nonempty, show that $\mathrm{ri}\,(C_1) \subset \mathrm{ri}\,(C_2)$
(d) Assuming that the set $C_1 \cap \mathrm{ri}\,(C_2)$ is nonempty, show that the set $\mathrm{ri}\,(C_1) \cap \mathrm{ri}\,(C_2)$ is nonempty.
(e) Show that the relative interior of a singleton $\{x_0\}$ is nonempty.

20 Let X_1 and X_2 be nonempty subsets of $\Re^n$, and let $X = \mathrm{conv}(X_1) + \mathrm{cone}(X_2)$. Show that every vector x in X can be represented in the form

$$x = \sum_{i=1}^{k} \alpha_i x_i + \sum_{i=k+1}^{m} \alpha_i y_i$$

where m is a positive integer with $m \leq n+1$, the vectors $x_1, \ldots, x_k$ belong to X_1, the vectors $y_{k+1}, \ldots, y_m$ belong to X_2, and the scalars $\alpha_1, \ldots, \alpha_m$ are nonnegative with $\alpha_1 + \cdots + \alpha_k = 1$. Furthermore, the vectors $x_2 - x_1, \ldots, x_k - x_1, y_{k+1}, \ldots, y_m$ are linearly independent.

21 (a) Let C_1 be a convex set with nonempty interior and C_2 be a nonempty convex set that does not intersect the interior of C_1. Show that there exists a hyperplane such that one of the associated closed halfspaces contains C_2, and does not intersect the interior of C_1.

(b) Show by an example that we cannot replace interior with relative interior in the statement of part (a).

22 Let C be a nonempty convex set in $\Re^n$, and let M be a nonempty affine set in $\Re^n$. Show that $M \cap \mathrm{ri}(C) = \varnothing$ is a necessary and sufficient condition for the existence of a hyperplane H containing M, and such that $\mathrm{ri}(C)$ is contained in one of the open halfspaces associated with H.

23 Let C_1 and C_2 be nonempty convex subsets of $\Re^n$ such that C_2 is a cone.

(a) Suppose that there exists a hyperplane that separates C_1 and C_2 properly. Show that there exists a hyperplane which separates C_1 and C_2 properly and passes through the origin.

(b) Suppose that there exists a hyperplane that separates C_1 and C_2 strictly. Show that there exists a hyperplane that passes through the origin such that one of the associated closed halfspaces contains the cone C_2 and does not intersect C_1.

24 Is the set $\{a \in \mathbf{R}^k | p(0) = 1, |p(t)| \leq 1 \text{ for } \alpha \leq t \leq \beta\}$, where

$$p(t) = a_1 + a_2 t + \cdots + a_k t^{k-1},$$

convex?

25 Let $\emptyset \neq X \subset \mathbb{R}^n$. Then every vector $x \in \mathrm{cone}(X)$ is a conic combination of *at most* n vectors from X.

Chapter 3
Convex Functions

3.1 Definition and examples

Let f be a real-valued function defined on a nonempty subset $\operatorname{dom} f$ in $\mathbb{R}^n$. f is called *convex*, if

a) $\operatorname{dom} f$ is a convex set,
b) for all $x, y \in \operatorname{dom} f$ and $\lambda \in [0,1]$ one has

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y). \tag{3.1.1}$$

Equivalently, let f be a real-valued function defined on a nonempty subset $\operatorname{dom} f$ in $\mathbb{R}^n$. The function is called convex, if its *epigraph* – the set

$$\operatorname{epi}\{f\} = \{(x,t) \in \mathbb{R}^{n+1} : f(x) \leq t\}$$

is a convex set in $\mathbb{R}^{n+1}$. We leave the proof for the equivalence of these two definitions as an exercise.

Note that the inequality in (3.1.1) is automatically satisfied when $x = y$ or when $\lambda = 0/1$. Thus, it says something only when the points x, y are distinct from each other and the point $z = \lambda x + (1-\lambda)y$ is a (relative) interior point of the segment $[x, y]$. What does (3.1.1) say in this case? Observe that $z = \lambda x + (1-\lambda)y = x + (1-\lambda)(y-x)$, whence

$$\|y-x\| : \|y-z\| : \|z-x\| = 1 : \lambda : (1-\lambda)$$

Therefore

$$\begin{gathered} f(z) \leq \lambda f(x) + (1-\lambda) f(y) \\ \Updownarrow \\ f(z) - f(x) \leq \underbrace{(1-\lambda)}_{\frac{\|z-x\|}{\|y-x\|}} (f(y) - f(x)) \\ \Updownarrow \\ \frac{f(z)-f(x)}{\|z-x\|} \leq \frac{f(y)-f(x)}{\|y-x\|} \end{gathered}$$

Similarly,

$$f(z) \le \lambda f(x) + (1-\lambda) f(y) \quad (*)$$
$$\Updownarrow$$
$$\underbrace{\lambda}_{\frac{\|y-z\|}{\|y-x\|}} (f(y) - f(x)) \le f(y) - f(z)$$
$$\Updownarrow$$
$$\frac{f(y)-f(x)}{\|y-x\|} \le \frac{f(y)-f(z)}{\|y-z\|}$$

We then conclude that f is convex iff for every three distinct points x, y, z such that $x, y \in \operatorname{dom} f$ and $z \in [x, y]$, we have $z \in \operatorname{dom} f$ and

$$\frac{f(z)-f(x)}{\|z-x\|} \le \frac{f(y)-f(x)}{\|y-x\|} \le \frac{f(y)-f(z)}{\|y-z\|}. \tag{3.1.2}$$

Some examples of convex or nonconvex functions are given as follows.

a) Functions convex on $\mathbb{R}$: x^2, x^4, x^6, $\exp\{x\}$.
b) Nonconvex functions on $\mathbb{R}$: x^3, $\sin(x)$.
c) Functions convex on $\mathbb{R}_+$: x^p, $p \ge 1$, $-x^p$, $0 \le p \le 1$, $x \ln x$.
d) Functions convex on $\mathbb{R}^n$: affine function $f(x) = f^T c$.
e) A norm $\|\cdot\|$ on $\mathbb{R}^n$ is a convex function.

3.2 Jensen's Inequality

Below we state the important Jensen's inequality.

Proposition 3.1. *Let $f(x)$ be a convex function. Then*

$$x_i \in \operatorname{dom} f, \lambda_i \ge 0, \sum_i \lambda_i = 1 \Rightarrow f(\sum_i \lambda_i x_i) \le \sum_i \lambda_i f(x_i)$$

Proof. The points $(x_i, f(x_i))$ belong to epi $\{f\}$. Since this set is convex, the point

$$(\sum_i \lambda_i x_i, \sum_i \lambda_i f(x_i)) \in \text{epi}\,\{f\}.$$

By definition of the epigraph, it follows that

$$f(\sum_i \lambda_i x_i) \le \sum_i \lambda_i f(x_i).$$

■

We can state Jensen's inequality in a more general form. Let f be convex, $\operatorname{dom} f$ be closed and f be continuous on $\operatorname{dom} f$. Consider a probability distribution $\pi(dx)$ supported on $\operatorname{dom} f$. Then

$$f(\mathbb{E}_\pi\{x\}) \le \mathbb{E}_\pi\{f(x)\}.$$

To illustrate the application of Jensen's Inequality, Let $p = \{p_i > 0\}_{i=1}^n$, $q = \{q_i > 0\}_{i=1}^n$ be two discrete probability distributions. We claim that the *Kullback-Liebler distance*

$$\sum_i p_i \ln \frac{p_i}{q_i}$$

between the distributions is ≥ 0. Indeed, the function $f(x) = -\ln x$, $\operatorname{dom} f = \{x > 0\}$, is convex. Setting $x_i = q_i/p_i$, $\lambda_i = p_i$ we have

$$\begin{aligned} 0 &= -\ln\left(\sum_i q_i\right) = f(\sum_i p_i x_i) \\ &\leq \sum_i p_i f(x_i) = \sum_i p_i(-\ln q_i/p_i) \\ &= \sum_i p_i \ln(p_i/q_i). \end{aligned}$$

3.3 Extended Real

What is the value of a convex function outside its domain? By convention, it is convenient to think that a convex function f is defined *everywhere* on $\mathbb{R}^n$ and takes real values *and value* $+\infty$. With this interpretation, f "remembers" its domain:

$$\begin{gathered} \operatorname{dom} f = \{x : f(x) \in \mathbb{R}\} \\ x \notin \operatorname{dom} f \Rightarrow f(x) = +\infty \end{gathered}$$

and the definition of convexity becomes

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \ \forall \begin{smallmatrix} x, y \in \mathbb{R}^n \\ \lambda \in [0,1] \end{smallmatrix},$$

where the arithmetics of $+\infty$ and reals is given by the rules

$$\begin{aligned} &+\infty \leq +\infty \\ &a \in \mathbb{R} \Rightarrow a + (+\infty) = (+\infty) + (+\infty) = +\infty \\ &0 \cdot (+\infty) = 0 \\ &\lambda > 0 \Rightarrow \lambda \cdot (+\infty) = +\infty. \end{aligned}$$

It should be noted that operations like $(+\infty) - (+\infty)$ or $(-5) \cdot (+\infty)$ are undefined.

3.4 Convexity-preserving Operations

In this section, we discuss a few operations that help to preserve the convexity of functions.

1. Taking conic combinations. If $f_i(x)$ are convex function on $\mathbb{R}^n$ and $\lambda_i \geq 0$, then the function $\sum\limits_i \lambda_i f_i(x)$ is convex.

2. Affine substitution of argument. If $f(x)$ is convex function on $\mathbb{R}^n$ and $x = Ay + b$ is an affine mapping from $\mathbb{R}^k$ to $\mathbb{R}^n$, then the function $g(y) = f(Ax + b)$ is convex on $\mathbb{R}^m$.

3. Taking supremum. If $f_\alpha(x)$, $\alpha \in \mathscr{A}$, is a family of convex function on $\mathbb{R}^n$, then the function $\sup\limits_{\alpha \in \mathscr{A}} f_\alpha(x)$ is convex. Indeed, $\mathrm{epi}\,\{\sup\limits_\alpha f_\alpha(\cdot)\} = \bigcap\limits_\alpha \mathrm{epi}\,\{f_\alpha(\cdot)\}$, and intersections of convex sets are convex.

4. Superposition Theorem. Let $f_i(x)$ be convex functions on $\mathbb{R}^n$, $i = 1, \ldots, m$, and $F(y_1, \ldots, y_m)$ be a convex and *monotone* function on $\mathbb{R}^m$. Then the function

$$g(x) = \begin{cases} F(f_1(x), \ldots, f_m(x)) & , x \in \mathrm{dom}\, f_i,\ \forall i \\ +\infty & , \text{otherwise} \end{cases}$$

is convex.

5. Partial minimization. Let $f(x,y)$ be a convex function of $z = (x,y) \in \mathbb{R}^n$, and let

$$g(x) = \inf_y f(x,y)$$

be $> -\infty$ for all x. Then the function $g(x)$ is convex. Indeed, g clearly takes real values and value $+\infty$. Let us check the Convexity Inequality

$$g(\lambda x' + (1-\lambda)x'') \leq \lambda g(x') + (1-\lambda) g(x'') \qquad [\lambda \in [0,1]]$$

There is nothing to check when $\lambda = 0$ or $\lambda = 1$, so let $0 < \lambda < 1$. In this case, there is nothing to check when $g(x')$ or $g(x'')$ is $+\infty$, so let $g(x') < +\infty$, $g(x'') < +\infty$. Since $g(x') < +\infty$, for every $\varepsilon > 0$ there exists y' such that $f(x', y') \leq g(x') + \varepsilon$. Similarly, there exists y'' such that $f(x'', y'') \leq g(x'') + \varepsilon$. Now,

$$\begin{aligned} & g(\lambda x' + (1-\lambda)x'') \\ \leq\ & f(\lambda x' + (1-\lambda)x'', \lambda y' + (1-\lambda)y'') \\ \leq\ & \lambda f(x', y') + (1-\lambda) f(x'', y'') \\ \leq\ & \lambda (g(x') + \varepsilon) + (1-\lambda)(g(x'') + \varepsilon) \\ =\ & \lambda g(x') + (1-\lambda) g(x'') + \varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, we get $g(\lambda x' + (1-\lambda)x'') \leq \lambda g(x') + (1-\lambda) g(x'')$.

6. Projective transformation. Let $f(x)$ be a convex function of $x \in \mathbb{R}^n$. Then the function $F(\alpha, x) = \alpha f(x/\alpha) : \{\alpha > 0\} \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is convex. Indeed, we need to verify that if $x, x' \in \mathbb{R}^n$, $\alpha, \alpha' > 0$ and $\lambda \in (0,1)$, the inequality

$$[\lambda\alpha + (1-\lambda)\alpha'] f([\lambda x + (1-\lambda)x'] / [\lambda\alpha + (1-\lambda)\alpha']) \leq \lambda \alpha f(x/\alpha) + (1-\lambda)\alpha' f(x/\alpha')$$

holds. The relations follows immediately from the convexity of f and the fact that $[\lambda\alpha+(1-\lambda)\alpha']/[\lambda\alpha+(1-\lambda)\alpha']=1$. As an illustration, the function $\alpha\ln(\alpha/\beta)$ is convex in the quadrant $\{\alpha>0,\beta>0\}$, as can be seen by applying the projective transformation to $\ln(1/\beta)=-\ln\beta$.

3.5 Detection of Convexity

First, we notice that convexity is one-dimensional property. A set $X\subset\mathbb{R}^n$ is convex if and only if the set

$$\{t : a+th\in X\}$$

is, for every (a,h), a convex set on the axis. Moreover, a function f on $\mathbb{R}^n$ is convex iff the function

$$\phi(t)=f(a+th) \tag{3.5.1}$$

is, for every (a,h), a convex function on the axis. Indeed, let us check the latter relation. The convexity of $\phi(t)$ follows from that of f by affine substitution. The convexity of f can be easily checked as follows. For any $x,y\in X$, let $h=y-x$. Then $\phi(0)=f(x+0(y-x))$, $\phi(1)=f(x+1(y-x))$. In addition for any $t\in[0,1]$, $\phi(t)=f(x+t(y-x))$. The convexity of ϕ implies that

$$f((1-t)x+ty)=f(x+t(y-x))=\phi(t)\le(1-t)\phi(0)+t\phi(1)=(1-t)f(x)+tf(y).$$

So now we consider when a function ϕ on the axis is convex. Let ϕ be convex and finite on (a,b). By (3.1.2), this is exactly the same as

$$\frac{\phi(z)-\phi(x)}{z-x}\le\frac{\phi(y)-\phi(x)}{y-x}\le\frac{\phi(y)-\phi(z)}{y-z}$$

when $a<x<z<y<b$. Assuming that $\phi'(x)$ and $\phi'(y)$ exist and passing to limits as $z\to x+0$ and $z\to y-0$, we get

$$\phi'(x)\le\frac{\phi(y)-\phi(x)}{y-x}\le\phi'(y)$$

that is, $\phi'(x)$ is nondecreasing on the set of points from (a,b) where it exists.

It turns out that the following conditions are necessary and sufficient for convexity of a univariate function.

a) The domain of the function ϕ should be an open interval $\Delta=(a,b)$, possibly with added endpoint(s) (provided that the corresponding endpoint(s) is/are finite).
b) ϕ should be continuous on (a,b) and differentiable everywhere, except, perhaps, a countable set, *and the derivative should be monotonically non-decreasing.*
c) at endpoint of (a,b) which belongs to $\mathrm{dom}\,\phi$, ϕ is allowed to “jump up”, but not to jump down.

In practice, it is more convenient to use the following sufficient condition for detecting the convexity of a univariate function ϕ: $\operatorname{dom}\phi$ is convex, ϕ is continuous on $\operatorname{dom}\phi$ and is twice differentiable, *with nonnegative* ϕ'', on $\operatorname{int}\operatorname{dom}\phi$. Indeed, we should prove that under the condition, if $x < z < y$ are in $\operatorname{dom}\phi$, then

$$\frac{\phi(z)-\phi(x)}{z-x} \leq \frac{\phi(y)-\phi(z)}{y-z}$$

By Lagrange Theorem, the left ratio is $\phi'(\xi)$ for certain $\xi \in (x,z)$, and the right ratio is $\phi'(\eta)$ for certain $\eta \in (z,y)$. Since $\phi''(\cdot) \geq 0$ and $\eta > \xi$, we have $\phi'(\eta) \geq \phi'(\xi)$. Similarly, a sufficient condition for convexity of a multivariate function f is given by: $\operatorname{dom} f$ is convex, f is continuous on $\operatorname{dom} f$ and is twice differentiable, *with positive semidefinite Hessian matrix* f'', on $\operatorname{int}\operatorname{dom} f$.

Example 3.1. Show that the function $f(x) = \ln(\sum_{i=1}^{n} \exp\{x_i\})$ is convex on $\mathbb{R}^n$.

Indeed,

$$\begin{aligned} h^T f'(x) &= \frac{\sum_i \exp\{x_i\}h_i}{\sum_i \exp\{x_i\}} \\ h^T f''(x)h &= -\frac{\left(\sum_i \exp\{x_i\}h_i\right)^2}{\left(\sum_i \exp\{x_i\}\right)^2} + \frac{\sum_i \exp\{x_i\}h_i^2}{\sum_i \exp\{x_i\}} \end{aligned}$$

Setting $p_i = \frac{\exp\{x_i\}}{\sum_j \exp\{x_j\}}$ and noting $\sum_i p_i = 1$, we have

$$\begin{aligned} h^T f''(x)h &= \sum_i p_i h_i^2 - \left(\sum_i p_i h_i\right)^2 \\ &= \sum_i p_i h_i^2 - \left(\sum_i \sqrt{p_i}(\sqrt{p_i}h_i)\right)^2 \\ &\geq \sum_i p_i h_i^2 - \left(\sum_i (\sqrt{p_i})^2\right)\left(\sum_i (\sqrt{p_i}h_i)^2\right) \\ &= \sum_i p_i h_i^2 - \left(\sum_i p_i h_i^2\right) = 0. \end{aligned}$$

Note that there exists a shortcut to prove the convexity of f in the example with no computations. It can be easily shown that $\ln(s) = \min_z[s\exp(z) - z - 1]$ for any $s > 0$. Now $\ln(\sum_i \exp(x_i)) = \min_z[\sum_i \exp(z)\exp(x_i) - z - 1]$ since the objective function in the latter relation is convex w.r.t. z and x_i's, it remains to use the rule on preserving convexity by partial minimization.

Below we prove an important inequality for convex functions.

Proposition 3.2. *Let f be a function, x be an interior point of the domain of f and Q, $x \in Q$, be a convex set such that f is convex on Q. Assume that f is differentiable at x. Then*

$$\forall y \in Q : f(y) \geq f(x) + (y-x)^T f'(x). \tag{3.5.2}$$

Proof. Let $y \in Q$. There is nothing to prove when $y = x$ or $f(y) = +\infty$, thus, assume that $f(y) < \infty$ and $y \neq x$. Let $z_\varepsilon = x + \varepsilon(y - x)$, $0 < \varepsilon < 1$. Then z_ε is an interior point of the segment $[x, y]$. Since f is convex, we have

$$\frac{f(y) - f(x)}{\|y - x\|} \geq \frac{f(z_\varepsilon) - f(x)}{\|z_\varepsilon - x\|} = \frac{f(x + \varepsilon(y - x)) - f(x)}{\varepsilon\|y - x\|}$$

Passing to limit as $\varepsilon \to +0$, we arrive at

$$\frac{f(y) - f(x)}{\|y - x\|} \geq \frac{(y - x)^T f'(x)}{\|y - x\|},$$

as required by (3.5.2). ■

3.6 Lipschitz continuity of convex functions*

Our goal in this section is to show that convex functions are Lipschitz continuous inside the interior of its domain.

We will first show that a convex function is locally bounded.

Lemma 3.1. *Let f be convex and $x_0 \in \operatorname{int}\operatorname{dom} f$. Then f is locally bounded, i.e., $\exists \varepsilon > 0$ and $M(x_0, \varepsilon) > 0$ such that*

$$f(x) \leq M(x_0, \varepsilon) \; \forall \, x \in B_\varepsilon(x_0) := \{x \in \mathbb{R}^n : \|x - x_0\|_2 \leq \varepsilon\}.$$

Proof. Since $x_0 \in \operatorname{int}\operatorname{dom} f$, $\exists \varepsilon > 0$ such that the vectors $x_0 \pm \varepsilon e_i \in \operatorname{int}\operatorname{dom} f$ for $i = 1, \ldots, n$, where e_i denotes the unit vector along coordinate i. Also let $H_\varepsilon(x_0) := \{x \in \mathbb{R}^n : \|x - x_0\|_\infty \leq \varepsilon\}$ denote the hypercube formed by the vectors $x_0 \pm \varepsilon e_i$. It can be easily seen that $B_\varepsilon(x_0) \subseteq H_\varepsilon(x_0)$ and hence that

$$\max_{x \in B_\varepsilon(x_0)} f(x) \leq \max_{x \in H_\varepsilon(x_0)} f(x) \leq \max_{i=1,\ldots,n} f(x_0 \pm \varepsilon e_i) =: M(x_0, \varepsilon).$$

■

Next we show that f is locally Lipschitz continuous.

Lemma 3.2. *Let f be convex and $x_0 \in \operatorname{int}\operatorname{dom} f$. Then f is locally Lipschitz, i.e., $\exists \varepsilon > 0$ and $\bar{M}(x_0, \varepsilon) > 0$ such that*

$$|f(y) - f(x_0)| \leq \bar{M}(x_0, \varepsilon)\|x - y\|, \; \forall y \in B_\varepsilon(x_0) := \{x \in \mathbb{R}^n : \|x - x_0\|_2 \leq \varepsilon\}. \quad (3.6.1)$$

Proof. We assume that $y \neq x_0$ (otherwise, the result is obvious). Let $\alpha = \|y - x_0\|_2/\varepsilon$. We extend the line segment connecting x_0 and y so that it intersects the ball $B_\varepsilon(x_0)$, and then obtain two intersection points z and u (see Figure 3.1). It can be easily seen that

$$y = (1-\alpha)x_0 + \alpha z, \tag{3.6.2}$$
$$x_0 = [y + \alpha u]/(1+\alpha). \tag{3.6.3}$$

It then follows from the convexity of f and (3.6.2) that

$$\begin{aligned} f(y) - f(x_0) &\le \alpha[f(z) - f(x_0)] = \tfrac{f(z)-f(x_0)}{\varepsilon}\|y - x_0\|_2 \\ &\le \tfrac{M(x_0,\varepsilon)-f(x_0)}{\varepsilon}\|y - x_0\|_2, \end{aligned}$$

where the last inequality follows from Lemma 3.1. Similarly, by the convexity f, (3.6.2) and Lemma 3.1, we have

$$f(x_0) - f(y) \le \|y - x_0\|_2 \tfrac{M(x_0,\varepsilon)-f(x_0)}{\varepsilon}.$$

Combining the previous two inequalities, we show (3.6.1) holds with $\bar{M}(x_0,\varepsilon) = [M(x_0,\varepsilon) - f(x_0)]/\varepsilon$. ■

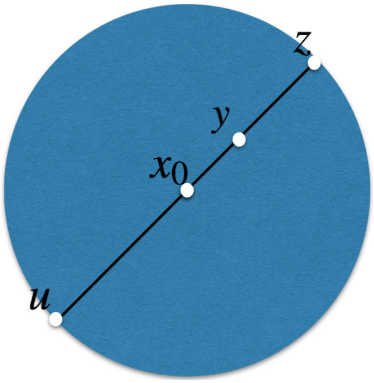


Fig. 3.1: Local Lipschitz continuity of a convex function

The following simple result shows the relation between the Lipschitz continuity of f and the boundedness of subgradients.

Lemma 3.3. *The following statements hold for a convex function f.*

a) If $x_0 \in \operatorname{int}\operatorname{dom} f$ and f is locally Lipschitz (i.e., (3.6.1) holds), then $\|g(x_0)\| \le \bar{M}(x_0,\varepsilon)$ for any $g(x_0) \in \partial f(x_0)$.

b) If $\exists g(x_0) \in \partial f(x_0)$ and $\|g(x_0)\|_2 \le \bar{M}(x_0,\varepsilon)$, then $f(x_0) - f(y) \le \bar{M}(x_0,\varepsilon)\|x_0 - y\|_2$.

Proof. We first show part a). Let $y = x_0 + \varepsilon g(x_0)/\|g(x_0)\|_2$. By the convexity of f and (3.6.1), we have

$$\varepsilon\|g(x_0)\|_2 = \langle g(x_0), y - x_0\rangle \le f(y) - f(x_0) \le \bar{M}(x_0,\varepsilon)\|y - x_0\| = \varepsilon\bar{M}(x_0,\varepsilon),$$

which implies part a). Part b) simply follows the convexity of f, i.e.,

$$f(x_0) - f(y) \leq \langle g(x_0), x_0 - y \rangle \leq \bar{M}(x_0, \varepsilon)\|x_0 - y\|_2.$$

■

Below we state the global Lipschitz continuity of a convex function in its interior of domain.

Theorem 3.1. *Let f be a convex function and let K be a closed and bounded set contained in the relative interior of the domain* $\operatorname{dom} f$ *of f. Then f is Lipschitz continuous on K, i.e., there exists constant M such that*

$$|f(x) - f(y)| \leq M_K \|x - y\|_2 \quad \forall x, y \in K. \tag{3.6.4}$$

Proof. The result directly follows from the local Lipschitz continuity of a convex function (see Lemmas 3.2 and 3.3) and the boundedness of K. ■

Remark 3.1. All three assumptions on K, i.e., (a) closedness, (b) boundedness, and (c) $K \subset \operatorname{ri} \operatorname{dom} f$ – are essential, as it is seen from the following three examples:

- $f(x) = 1/x$, $\operatorname{dom} f = (0, +\infty)$, $K = (0, 1]$. We have (b), (c) but not (a); f is neither bounded, nor Lipschitz continuous on K.
- $f(x) = x^2$, $\operatorname{dom} f = \mathbb{R}$, $K = \mathbb{R}$. We have (a), (c) and not (b); f is neither bounded nor Lipschitz continuous on K.
- $f(x) = -\sqrt{x}$, $\operatorname{dom} f = [0, +\infty)$, $K = [0, 1]$. We have (a), (b) and not (c); f is not Lipschitz continuous on K although is bounded. Indeed, we have $\lim_{t \to +0} \frac{f(0) - f(t)}{t} = \lim_{t \to +0} t^{-1/2} = +\infty$, while for a Lipschitz continuous f the ratios $t^{-1}(f(0) - f(t))$ should be bounded.

■

3.7 Minima and Maxima of Convex Functions

We first show the unimodality of a convex function.

Proposition 3.3. *Let f be a convex function and x_* be a local minimizer of f s.t. $x_* \in$* $\operatorname{dom} f$ *and*

$$\exists r > 0 : f(x) \geq f(x_*) \ \forall (x : \|x - x_*\| \leq r).$$

Then x_* is a global minimizer of f: $f(x) \geq f(x_*)$ $\forall x$.

Proof. All we need to prove is that if $x \neq x_*$ and $x \in \operatorname{dom} f$, then $f(x) \geq f(x_*)$. To this end let $z \in (x_*, x)$. By convexity we have

$$\frac{f(z) - f(x_*)}{\|z - x_*\|} \leq \frac{f(x) - f(x_*)}{\|x - x_*\|}.$$

When $z \in (x_*, x)$ is close enough to x_*, we have $\frac{f(z)-f(x_*)}{\|z-x_*\|} \geq 0$, whence $\frac{f(x)-f(x_*)}{\|x-x_*\|} \geq 0$, that is, $f(x) \geq f(x_*)$. ■

Now we show that the set of global optimizers of f is a convex set.

Proposition 3.4. *Let f be a convex function. The set of X_* of global minimizers is convex.*

Proof. This is an immediate corollary of important lemma. Lemma: Let f be a convex function. Then the level sets of f, that is, the sets

$$X_a = \{x : f(x) \leq a\}$$

where a is a real, are convex.

Proof of Lemma: If $x, y \in X_a$ and $\lambda \in [0,1]$, then

$$\begin{aligned} f(\lambda x + (1-\lambda)y) &\leq \lambda f(x) + (1-\lambda) f(y) \\ &\leq \lambda a + (1-\lambda)a = a. \end{aligned}$$

Thus, $[x,y] \subset X_a$. ■

We set out to understand when the minimizer of a convex function is unique. To this end, we say that a convex function *strictly convex*, if

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y)$$

whenever $x \neq y$ and $\lambda \in (0,1)$. Note that if a convex function f has *open* domain and is twice continuously differentiable on this domain with

$$h^T f''(x) h > 0 \quad \forall (x \in \operatorname{dom} f, h \neq 0),$$

then f is strictly convex.

Proposition 3.5. *For a strictly convex function f a minimizer, if it exists, is unique.*

Proof. Assume that $X_* = \text{Argmin} f$ contains two distinct points x', x''. By strict convexity,

$$f(\tfrac{1}{2}x' + \tfrac{1}{2}x'') < \tfrac{1}{2}\left[f(x') + f(x'')\right] = \inf_x f,$$

which is impossible. ■

We now state the optimality conditions in convex minimization.

Theorem 3.2. *Let f be a function which is differentiable at a point x_* and is convex on a convex set $Q \subset \operatorname{dom} f$ which contains x_*. A necessary and sufficient condition for f to attain its minimum on Q at x_* is*

$$(x - x_*)^T f'(x_*) \geq 0 \quad \forall x \in Q. \tag{3.7.1}$$

Proof. $\Leftarrow$: Assume that (3.7.1) is valid, and let us verify that $f(x) \geq f(x_*)$ for every $x \in Q$. There is nothing to prove when $x = x_*$, thus, let $f(x) < \infty$ and $x \neq x_*$. For $z_\lambda = x_* + \lambda(x - x_*)$ we have

$$\frac{f(z_\lambda) - f(x_*)}{\|z_\lambda - x_*\|} \leq \frac{f(x) - f(x_*)}{\|x - x_*\|} \quad \forall \lambda \in (0,1)$$

or, which is the same,

$$\frac{f(x_* + \lambda[x - x_*]) - f(x_*)}{\lambda\|x - x_*\|} \leq \frac{f(x) - f(x_*)}{\|x - x_*\|} \forall \lambda \in (0,1)$$

As $\lambda \to +0$, the left ratio converges to $(x - x_*)^T f'(x_*)/\|x - x_*\| \geq 0$; thus, $\frac{f(x)-f(x_*)}{\|x-x_*\|} \geq 0$, whence $f(x) \geq f(x_*)$.

$\Rightarrow$**:** Given that $x_* \in \mathrm{Argmin}_{y \in Q} f(y)$, let $x \in Q$. Then

$$0 \leq \frac{f(x_* + \lambda[x - x_*]) - f(x_*)}{\lambda} \quad \forall \lambda \in (0,1),$$

whence $(x - x_*)^T f'(x_*) \geq 0$.

■

We discuss an equivalent form of the above optimality condition. Let f be a function which is differentiable at a point x_* and is convex on a convex set $Q \subset \mathrm{dom}\, f$, $x_* \in Q$. Consider the *radial cone* of Q at x_*:

$$T_Q(x_*) = \{h : \exists t > 0 : x_* + th \in Q\}$$

Note that $T_Q(x_*)$ is indeed a cone which is comprised of all vectors of the form $s(x - x_*)$, where $x \in Q$ and $s \geq 0$. Then f attains its minimum on Q at x_* if and only if

$$h^T f'(x_*) \geq 0 \; \forall h \in T_Q(x_*),$$

or, which is the same, if and only if

$$f'(x_*) \in \underbrace{N_Q(x_*) = \{g : g^T h \geq 0 \forall h \in T_Q(x_*)\}}_{\text{normal cone of } Q \text{ at } x_*}. \tag{3.7.2}$$

Example 3.2. $x_* \in \mathrm{int}\, Q$. Here $T_Q(x_*) = \mathbb{R}^n$, whence $N_Q(x_*) = \{0\}$, and (3.7.2) becomes the Fermat equation $f'(x_*) = 0$.

Example 3.3. $x_* \in \mathrm{rint}\, Q$. Let $\mathrm{aff}(Q) = x_* + L$, where L is a linear subspace in $\mathbb{R}^n$. Here $T_Q(x_*) = L$, whence $N_Q(x_*) = L^\perp$. (3.7.2) becomes the condition

$$f'(x_*) \text{ is orthogonal to } L.$$

Equivalently, let $\mathrm{aff}(Q) = \{x : Ax = b\}$. Then $L = \{x : Ax = 0\}$, $L^\perp = \{y = A^T \lambda\}$, and the optimality condition becomes

$$\exists\lambda^* : \begin{array}{c} \nabla\big|_{x=x_*}[f(x)+(\lambda^*)^T(Ax-b)]=0 \\ \Updownarrow \\ f'(x_*)+\sum\limits_i \lambda_i^* \nabla(a_i^T x - b_i)=0 \end{array} \qquad [A = \begin{bmatrix} a_1^T \\ \vdots \\ a_m^T \end{bmatrix}].$$

Example 3.4. Let us solve the problem

$$\min_x \left\{ c^T x + \sum_{i=1}^m x_i \ln x_i : x \geq 0, \sum_i x_i = 1 \right\}.$$

The objective is convex, the domain $Q = \{x \geq 0, \sum\limits_i x_i = 1\}$ is convex (and even polyhedral). *Assuming* that the minimum is achieved at a point $x_* \in \operatorname{rint} Q$, the optimality condition becomes

$$\begin{array}{c} \nabla \left[c^T x + \sum\limits_i x_i \ln x_i + \lambda [\sum\limits_i x_i - 1] \right] = 0 \\ \Updownarrow \\ \ln x_i = -c_i - \lambda - 1 \ \forall i \\ \Updownarrow \\ x_i = \exp\{1-\lambda\}\exp\{-c_i\} \end{array}$$

Since $\sum\limits_i x_i$ should be 1, we arrive at

$$x_i = \frac{\exp\{-c_i\}}{\sum\limits_j \exp\{-c_j\}}.$$

At this point, the optimality condition is satisfied, so that the point indeed is a minimizer.

We end this section by a brief discussion about the maxima of convex functions. Let f be a convex function. Then

- If f attains its maximum over $\operatorname{dom} f$ at a point $x^* \in \operatorname{rint}\operatorname{dom} f$, then f is constant on $\operatorname{dom} f$.
- If $\operatorname{dom} f$ is closed and does not contain lines and f attains its maximum on $\operatorname{dom} f$, then among the maximizers there is an extreme point of $\operatorname{dom} f$.
- If $\operatorname{dom} f$ is polyhedral and f is bounded from above on $\operatorname{dom} f$, then f attains its maximum on $\operatorname{dom} f$.

3.8 Optimality Condition over Polyhedron

Below we discuss the optimality conditions of minimizing a convex function over a polyhedron. We need to use an important result shown in linear optimization, namely the homogeneous Farkas Lemma.

Consider a homogeneous linear inequality

$$a^T x \geq 0 \tag{3.8.1}$$

along with a finite system of similar inequalities:

$$a_i^T x \geq 0, \; 1 \leq i \leq m. \tag{3.8.2}$$

Our question is: when (3.8.1) is a consequence of (3.8.2), or equivalently, what kind of conditions will guarantee that every x satisfying (3.8.2) satisfies (3.8.1) as well? One immediate observation is that if a is a conic combination of $a_1,...,a_m$, i.e.,

$$\exists \lambda_i \geq 0 : a = \sum_i \lambda_i a_i, \tag{3.8.3}$$

then (3.8.1) is a consequence of (3.8.2). Indeed, (3.8.3) implies that

$$a^T x = \sum_i \lambda_i a_i^T x \; \forall x,$$

and thus for every x with $a_i^T x \geq 0 \forall i$ one has $a^T x \geq 0$.

Homogeneous Farkas Lemma below states a much stronger result than the above observation.

Theorem 3.3. *(3.8.1) is a consequence of (3.8.2) if and only if a is a conic combination of $a_1,...,a_m$.*

Proof. All we need to prove is that if a is not a conic combination of $a_1,...,a_m$, then there exists d such that $a^T d < 0$ and $a_i^T d \geq 0$, $i = 1,...,m$. Observe first that the set $K = \text{cone}\{a_1,...,a_m\}$ is polyhedrally representable:

$$\text{cone}\{a_1,...,a_m\} = \left\{ x : \exists \lambda \in \mathbb{R}^m : \begin{array}{l} x = \sum_i \lambda_i a_i \\ \lambda \geq 0 \end{array} \right\}.$$

Hence, by Fourier-Motzkin elimination, K is polyhedral:

$$K = \{x : d_\ell^T x \geq c_\ell, 1 \leq \ell \leq L\}.$$

Now notice that $0 \in K$ and hence that $c_\ell \leq 0 \forall \ell$. Moreover, using the fact $\lambda a_i \in \text{cone}\{a_1,...,a_m\} \forall \lambda > 0$, we have $\lambda d_\ell^T a_i \geq c_\ell \forall \lambda \geq 0$. Dividing both sides by λ and as λ tends to ∞, we have $d_\ell^T a_i \geq 0 \forall i, \ell$. Now, $a \notin \text{cone}\{a_1,...,a_m\}$ (i.e., $a \notin K$), hence there exists $\ell = \ell_*$ such that $d_{\ell_*}^T a < c_{\ell_*} \leq 0$. Therefore, $d = d_{\ell_*}$ satisfies $a^T d < 0$, $a_i^T d \geq 0, i = 1,...,m$. ■

Theorem 3.3 can be stated equivalently as follows. Given vectors $a_1,...,a_m \in \mathbb{R}^n$, let $K = \text{cone}\{a_1,...,a_m\} = \{\sum_i \lambda_i a_i : \lambda \geq 0\}$ be the conic hull of the vectors. Given a vector a,

a) it is easy to certify that $a \in \text{cone}\{a_1,...,a_m\}$: a certificate is a collection of weights $\lambda_i \geq 0$ such that $\sum_i \lambda_i a_i = a$;

b) it is easy to certify that $a \notin \mathrm{cone}\{a_1,...,a_m\}$: a certificate is a vector d such that $a_i^T d \geq 0 \,\forall i$ and $a^T d < 0$.

We now consider minimize a convex function f over the polyhedron $Q = \{x : Ax - b \leq 0\}$. In this case, the tangent cone

$$T_Q(x_*) = \left\{h : a_i^T h \leq 0 \;\forall i \in I(x_*) = \{i : a_i^T x_* - b_i = 0\}\right\}.$$

By Homogeneous Farkas Lemma,

$$\begin{aligned} N_Q(x_*) &\equiv \{y : a_i^T h \leq 0, i \in I(x_*) \Rightarrow y^T h \geq 0\} \\ &= \{y = -\sum_{i \in I(x_*)} \lambda_i a_i : \lambda_i \geq 0\} \end{aligned}$$

and the optimality condition becomes

$$\exists(\lambda_i^* \geq 0, i \in I(x_*)) : f'(x_*) + \sum_{i \in I(x_*)} \lambda_i^* a_i = 0$$

or, which is the same:

$$\exists \lambda^* \geq 0 : \begin{cases} f'(x_*) + \sum_{i=1}^m \lambda_i^* a_i = 0 \\ \quad \lambda_i^*(a_i^T x_* - b_i) = 0, i = 1,...,m \end{cases}$$

The point is that in the *convex* case these conditions are necessary *and sufficient* for x_* to be a minimizer of f on Q. The optimality condition can be extended to the case when Q contains linear equality constraints. We leave this development as an exercise.

3.9 Subgradients

Let f be a convex function and $\bar{x} \in \mathrm{int\,dom}\, f$. If f differentiable at $\bar{x}$, then, by gradient inequality, there exists an affine function, specifically,

$$h(x) = f(\bar{x}) + (x - \bar{x})^T f'(\bar{x}),$$

such that

$$f(x) \geq h(x) \forall x \;\&\; f(\bar{x}) = h(\bar{x}). \tag{3.9.1}$$

Affine function with this property may exist also in the case when f is *not* differentiable at $\bar{x} \in \mathrm{dom}\, f$. (3.9.1) implies that

$$h(x) = f(\bar{x}) + (x - \bar{x})^T g \tag{3.9.2}$$

for certain g. Function (3.9.2) indeed satisfies (3.9.1) if and only if g is such that

$$f(x) \geq f(\bar{x}) + (x-\bar{x})^T g \quad \forall x. \tag{3.9.3}$$

Let f be a convex function and $\bar{x} \in \operatorname{dom} f$. Every vector g satisfying (3.9.3) is called a *subgradient* of f at $\bar{x}$. The set of all subgradients, if any, of f at $\bar{x}$ is called *subdifferential* $\partial f(\bar{x})$ of f at $\bar{x}$.

Example 3.5. By Gradient Inequality, if convex function f is differentiable at $\bar{x}$, then $\nabla f(\bar{x}) \in \partial f(\bar{x})$. If, in addition, $\bar{x} \in \operatorname{int}\operatorname{dom} f$, then $\nabla f(\bar{x})$ is the *unique* element of $\partial f(\bar{x})$.

Example 3.6. Let $f(x) = |x|$ $(x \in \mathbb{R})$. When $\bar{x} \neq 0$, f is differentiable at $\bar{x}$, whence $\partial f(\bar{x}) = f'(\bar{x})$. When $\bar{x} = 0$, subgradients g are given by

$$|x| \geq 0 + gx = gx \,\forall x,$$

that is, $\partial f(0) = [-1,1]$. Note that in the case in question, f has directional derivative

$$Df(x)[h] = \lim_{t \to +0} \frac{f(x+th) - f(x)}{t}$$

at every point $x \in \mathbb{R}$ along every direction $h \in \mathbb{R}$, and this derivative is nothing but

$$Df(x)[h] = \max_{g \in \partial f(x)} g^T h.$$

The next result establishes the existence of subgradients for convex functions.

Proposition 3.6. *Let $X \subseteq \mathbb{R}^n$ be convex and $f: X \to \mathbb{R}$. If $\forall x \in X$, $\partial f(x) \neq \emptyset$ then f is convex. Moreover, if f is convex then for any $x \in \operatorname{ri}(X)$, $\partial f(x) \neq \emptyset$.*

Proof. The first claim is obvious. Let $g \in \partial f(\lambda x + (1-\lambda)y)$ for some $\lambda \in [0,1]$. Then by definition we have

$$f(y) \geq f(\lambda x + (1-\lambda)y) + \lambda \langle g, y-x \rangle,$$
$$f(x) \geq f(\lambda x + (1-\lambda)y) + (1-\lambda)\langle g, x-y \rangle.$$

Multiplying the first inequality by $1-\lambda$ and the second one by λ, and then summing them up, we show the convexity of f.

We now show that f has subgradients in the interior of X. We will construct such a subgradient by using a supporting hyperplane to the epigraph of f. Let $x \in X$. Then $(x, f(x)) \in \operatorname{epi}(f)$. By the convexity of $\operatorname{epi}(f)$ and the separating hyperplane theorem, there exists $(w,v) \in \mathbb{R}^n \times \mathbb{R}$ $((w,v) \neq 0)$ such that

$$\langle w, x \rangle + vf(x) \geq \langle w, y \rangle + vt, \ \ \forall (y,t) \in \operatorname{epi}(f). \tag{3.9.4}$$

Clearly, by tending t to infinity, we can see that $v \leq 0$. Now let us assume that x is in the interior of X. Then for $\varepsilon > 0$ small enough, $y = x + \varepsilon w \in X$, which implies that $v \neq 0$, since otherwise, we have $0 \geq \varepsilon \|w\|_2^2$ and hence $w = 0$, contradicting with the fact that $(w,v) \neq 0$. Letting $t = f(y)$ in (3.9.4), we obtain

$$f(y) \geq f(x) + \tfrac{1}{v}\langle w, x - y\rangle,$$

which implies that $-w/v$ is a subgradient of f at x. ■

Below we provide some basic subgradient calculus for convex functions. Observe that many of them mimic the calculus for gradient computation.

a) Scaling: $\partial(af) = a\partial f$ provided $a > 0$. The condition $a > 0$ makes function f remain convex.
b) Addition: $\partial(f_1 + f_2) = \partial(f_1) + \partial(f_2)$.
c) Affine composition: if $g(x) = f(Ax + b)$, then $\partial g(x) = A^T \partial f(Ax + b)$.
d) Finite pointwise maximum: if $f(x) = \max_{i=1,\dots,m} f_i(x)$, then

$$\partial f(x) = \operatorname{conv}\left\{\cup_{i:f_i(x)=f(x)} \partial f_i(x)\right\},$$

which is the convex hull of union of subdifferentials of all active $i : f_i(x) = f(x)$ functions at x.
e) General pointwise maximum: if $f(x) = \max_{s\in S} f_s(x)$, then under some regularity conditions (on S and f_s),

$$\partial f(x) = \operatorname{cl}\left\{\operatorname{conv}\left(\cup_{s:f_s(x)=f(x)} \partial f_s(x)\right)\right\}.$$

f) Norms: important special case, $f(x) = \|x\|_p$. Let q be such that $1/p + 1/q = 1$, then

$$\partial f(x) = \{y : \|y\|_q \leq 1 \,\text{and}\, y^T x = \max\{z^T x : \|z\|_q \leq 1\}.$$

We conclude this section by stating some important results of subgradients without providing the proof. Let f be convex. Then

a) For every $x \in \operatorname{dom} f$, the subdifferential $\partial f(x)$ is closed convex set.
b) If $x \in \operatorname{rint}\operatorname{dom} f$, then, for every $h \in \mathbb{R}^n$,

$$\exists Df(x)[h] \equiv \lim_{t\to+0} \frac{f(x+th) - f(x)}{t} = \max_{g\in\partial f(x)} g^T h.$$

c) Assume that $\bar{x} \in \operatorname{dom} f$ is represented as $\lim_{i\to\infty} x_i$ with $x_i \in \operatorname{dom} f$ and that

$$f(\bar{x}) \leq \lim \inf_{i\to\infty} f(x_i)$$

If a sequence $g_i \in \partial f(x_i)$ converges to certain vector g, then $g \in \partial f(\bar{x})$.
d) The multi-valued mapping $x \mapsto \partial f(x)$ is locally bounded at every point $\bar{x} \in \operatorname{int}\operatorname{dom} f$, that is, whenever $\bar{x} \in \operatorname{int}\operatorname{dom} f$, there exist $r > 0$ and $R < \infty$ such that

$$\|x - \bar{x}\|_2 \leq r, g \in \partial f(x) \Rightarrow \|g\|_2 \leq R.$$

3.10 Exercises

1. *Inverse of an increasing convex function*. Suppose $f : \mathbf{R} \to \mathbf{R}$ is increasing and convex on its domain (a,b). Let g denote its inverse, *i.e.*, the function with domain $(f(a), f(b))$ and $g(f(x)) = x$ for $a < x < b$. What can you say about convexity or concavity of g ?
2. [RV73, page15] Show that a continuous function $f : \mathbf{R}^n \to \mathbf{R}$ is convex if and only if for every line segment, its average value on the segment is less than or equal to the average of its values at the endpoints of the segment: For every $x, y \in \mathbf{R}^n$,
$$\int_0^1 f(x+\lambda(y-x))d\lambda \le \frac{f(x)+f(y)}{2}$$
3. *Running average of a convex function*. Suppose $f : \mathbf{R} \to \mathbf{R}$ is convex, with $\mathbf{R}_+ \subseteq \mathbf{dom} f$. Show that *its running average* F, defined as
$$F(x) = \frac{1}{x}\int_0^x f(t)dt, \quad \mathbf{dom} F = \mathbf{R}_{++},$$
is convex. You can assume f is differentiable.
4. Suppose $f : \mathbf{R}^n \to \mathbf{R}$ is convex with $\mathbf{dom} f = \mathbf{R}^n$, and bounded above on $\mathbf{R}^n$. Show that f is constant.
5. *Second-order conditions for convexity on an affine set*. Let $F \in \mathbf{R}^{n\times m}, \hat{x} \in \mathbf{R}^n$. The restriction of $f : \mathbf{R}^n \to \mathbf{R}$ to the affine set $\{Fz + \hat{x} \mid z \in \mathbf{R}^m\}$ is defined as the function $\tilde{f} : \mathbf{R}^m \to \mathbf{R}$ with
$$\tilde{f}(z) = f(Fz+\hat{x}), \quad \mathbf{dom}\tilde{f} = \{z \mid Fz+\hat{x} \in \mathbf{dom} f\}$$
Suppose f is twice differentiable with a convex domain.
   a. Show that $\tilde{f}$ is convex if and only if for all $z \in \mathbf{dom}\tilde{f}$
   $$F^T\nabla^2 f(Fz+\hat{x})F \succeq 0.$$
   b. Suppose $A \in \mathbf{R}^{p\times n}$ is a matrix whose nullspace is equal to the range of F, i.e., $AF = 0$ and $\mathbf{rank} A = n - \mathbf{rank} F$. Show that $\tilde{f}$ is convex if and only if for all $z \in \mathbf{dom}\tilde{f}$ there exists a $\lambda \in \mathbf{R}$ such that
   $$\nabla^2 f(Fz+\hat{x}) + \lambda A^T A \succeq 0.$$
   Hint. Use the following result: If $B \in \mathbf{S}^n$ and $A \in \mathbf{R}^{p\times n}$, then $x^T Bx \ge 0$ for all $x \in \mathcal{N}(A)$ if and only if there exists a λ such that $B + \lambda A^T A \succeq 0$.
6. *An extension of Jensen's inequality*. One interpretation of Jensen's inequality is that randomization or dithering hurts, i.e., raises the average value of a convex function: For f convex and v a zero mean random variable, we have $\mathbf{E} f(x_0+v) \ge f(x_0)$. This leads to the following conjecture. If f_0 is convex, then the larger the variance of v, the larger $\mathbf{E} f(x_0+v)$.

a. Give a counterexample that shows that this conjecture is false. Find zero mean random variables v and w, with $\mathbf{var}(v) > \mathbf{var}(w)$, a convex function f, and a point x_0, such that $\mathbf{E}f(x_0+v) < \mathbf{E}f(x_0+w)$.
b. The conjecture is true when v and w are scaled versions of each other. Show that $\mathbf{E}f(x_0+tv)$ is monotone increasing in $t \geq 0$, when f is convex and v is zero mean.

7. *Monotone mappings*. A function $\psi : \mathbf{R}^n \to \mathbf{R}^n$ is called monotone if for all $x, y \in \mathbf{dom}\psi$,

$$(\psi(x) - \psi(y))^T (x-y) \geq 0.$$

Suppose $f : \mathbf{R}^n \to \mathbf{R}$ is a differentiable convex function. Show that its gradient ∇f is monotone. Is the converse true, *i.e.*, is every monotone mapping the gradient of a convex function?

8. Suppose $f : \mathbf{R}^n \to \mathbf{R}$ is convex, $g : \mathbf{R}^n \to \mathbf{R}$ is concave, $\mathbf{dom}f = \mathbf{dom}g = \mathbf{R}^n$, and for all $x, g(x) \leq f(x)$. Show that there exists an affine function h such that for all x, $g(x) \leq h(x) \leq f(x)$. In other words, if a concave function g is an underestimator of a convex function f, then we can fit an affine function between f and g.

9. A family of concave utility functions. For $0 < \alpha \leq 1$ let

$$u_\alpha(x) = \frac{x^\alpha - 1}{\alpha},$$

with $\mathbf{dom}u_\alpha = \mathbf{R}_+$. We also define $u_0(x) = \log x$ (with $\mathbf{dom}u_0 = \mathbf{R}_{++}$).

a. Show that for $x > 0, u_0(x) = \lim_{\alpha \to 0} u_\alpha(x)$.
b. Show that u_α are concave, monotone increasing, and all satisfy $u_\alpha(1) = 0$.

These functions are often used in economics to model the benefit or utility of some quantity of goods or money. Concavity of u_α means that the marginal utility (*i.e.*, the increase in utility obtained for a fixed increase in the goods) decreases as the amount of goods increases. In other words, concavity models the effect of satiation.

10. *Nonnegative weighted sums and integrals*.

a. Show that $f(x) = \sum_{i=1}^r \alpha_i x_{[i]}$ is a convex function of x, where $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_r \geq 0$, and $x_{[i]}$ denotes the i th largest component of x. (You can use the fact that $f(x) = \sum_{i=1}^k x_{[i]}$ is convex on $\mathbf{R}^n$.)
b. Let $T(x,\omega)$ denote the trigonometric polynomial

$$T(x,\omega) = x_1 + x_2 \cos\omega + x_3 \cos 2\omega + \cdots + x_n \cos(n-1)\omega.$$

Show that the function

$$f(x) = -\int_0^{2\pi} \log T(x,\omega) d\omega$$

is convex on $\{x \in \mathbf{R}^n \mid T(x,\omega) > 0, 0 \leq \omega \leq 2\pi\}$.

11. *Some functions on the probability simplex.* Let x be a real-valued random variable which takes values in $\{a_1, \ldots, a_n\}$ where $a_1 < a_2 < \cdots < a_n$, with $\mathbf{prob}\,(x = a_i) = p_i, i = 1, \ldots, n$. For each of the following functions of p (on the probability simplex $\{p \in \mathbf{R}^n_+ \mid \mathbf{1}^T p = 1\}$), determine if the function is convex, concave, quasiconvex, or quasiconcave.

    a. $\mathbf{E}x$.
    b. $\mathbf{prob}(x \geq \alpha)$.
    c. $\mathbf{prob}(\alpha \leq x \leq \beta)$.
    d. $\sum_{i=1}^n p_i \log p_i$, the negative entropy of the distribution.
    e. $\mathbf{var}x = \mathbf{E}(x - \mathbf{E}x)^2$.
    f. $\mathbf{quartile}(x) = \inf\{\beta \mid \mathbf{prob}(x \leq \beta) \geq 0.25\}$.
    g. The cardinality of the smallest set $\mathscr{A} \subseteq \{a_1, \ldots, a_n\}$ with probability $\geq 90\%$. (By cardinality we mean the number of elements in $\mathscr{A}$.)
    h. The minimum width interval that contains 90% of the probability, *i.e.*,

    $$\inf\{\beta - \alpha \mid \mathbf{prob}(\alpha \leq x \leq \beta) \geq 0.9\}$$

12. *Convex hull or envelope of a function.* The *convex hull* or *convex envelope* of a function $f : \mathbf{R}^n \to \mathbf{R}$ is defined as

    $$g(x) = \inf\{t \mid (x,t) \in \mathbf{conv}\,\mathbf{epi}\,f\}.$$

    Geometrically, the epigraph of g is the convex hull of the epigraph of f.
    Show that g is the largest convex underestimator of f. In other words, show that if h is convex and satisfies $h(x) \leq f(x)$ for all x, then $h(x) \leq g(x)$ for all x.
13. *Products and ratios of convex functions.* In general the product or ratio of two convex functions is not convex. However, there are some results that apply to functions on $\mathbf{R}$. Prove the following.

    a. If f and g are convex, both nondecreasing (or nonincreasing), and positive functions on an interval, then fg is convex.
    b. If f, g are concave, positive, with one nondecreasing and the other nonincreasing, then fg is concave.
    c. If f is convex, nondecreasing, and positive, and g is concave, nonincreasing, and positive, then f/g is convex.

14. *Representation of piecewise-linear convex functions.* A convex function $f : \mathbf{R}^n \to \mathbf{R}$, with $\mathbf{dom}f = \mathbf{R}^n$, is called *piecewise-linear* if there exists a partition of $\mathbf{R}^n$ as

    $$\mathbf{R}^n = X_1 \cup X_2 \cup \cdots \cup X_L,$$

    where $\text{int}\, X_i \neq \emptyset$ and $\text{int}\, X_i \cap \mathbf{int} X_j = \emptyset$ for $i \neq j$, and a family of affine functions $a_1^T x + b_1, \ldots, a_L^T x + b_L$ such that $f(x) = a_i^T x + b_i$ for $x \in X_i$
    Show that such a function has the form $f(x) = \max\left\{a_1^T x + b_1, \ldots, a_L^T x + b_L\right\}$.

15. *More functions of eigenvalues*. Let $\lambda_1(X) \geq \lambda_2(X) \geq \cdots \geq \lambda_n(X)$ denote the eigenvalues of a matrix $X \in \mathbf{S}^n$. We have already seen several functions of the eigenvalues that are convex or concave functions of X.

    - The maximum eigenvalue $\lambda_1(X)$ is convex. The minimum eigenvalue $\lambda_n(X)$ is concave.
    - The sum of the eigenvalues (or trace), $\mathbf{tr} X = \lambda_1(X) + \cdots + \lambda_n(X)$, is linear.
    - The sum of the inverses of the eigenvalues (or trace of the inverse),$\mathbf{tr}\left(X^{-1}\right) = \sum_{i=1}^{n} 1/\lambda_i(X)$, is convex on $\mathbf{S}^n_{++}$ (exercise 3.18).
    - The geometric mean of the eigenvalues, $(\det X)^{1/n} = \left(\prod_{i=1}^{n} \lambda_i(X)\right)^{1/n}$, and the logarithm of the product of the eigenvalues, $\log\det X = \sum_{i=1}^{n} \log \lambda_i(X)$, are concave on $X \in \mathbf{S}^n_{++}$ (exercise 3.18 and page 74).

    In this problem we explore some more functions of eigenvalues, by exploiting variational characterizations.

    a. *Sum of k largest eigenvalues*. Show that $\sum_{i=1}^{k} \lambda_i(X)$ is convex on $\mathbf{S}^n$. *Hint*. [HJ85, page 191] Use the variational characterization

    $$\sum_{i=1}^{k} \lambda_i(X) = \sup\left\{\mathbf{tr}\left(V^T X V\right) \mid V \in \mathbf{R}^{n \times k}, V^T V = I\right\}$$

    b. *Geometric mean of k smallest eigenvalues*. Show that $\left(\prod_{i=n-k+1}^{n} \lambda_i(X)\right)^{1/k}$ is concave on $\mathbf{S}^n_{++}$. *Hint*. [MO79, page 513] For $X \succ 0$, we have

    $$\left(\prod_{i=n-k+1}^{n} \lambda_i(X)\right)^{1/k} = \frac{1}{k} \inf\left\{\mathbf{tr}\left(V^T X V\right) \mid V \in \mathbf{R}^{n \times k}, \det V^T V = 1\right\}$$

    c. *Log of product of k smallest eigenvalues*. Show that $\sum_{i=n-k+1}^{n} \log \lambda_i(X)$ is concave on $\mathbf{S}^n_{++}$. *Hint*. [MO79, page 513] For $X \succ 0$,

    $$\prod_{i=n-k+1}^{n} \lambda_i(X) = \inf\left\{\prod_{i=1}^{k} \left(V^T X V\right)_{ii} \mid V \in \mathbf{R}^{n \times k}, V^T V = I\right\}$$

16. **(Convexity under Composition)**
    Let C be a nonempty convex subset of $\Re^n$. Let also $f = (f_1, \ldots, f_m)$, where $f_i : C \mapsto \Re, i = 1, \ldots, m$, are convex functions, and let $g : \Re^m \mapsto \Re$ be a function that is convex and monotonically nondecreasing over a convex set that contains the set $\{f(x) \mid x \in C\}$, in the sense that for all $u, \bar{u}$ in this set such that $u \leq \bar{u}$, we have $g(u) \leq g(\bar{u})$. Show that the function h defined by $h(x) = g(f(x))$ is convex over C. If in addition, $m = 1, g$ is monotonically increasing and f is strictly convex, then h is strictly convex.
17. **(Posynomials)**
    A *posynomial* is a function of positive scalar variables $y_1, \ldots, y_n$ of the form

$$g(y_1,\dots,y_n)=\sum_{i=1}^{m}\beta_i y_1^{a_{i1}}\cdots y_n^{a_{in}}$$

where a_{ij} and β_i are scalars, such that $\beta_i > 0$ for all i. Show the following:

a. A posynomial need not be convex.

b. By a logarithmic change of variables, where we set

$$f(x)=\ln\left(g(y_1,\dots,y_n)\right), \quad b_i=\ln\beta_i, \forall i, \quad x_j=\ln y_j, \forall j,$$

we obtain a convex function

$$f(x)=\ln\exp(Ax+b), \quad \forall x\in\Re^n$$

where $\exp(z)=e^{z_1}+\cdots+e^{z_m}$ for all $z\in\Re^m$,A is an $m\times n$ matrix with components a_{ij}, and $b\in\Re^m$ is a vector with components b_i.

c. Every function $g:\Re^n\mapsto\Re$ of the form

$$g(y)=g_1(y)^{\gamma_1}\cdots g_r(y)^{\gamma_r}$$

where g_k is a posynomial and $\gamma_k>0$ for all k, can be transformed by a logarithmic change of variables into a convex function f given by

$$f(x)=\sum_{k=1}^{r}\gamma_k\ln\exp\left(A_kx+b_k\right)$$

with the matrix A_k and the vector b_k being associated with the posynomial g_k for each k.

18. **(Examples of Convex Functions)**
Show that the following functions from $\Re^n$ to $(-\infty,\infty]$ are convex:

a.
$$f_1(x_1,\dots,x_n)=\begin{cases}-\left(x_1x_2\cdots x_n\right)^{\frac{1}{n}} & \text{if } x_1>0,\cdots,x_n>0\\ \infty & \text{otherwise}\end{cases}$$

b. $f_2(x)=\ln\left(e^{x_1}+\cdots+e^{x_n}\right)$.

c. $f_3(x)=\|x\|^p$ with $p\geq 1$.

d. $f_4(x)=\frac{1}{f(x)}$, where f is concave and $0<f(x)<\infty$ for all x.

e. $f_5(x)=\alpha f(x)+\beta$, where $f:\Re^n\mapsto\Re$ is a convex function, and α and β are scalars, with $\alpha\geq 0$.

f. $f_6(x)=e^{\beta x'Ax}$, where A is a positive semidefinite symmetric $n\times n$ matrix and β is a positive scalar.

g. $f_7(x)=f(Ax+b)$, where $f:\Re^m\mapsto\Re$ is a convex function, A is an $m\times n$ matrix, and b is a vector in $\Re^m$.

19. **(Arithmetic-Geometric Mean Inequality)**

Show that if $\alpha_1,\dots,\alpha_n$ are positive scalars with $\sum_{i=1}^n \alpha_i = 1$, then for every set of positive scalars $x_1,\dots,x_n$, we have

$$x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n} \le \alpha_1 x_1 + a_2 x_2 + \cdots + \alpha_n x_n,$$

with equality if and only if $x_1 = x_2 = \cdots = x_n$. *Hint*: Show that $(-\ln x)$ is a strictly convex function on $(0,\infty)$.

20. **(Characterization of Differentiable Convex Functions)**
Let $f : \Re^n \mapsto \Re$ be a differentiable function. Show that f is convex over a nonempty convex set C if and only if

$$(\nabla f(x) - \nabla f(y))'(x-y) \ge 0, \quad \forall x,y \in C.$$

Note: The condition above says that the function f, restricted to the line segment connecting x and y, has monotonically nondecreasing gradient.

21. Let $f : \Re^n \mapsto (-\infty,\infty]$ be a convex function, let γ be a scalar, and let C be a nonempty convex subset of $\Re^n$.

a. Show that if $f(x) < \gamma$ for some x, then $f(x) < \gamma$ for some $x \in \mathrm{ri}(\mathrm{dom}(f))$.
b. Show that if $C \subset \mathrm{ri}(\mathrm{dom}(f))$ and $f(x) < \gamma$ for some $x \in \mathrm{cl}(C)$, then $f(x) < \gamma$ for some $x \in \mathrm{ri}(C)$.
c. Show that if $C \subset \mathrm{dom}(f)$ and $f(x) \ge \gamma$ for all $x \in C$, then $f(x) \ge \gamma$ for all $x \in \mathrm{cl}(C)$.

22. **(Strong Convexity)**
Let $f : \Re^n \mapsto \Re$ be a function that is continuous over a closed convex set $C \subset \mathrm{dom}(f)$, and let $\sigma > 0$. We say that *f is strongly convex over C with coefficient σ* if for all $x,y \in C$ and all $\alpha \in [0,1]$, we have

$$f(\alpha x + (1-\alpha)y) + \frac{\sigma}{2}\alpha(1-\alpha)\|x-y\|^2 \le \alpha f(x) + (1-\alpha) f(y).$$

a. Show that if f is strongly convex over C with coefficient σ, then f is strictly convex over C. Furthermore, there exists a unique $x^* \in C$ that minimizes f over C, and we have

$$f(x) \ge f(x^*) + \frac{\sigma}{2}\|x-x^*\|^2, \quad \forall x \in C.$$

b. Assume that int (C), the interior of C, is nonempty, and that f is continuously differentiable over int (C). Show that the following are equivalent:
(I) f is strongly convex with coefficient σ over C.
(II) We have

$$(\nabla f(x) - \nabla f(y))'(x-y) \ge \sigma\|x-y\|^2, \quad \forall x,y \in \mathrm{int}(C).$$

(III) We have

$$f(y) \geq f(x) + \nabla f(x)'(y-x) + \frac{\sigma}{2}\|x-y\|^2, \quad \forall x, y \in \mathrm{int}(C).$$

Furthermore, if f is twice continuously differentiable over int (C), the above three properties are equivalent to:

(IV) The matrix $\nabla^2 f(x) - \sigma I$ is positive semidefinite for every $x \in \mathrm{int}(C)$, where I is the identity matrix.

23. *Maximum of a convex function over a polyhedron*. Show that the maximum of a convex function f over the polyhedron $\mathscr{P} = \mathbf{conv}\{v_1, \ldots, v_k\}$ is achieved at one of its vertices, i.e.,

$$\sup_{x \in \mathscr{P}} f(x) = \max_{i=1,\ldots,k} f(v_i).$$

(A stronger statement is: the maximum of a convex function over a closed bounded convex set is achieved at an extreme point, *i.e.*, a point in the set that is not a convex combination of any other points in the set.) *Hint*. Assume the statement is false, and use Jensen's inequality.

24. *A quadratic-over-linear composition theorem*. Suppose that $f : \mathbf{R}^n \to \mathbf{R}$ is nonnegative and convex, and $g : \mathbf{R}^n \to \mathbf{R}$ is positive and concave. Show that the function f^2/g, with domain $\mathbf{dom} f \cap \mathbf{dom} g$, is convex.

25. Show that the function

$$f(x) = \frac{\|Ax-b\|_2^2}{1-x^T x}$$

is convex on $\{x \mid \|x\|_2 < 1\}$.

26. *Circularly symmetric convex functions*. Suppose $f : \mathbf{R}^n \to \mathbf{R}$ is convex and symmetric with respect to rotations, i.e., $f(x)$ depends only on $\|x\|_2$. Show that f must have the form $f(x) = \phi(\|x\|_2)$, where $\phi : \mathbf{R} \to \mathbf{R}$ is nondecreasing and convex, with dom $f = \mathbf{R}$. (Conversely, any function of this form is symmetric and convex, so this form characterizes such functions.)

27. *Infimal convolution*. Let $f_1, \ldots, f_m$ be convex functions on $\mathbf{R}^n$. Their *infimal convolution*, denoted $g = f_1 \diamond \cdots \diamond f_m$ (several other notations are also used), is defined as

$$g(x) = \inf\{f_1(x_1) + \cdots + f_m(x_m) \mid x_1 + \cdots + x_m = x\},$$

with the natural domain (*i.e.*, defined by $g(x) < \infty$). In one simple interpretation, $f_i(x_i)$ is the cost for the i th firm to produce a mix of products given by $x_i; g(x)$ is then the optimal cost obtained if the firms can freely exchange products to produce, all together, the mix given by x. (The name 'convolution' presumably comes from the observation that if we replace the sum above with the product, and the infimum above with integration, then we obtain the normal convolution.) Show that g is convex.

28. Suppose $\lambda_1, \ldots, \lambda_n$ are positive. Show that the function $f : \mathbf{R}^n \to \mathbf{R}$, given by

$$f(x) = \prod_{i=1}^{n} \left(1 - e^{-x_i}\right)^{\lambda_i}$$

is concave on

$$\mathbf{dom} f = \left\{ x \in \mathbf{R}^n_{++} \mid \sum_{i=1}^{n} \lambda_i e^{-x_i} \leq 1 \right\}$$

Hint. The Hessian is given by

$$\nabla^2 f(x) = f(x) \left(yy^T - \mathbf{diag}(z) \right)$$

where $y_i = \lambda_i e^{-x_i} / (1 - e^{-x_i})$ and $z_i = y_i / (1 - e^{-x_i})$

29. Show that the following functions $f : \mathbf{R}^n \to \mathbf{R}$ are convex.

   a. The difference between the maximum and minimum value of a polynomial on a given interval, as a function of its coefficients:

   $$f(x) = \sup_{t \in [a,b]} p(t) - \inf_{t \in [a,b]} p(t) \quad \text{where} \quad p(t) = x_1 + x_2 t + x_3 t^2 + \cdots + x_n t^{n-1}.$$

   a, b are real constants with $a < b$.

   b. The 'exponential barrier' of a set of inequalities:

   $$f(x) = \sum_{i=1}^{m} e^{-1/f_i(x)}, \quad \mathbf{dom} f = \{x \mid f_i(x) < 0, i = 1, \ldots, m\}.$$

   The functions f_i are convex.

   c. The function

   $$f(x) = \inf_{\alpha > 0} \frac{g(y + \alpha x) - g(y)}{\alpha}$$

   if g is convex and $y \in \mathbf{dom} g$. (It can be shown that this is the directional derivative of g at y in the direction x.)

30. Show that the following functions $f : \mathbf{R}^n \to \mathbf{R}$ are convex.

   a. $f(x) = -\exp(-g(x))$ where $g : \mathbf{R}^n \to \mathbf{R}$ has a convex domain and satisfies

   $$\begin{bmatrix} \nabla^2 g(x) & \nabla g(x) \\ \nabla g(x)^T & 1 \end{bmatrix} \succeq 0$$

   for $x \in \mathbf{dom} g$.

   b. The function

   $$f(x) = \max\{ \|APx - b\| \mid P \text{is a permutation matrix} \}$$

   with $A \in \mathbf{R}^{m \times n}, b \in \mathbf{R}^m$.

31. Show that a function $f : \mathbf{R} \to \mathbf{R}$ is convex if and only if $\mathbf{dom} f$ is convex and

$$\det \begin{bmatrix} 1 & 1 & 1 \\ x & y & z \\ f(x) & f(y) & f(z) \end{bmatrix} \geq 0$$

for all $x, y, z \in \mathbf{dom} f$ with $x < y < z$.

32. Mark by "c" those of the following functions which are convex on the indicated domains:

- $f(x) \equiv 1$ on $\mathbf{R}$
- $f(x) = x$ on $\mathbf{R}$
- $f(x) = |x|$ on $\mathbf{R}$
- $f(x) = -|x|$ on $\mathbf{R}$
- $f(x) = -|x|$ on $\mathbf{R}_+ = \{x \geq 0\}$
- $\exp\{x\}$ on $\mathbf{R}$
- $\exp\left\{x^2\right\}$ on $\mathbf{R}$
- $\exp\left\{-x^2\right\}$ on $\mathbf{R}$
- $\exp\left\{-x^2\right\}$ on $\{x \mid x \geq 100\}$

33. Prove that the following functions are convex on the indicated domains:

- $\frac{x^2}{y}$ on $\left\{(x,y) \in \mathbf{R}^2 \mid y > 0\right\}$
- $\ln(\exp\{x\} + \exp\{y\})$ on the $2D$ plane.

34. A function f defined on a convex set Q is called log-convex on Q, if it takes real positive values on Q and the function $\ln f$ is convex on Q. Prove that

- a log-convex on Q function is convex on Q
- the sum (more generally, linear combination with positive coefficients) of two log-convex functions on Q also is log-convex on the set.
  Hint: use the result of the previous Exercise + your knowledge on operations preserving convexity

35. For n-dimensional vector x, let $\widehat{x} = \left(\widehat{x}^1, \ldots, \widehat{x}^n\right)^T$ be the vector obtained from x by rearranging the coordinates in the non-ascending order. E.g., with $x = (2,1,3,1)^T, \widehat{x} = (3,2,1,1)^T$. Let us fix $k, 1 \leq k \leq n$.

- Is the function $\widehat{x}^k$ (k-th largest entry in x) a convex function of x ?
- Is the function $s_k(x) = \widehat{x}^1 + \ldots + \widehat{x}^k$ (the sum of k largest entries in x) convex?

36. Consider a Linear Programming program

$$\min_x \left\{c^T x : Ax \leq b\right\}$$

with $m \times n$ matrix A, and let x^* be an optimal solution to the problem. It means that x^* is a minimizer of differentiable convex function $f(x) = c^T x$ on convex set $Q = \{x \mid Ax \leq b\}$ and therefore, $\nabla f(x^*)$ should belong to the normal cone of A at x^* - this is the necessary and sufficient condition for optimality of x^*. What does this condition mean in terms of the data A, b, c ?

37. Let $f(x)$ be a convex symmetric function of $x \in \mathbf{R}^n$ (symmetry means that f remains unchanged when permuting the coordinates of the argument, as it is the case with $\sum_i x_i$, or $\max_i x_i$). Prove that if π is a double stochastic $n \times n$ matrix, then
$$f(\pi x) \leq f(x) \forall x$$

38. Let $f(x)$ be a convex symmetric function on $\mathbf{R}^n$. For a symmetric $n \times n$ matrix X, let $\lambda_1(X) \geq \lambda_2(X) \geq \ldots \geq \lambda_n(X)$ be the eigenvalues of X taken with their multiplicities and arranged in the nondecreasing order. Prove

    1. For every orthogonal $n \times n$ matrix U and symmetric $n \times n$ matrix X,
    $$f\left(\operatorname{diag}\left(UXU^T\right)\right) \leq f(\lambda(X))$$
    where $\operatorname{diag}(A)$ stands for the diagonal of square matrix;
    2. The function
    $$F(X) = f(\lambda(X))$$
    of symmetric $n \times n$ matrix X is convex.

39. For a 10×10 symmetric matrix X, what is larger - the sum of two largest diagonal entries or the sum of two largest eigenvalues?

Chapter 4
Duality and Optimality Conditions

4.1 Convex Programming

A mathematical optimization problem is

$$f_* = \min_x \left\{ f(x) : \begin{array}{r} g(x) \equiv (g_1(x),...,g_m(x))^T \le 0 \\ h(x) = (h_1(x),...,h_k(x))^T = 0 \\ x \in X \end{array} \right\} \qquad (P)$$

Here x is the *design vector*. Values of x are called *solutions* to (P), $f(x)$ is the *objective*, $g(x) \equiv (g_1(x),...,g_m(x))^T \le 0$ are *inequality constraints*, $h(x) = (h_1(x),...,h_k(x))^T = 0$ are *equality constraints*, $X \subset \mathbb{R}^n$ is the *domain*. We always assume that the objective and the constraints are well-defined on X.

A solution x is called *feasible*, if it satisfies all the constraints. Problem which has feasible solutions is called *feasible*. If the objective is (below) bounded on the set of feasible solutions, (P) is called *bounded*. The *optimal value* f_* is

$$f_* = \begin{cases} \inf_x \{f(x) : \ x \text{ is feasible}\}, & (P) \text{ is feasible} \\ +\infty, & \text{otherwise} \end{cases}$$

f_* is a real for feasible and bounded problem, is $-\infty$ for feasible unbounded problem, and is $+\infty$ for infeasible problem. *Optimal solution* of (P) is a feasible solution x_* such that $f(x_*) = f_*$. Problem which has optimal solutions is called *solvable*.

Problem (P) is called *convex*, if X is a convex subset of $\mathbb{R}^n$, $f(\cdot)$, $g_1(\cdot)$,...,$g_m(\cdot)$ are *convex real-valued* functions on X, and there are no equality constraints. Note that we could allow *linear* equality constraints, but this does not add generality.

4.2 Convex Theorem on Alternative

In this section, we generalize our discussions on how to certify the insolvability of linear systems in Theorem 3.3. Our present question is how to certify insolvability of the following nonlinear system:

$$\begin{aligned} f(x) &< c \\ g_j(x) &\le 0,\ j=1,\ldots,m \\ x &\in X. \end{aligned} \tag{4.2.1}$$

Assume that there exist nonnegative weights λ_j, $j=1,\ldots,m$, such that the inequality

$$f(x)+\sum_{j=1}^{m}\lambda_j g_j(x) < c$$

has no solutions in X, i.e.,

$$\lambda_j \ge 0: \quad \inf_{x\in X}[f(x)+\sum_{j=1}^{m}\lambda_j g_j(x)] \ge c. \tag{4.2.2}$$

Then, clearly (4.2.1) is insolvable.

Much stronger results can be obtained for convex system of inequalities. We say that system (4.2.1) is convex if

a) X is convex set.
b) $f, g_1,\ldots,g_m$ are real-valued convex functions on X.

Moreover, system (4.2.1) satisfies the so-called *Slater condition* if the subsystem

$$\begin{aligned} g_j(x) &< 0,\ j=1,\ldots,m, \\ x &\in X \end{aligned}$$

is solvable.

Theorem 4.1. *Consider a system of constraints in (4.2.1) along with system of constraints on λ in (4.2.2).*

a) If (4.2.2) is solvable, then (4.2.1) is insolvable.
b) If (4.2.1) is insolvable, convex, and satisfies the Slater condition, then (4.2.2) is solvable.

In addition, part b) still holds when the Slater condition is replaced with a relaxed Slater Condition: $\exists \bar{x} \in \mathrm{rint} X$ such that $g_i(\bar{x}) \le 0$ for all i and $g_i(\bar{x}) < 0$ for those i for which $g_i(\cdot)$ are not affine functions.

Proof. We only need to prove b). For simplicity, we prove b) under the slater condition and leave the one under the relaxed slater condition as an exercise. Assume that (I) has no solutions. Consider two sets in $\mathbb{R}^{m+1}$:

$$\overbrace{\left\{ u \in \mathbb{R}^{m+1} : \exists x \in X : \begin{array}{r} f(x) \leq u_0 \\ g_1(x) \leq u_1 \\ \cdots\cdots\cdots \\ g_m(x) \leq u_m \end{array} \right\}}^{T}$$

$$\underbrace{\left\{ u \in \mathbb{R}^{m+1} : u_0 < c, u_1 \leq 0, \ldots, u_m \leq 0 \right\}}_{S}$$

Observe that S and T are convex and nonempty, S and T do not intersect (otherwise (4.2.1) would have a solution). Therefore, S and T can be separated:

$$\exists (a_0, \ldots, a_m) \neq 0 : \inf_{u \in T} a^T u \geq \sup_{u \in S} a^T u.$$

Equivalently,

$$\begin{array}{l} \exists (a_0, \ldots, a_m) \neq 0 : \\ \quad \inf_{x \in X} \inf_{\substack{u_0 \geq f(x) \\ u_1 \geq g_1(x) \\ \vdots \\ u_m \geq g_m(x)}} [a_0 u_0 + a_1 u_1 + \ldots + a_m u_m] \\ \geq \sup_{\substack{u_0 < c \\ u_1 \leq 0 \\ \vdots \\ u_m \leq 0}} [a_0 u_0 + a_1 u_1 + \ldots + a_m u_m], \end{array}$$

which, in view of $a \geq 0$, implies that

$$\inf_{x \in X} [a_0 f(x) + a_1 g_1(x) + \ldots + a_m g_m(x)] \geq a_0 c.$$

Observe that we must have $a_0 > 0$. Otherwise $0 \neq (a_1, \ldots, a_m) \geq 0$ and

$$\inf_{x \in X} [a_1 g_1(x) + \ldots + a_m g_m(x)] \geq 0,$$

while by the slater condition, there exists $\bar{x} \in X : g_j(\bar{x}) < 0$ for all j. Therefore, we have

$$\inf_{x \in X} \left[f(x) + \sum_{j=1}^{m} \underbrace{\left[\frac{a_j}{a_0} \right]}_{\lambda_j \geq 0} g_j(x) \right] \geq c.$$

■

4.3 Lagrange Duality

Consider the optimization program

$$\mathrm{Opt}\,(P)=\min\left\{f(x):g_j(x)\le 0,\ j\le m,\ x\in X\right\}. \tag{4.3.1}$$

We associate (4.3.1) with the *Lagrange function*

$$L(x,\lambda)=f(x)+\sum_{j=1}^{m}\lambda_j g_j(x)$$

along with the *Lagrange Dual problem*

$$\mathrm{Opt}\,(D)=\max_{\lambda\ge 0}\underline{L}(\lambda),\ \underline{L}(\lambda)=\inf_{x\in X}L(x,\lambda). \tag{4.3.2}$$

We intend to show the following important duality theorem for convex programming.

Theorem 4.2. *a) [Weak Duality] For every $\lambda\ge 0$, $\underline{L}(\lambda)\le\mathrm{Opt}\,(P)$. In particular,*

$$\mathrm{Opt}\,(D)\le\mathrm{Opt}\,(P).$$

b) [Strong Duality] If (P) is convex and below bounded and satisfies Slater condition, then (D) is solvable, and

$$\mathrm{Opt}\,(D)=\mathrm{Opt}\,(P).$$

Proof. a) Weak Duality:"$\mathrm{Opt}\,(D)\le\mathrm{Opt}\,(P)$". There is nothing to prove when (P) is infeasible, that is, when $\mathrm{Opt}\,(P)=\infty$. If x is feasible for (P) and $\lambda\ge 0$, then $L(x,\lambda)\le f(x)$, whence

$$\begin{array}{rcl}\lambda\ge 0\Rightarrow\underline{L}(\lambda) & \equiv & \inf\limits_{x\in X}L(x,\lambda)\\ & \le & \inf\limits_{x\in X\text{ is feasible}}L(x,\lambda)\\ & \le & \inf\limits_{x\in X\text{ is feasible}}f(x)\\ & = & \mathrm{Opt}\,(P)\\ \Rightarrow\mathrm{Opt}\,(D) & = & \sup\limits_{\lambda\ge 0}\underline{L}(\lambda)\le\mathrm{Opt}\,(P).\end{array}$$

b) Strong Duality: "If (P) is convex and below bounded and satisfies Slater condition, then (D) is solvable and $\mathrm{Opt}\,(D)=\mathrm{Opt}\,(P)$". The system

$$f(x)<\mathrm{Opt}\,(P),\ g_j(x)\le 0,\ j=1,...,m,\ x\in X$$

has no solutions, while the system $g_j(x)<0,\ j=1,...,m,\ x\in X$ has a solution. By Theorem 4.1,

$$\exists\lambda^*\ge 0:f(x)+\sum_j\lambda_j^*g_j(x)\ge\mathrm{Opt}\,(P)\ \forall x\in X,$$

whence

$$\underline{L}(\lambda^*) \geq \mathrm{Opt}\,(P).$$

Combined with Weak Duality, we have

$$\mathrm{Opt}\,(D) = \underline{L}(\lambda^*) = \mathrm{Opt}\,(P).$$

■

Note that the Lagrange function "remembers", up to equivalence, both (P) and (D). Indeed,

$$\mathrm{Opt}\,(D) = \sup_{\lambda \geq 0} \inf_{x \in X} L(x, \lambda)$$

is given by the Lagrange function. Now consider the function

$$\overline{L}(x) = \sup_{\lambda \geq 0} L(x, \lambda) = \begin{cases} f(x), \; g_j(x) \leq 0, \; j \leq m \\ +\infty, \;\; \text{otherwise.} \end{cases}$$

(P) clearly is equivalent to the problem of minimizing $\overline{L}(x)$ over $x \in X$:

$$\mathrm{Opt}\,(P) = \inf_{x \in X} \sup_{\lambda \geq 0} L(x, \lambda). \tag{4.3.3}$$

4.4 Saddle Points

We now consider a more general form of minimax problem than (4.3.3). Let $X \subset \mathbb{R}^n$, $\Lambda \subset \mathbb{R}^m$ be nonempty sets, and let $F(x, \lambda)$ be a real-valued function on $X \times \Lambda$. This function gives rise to two optimization problems

$$\begin{array}{lll} \mathrm{Opt}\,(P') = \inf\limits_{x \in X} \overbrace{\sup\limits_{\lambda \in \Lambda} F(x, \lambda)}^{\overline{F}(x)} & & (P') \\ \mathrm{Opt}\,(D') = \sup\limits_{\lambda \in \Lambda} \underbrace{\inf\limits_{x \in X} F(x, \lambda)}_{\underline{F}(\lambda)} & & (D') \end{array} \tag{4.4.1}$$

The above problem has a game interpretation. Player I chooses $x \in X$, player II chooses $\lambda \in \Lambda$. With choices of the players x, λ, player I pays to player II the sum of $F(x, \lambda)$. So, what should the players do to optimize their wealth?

a) If Player I chooses x first, and Player II knows this choice when choosing λ, Player II will maximize her profit, and the loss of Player I will be $\overline{F}(x)$. To minimize his loss, Player I should solve (P'), thus ensuring himself loss $\mathrm{Opt}\,(P')$ or less.
b) If Player II chooses λ first, and Player I knows this choice when choosing x, Player I will minimize his loss, and the profit of Player II will be $\underline{F}(\lambda)$. To maximize her profit, Player II should solve (D'), thus ensuring herself profit $\mathrm{Opt}\,(D')$ or more.

Intuitively, the second situation seems better for Player I, so that it is natural to guess that his anticipated loss in this situation is $\leq$ his anticipated loss in the first situation:

$$\mathrm{Opt}\,(D') \equiv \sup_{\lambda\in\Lambda}\inf_{x\in X} F(x,\lambda) \leq \inf_{x\in X}\sup_{\lambda\in\Lambda} F(x,\lambda) \equiv \mathrm{Opt}\,(P').$$

This indeed is true: assuming $\mathrm{Opt}\,(P') < \infty$ (otherwise the inequality is evident),

$$\begin{array}{l}\forall(\varepsilon>0): \quad \exists x_\varepsilon \in X : \sup\limits_{\lambda\in\Lambda} F(x_\varepsilon,\lambda) \leq \mathrm{Opt}\,(P') + \varepsilon \\ \Rightarrow \forall\lambda\in\Lambda : \underline{F}(\lambda) = \inf\limits_{x\in X} F(x,\lambda) \leq F(x_\varepsilon,\lambda) \leq \mathrm{Opt}\,(P') + \varepsilon \\ \Rightarrow \mathrm{Opt}\,(D') \equiv \sup\limits_{\lambda\in\Lambda} \underline{F}(\lambda) \leq \mathrm{Opt}\,(P') + \varepsilon \\ \Rightarrow \mathrm{Opt}\,(D') \leq \mathrm{Opt}\,(P').\end{array}$$

So, what should the players do when making their choices simultaneously? A "good case" when we can answer this question happens when F has a *saddle point*. We call a point $(x_*,\lambda_*) \in X\times\Lambda$ a *saddle point* of F, if

$$F(x,\lambda_*) \geq F(x_*,\lambda_*) \geq F(x_*,\lambda)\ \forall(x\in X, \lambda\in\Lambda).$$

In game terms, a saddle point is an *equilibrium*, meaning no one of the players can improve his wealth, provided the adversary keeps his choice unchanged.

Proposition 4.1. *F has a saddle point if and only if both (P') and (D') are solvable with equal optimal values. In this case, the saddle points of F are exactly the pairs (x_*,λ_*), where x_* is an optimal solution to (P'), and λ_* is an optimal solution to (D').*

Proof. $\Rightarrow$ Assume that (x_*,λ_*) is a saddle point of F, and let us prove that x_* solves (P'), λ_* solves (D'), and $\mathrm{Opt}\,(P') = \mathrm{Opt}\,(D')$. Indeed, we have

$$F(x,\lambda_*) \geq F(x_*,\lambda_*) \geq F(x_*,\lambda)\ \forall(x\in X, \lambda\in\Lambda),$$

whence

$$\begin{array}{c}\mathrm{Opt}\,(P') \leq \overline{F}(x_*) = \sup\limits_{\lambda\in\Lambda} F(x_*,\lambda) = F(x_*,\lambda_*) \\ \mathrm{Opt}\,(D') \geq \underline{F}(\lambda_*) = \inf\limits_{x\in X} F(x,\lambda_*) = F(x_*,\lambda_*).\end{array}$$

Since $\mathrm{Opt}\,(P') \geq \mathrm{Opt}\,(D')$, we see that all inequalities in the chain

$$\mathrm{Opt}\,(P') \leq \overline{F}(x_*) = F(x_*,\lambda_*) = \underline{F}(\lambda_*) \leq \mathrm{Opt}\,(D')$$

are equalities. Thus, x_* solves (P'), λ_* solves (D') and $\mathrm{Opt}\,(P') = \mathrm{Opt}\,(D')$.

$\Leftarrow$ Assume that (P') and (D') have optimal solutions x_*,λ_* and $\mathrm{Opt}\,(P') = \mathrm{Opt}\,(D')$, and let us prove that (x_*,λ_*) is a saddle point. We have

$$\begin{aligned}\operatorname{Opt}(P') &= \overline{F}(x_*) = \sup_{\lambda\in\Lambda} F(x_*,\lambda) \geq F(x_*,\lambda_*)\\ \operatorname{Opt}(D') &= \underline{F}(\lambda_*) = \inf_{x\in X} F(x,\lambda_*) \leq F(x_*,\lambda_*).\end{aligned}$$

Since $\operatorname{Opt}(P') = \operatorname{Opt}(D')$, all inequalities in the above relation are equalities, so that

$$\sup_{\lambda\in\Lambda} F(x_*,\lambda) = F(x_*\lambda_*) = \inf_{x\in X} F(x,\lambda_*).$$

■

In view of Proposition 4.1, whenever F has a saddle point, the order of playing this game does not matter anymore since both (P') and (D') have equal optimal values.

4.5 Saddle Point Form of Optimality Conditions

We now turn out attention to the minimax problem defined in (4.3.3) and discuss the *saddle point form of optimality conditions in convex programming.*

$$\begin{gathered}\operatorname{Opt}(P) = \min_x \left\{ f(x) : g_j(x) \leq 0, j \leq m, x \in X \right\} \ (P)\\ \Downarrow\\ L(x,\lambda) = f(x) + \sum_{j=1}^{m} \lambda_j g_j(x)\end{gathered}$$

Theorem 4.3. *Let $x_* \in X$ be given.*

a) [Sufficient optimality condition] If x_* can be extended, by a $\lambda^* \geq 0$, to a saddle point of the Lagrange function on $X \times \{\lambda \geq 0\}$:

$$L(x,\lambda^*) \geq L(x_*,\lambda^*) \geq L(x_*,\lambda) \ \forall (x \in X, \lambda \geq 0),$$

then x_* is optimal for (P).

b) [Necessary optimality condition] If x_* is optimal for (P) and (P) is convex and satisfies the Slater condition, then x_* can be extended, by a $\lambda^* \geq 0$, to a saddle point of the Lagrange function on $X \times \{\lambda \geq 0\}$.

Proof. ⇒ Clearly, $\sup_{\lambda\geq 0} L(x_*,\lambda) = \begin{cases} +\infty, & x_* \text{ is infeasible}\\ f(x_*), & \text{otherwise}\end{cases}$

Thus, $\lambda^* \geq 0$ & $L(x_*,\lambda^*) \geq L(x_*,\lambda)\ \forall \lambda \geq 0$ is equivalent to

$$g_j(x_*) \leq 0 \forall j \ \&\ \lambda_j^* g_j(x_*) = 0 \forall j.$$

Consequently, $L(x_*,\lambda^*) = f(x_*)$, whence

$$L(x,\lambda^*) \geq L(x_*,\lambda^*) \ \forall x \in X$$

reads as

$$L(x,\lambda^*) \geq f(x_*)\ \forall x.$$

Since for $\lambda \geq 0$ one has $f(x) \geq L(x,\lambda)$ for all feasible x, it follows from the above inequality that

$$x \text{ is feasible } \Rightarrow f(x) \geq f(x_*).$$

$\Leftarrow$ By Lagrange Duality Theorem, $\exists \lambda^* \geq 0$:

$$f(x_*) = \underline{L}(\lambda^*) \equiv \inf_{x\in X}\left[f(x) + \sum_j \lambda_j^* g_j(x)\right]. \tag{4.5.1}$$

Since x_* is feasible, we have

$$\inf_{x\in X}\left[f(x) + \sum_j \lambda_j^* g_j(x)\right] \leq f(x_*) + \sum_j \lambda_j^* g_j(x_*) \leq f(x_*).$$

By (4.5.1), the last " $\leq$ " here is " $=$ ", which with $\lambda^* \geq 0$ is possible iff $\lambda_j^* g_j(x_*) = 0\, \forall j$ which together with $g_j(x_*) \leq 0$ imply that

$$f(x_*) = L(x_*,\lambda^*) \geq L(x_*,\lambda)\ \forall \lambda \geq 0.$$

Now (4.5.1) reads $L(x,\lambda^*) \geq f(x_*) = L(x_*,\lambda^*)$. The result then follows by combining these two inequalities.

■

4.6 Karush-Kuhn-Tucker Optimality Condition

Suppose that the functions $f, g_1,\ldots,g_m$ are differentiable at x^*. We call the following requirement

$$\begin{array}{l} \exists \lambda^* \geq 0 \text{ s.t.} \\ (a)\, \nabla f(x_*) + \sum\limits_{j=1}^{m} \lambda_j^* \nabla g_j(x_*) \in N_X^*(x_*) \\ (b)\, \lambda_j^* g_j(x_*) = 0,\ j \leq m\ [\text{complementary slackness}] \end{array}$$

as the KKT condition.

Theorem 4.4. *Let x^* be a feasible solution of (P), and let the functions f, $g_1,\ldots,g_m$ be differentiable at x^*. Then the KKT condition is sufficient for x_* to be optimal. Moreover, if (P) satisfies* restricted Slater condition*, then the KKT is necessary and sufficient for x_* to be optimal.*

Proof. $\Rightarrow$ Note that part b) (i.e., complementary slackness) in the KKT condition plus $\lambda^* \geq 0$ ensure that

$$L(x_*,\lambda^*) \geq L(x_*,\lambda) \quad \forall \lambda \geq 0.$$

Further, $L(x,\lambda^*)$ is convex in $x \in X$ and differentiable at $x_* \in X$, so that part a) in the KKT condition implies that

$$L(x,\lambda^*) \geq L(x_*,\lambda^*) \quad \forall x \in X.$$

Thus, x_* can be extended to a saddle point of the Lagrange function and therefore is optimal for (P).

$\Leftarrow$ By the saddle point optimality condition, from optimality of x_* it follows that $\exists \lambda^* \geq 0$ such that (x_*,λ^*) is a saddle point of $L(x,\lambda)$ on $X \times \{\lambda \geq 0\}$. This is equivalent to

$$\lambda_j^* g_j(x_*) = 0 \,\forall j \ \& \ \underbrace{\min_{x\in X} L(x,\lambda^*) = L(x_*,\lambda^*)} \tag{4.6.1}$$

Since the function $L(x,\lambda^*)$ is convex in $x \in X$ and differentiable at $x_* \in X$, part a) in the KKT condition immediately follows from the second relation in (4.6.1). ■

Example 4.1. Assuming $a_i > 0$, $p \geq 1$, let us solve the problem

$$\min_x \left\{ \sum_i \frac{a_i}{x_i} : x > 0, \sum_i x_i^p \leq 1 \right\}$$

Assuming $x_* > 0$ is a solution such that $\sum\limits_i (x_i^*)^p = 1$, the KKT conditions read

$$\begin{aligned} \nabla_x \left\{ \textstyle\sum\limits_i \frac{a_i}{x_i} + \lambda(\sum\limits_i x_i^p - 1) \right\} &= 0 \Leftrightarrow \tfrac{a_i}{x_i^2} = p\lambda x_i^{p-1} \\ \textstyle\sum\limits_i x_i^p &= 1 \end{aligned}$$

whence $x_i = c(\lambda) a_i^{\frac{1}{p+1}}$. Since $\sum\limits_i x_i^p$ should be 1, we get

$$x_i^* = a_i^{\frac{1}{p+1}} \Big/ \left(\sum_j a_j^{\frac{p}{p+1}} \right)^{\frac{1}{p}}.$$

This point is feasible, problem is convex, KKT at the point is satisfied, and thus must be optimal.

4.7 Convex-Cancave Saddle Points*

In this section, we consider the general saddle point problem in (4.4.1) and establish the conditions that guarantee the existence of saddle points.

Theorem 4.5. *[Sion-Kakutani] Let $X \subset \mathbb{R}^n$, $\Lambda \subset \mathbb{R}^m$ be nonempty convex closed sets and $F(x,\lambda) : X \times \Lambda \to \mathbb{R}$ be a continuous function which is convex in $x \in X$ and* *concave in $\lambda \in \Lambda$. Assume that X is compact, and that there exists $\bar{x} \in X$ such that all the sets*

$$\Lambda_a := \{\lambda \in \Lambda : F(\bar{x}, \lambda) \geq a\}$$

are bounded (e.g., Λ is bounded). Then F possesses a saddle point on $X \times \Lambda$.

The proof of the above Sion-Kakutani theorem requires us to show the following MiniMax lemma.

Lemma 4.1. *Let $f_i(x)$, $i = 1, \ldots, m$, be convex continuous functions on a convex compact set $X \subset \mathbb{R}^n$. Then there exists $\mu^* \geq 0$ with $\sum\limits_i \mu_i^* = 1$ such that*

$$\min_{x \in X} \max_{1 \leq i \leq m} f_i(x) = \min_{x \in X} \sum_i \mu_i^* f_i(x)$$

Proof. Consider the optimization program

$$\begin{gathered}\min_{t,x} \{t : f_i(x) - t \leq 0, \, i \leq m, (t,x) \in X_+\}, \\ X_+ = \{(t,x) : x \in X\}\end{gathered} \qquad (P)$$

The optimal value in this problem clearly is

$$t_* = \min_{x \in X} \max_i f_i(x).$$

The program clearly is convex, solvable and satisfies the Slater condition, whence there exists $\lambda^* \geq 0$ and an optimal solution (x_*, t_*) to (P) such that $(x_*, t_*; \lambda^*)$ is the saddle point of the Lagrange function on $X^+ \times \{\lambda \geq 0\}$:

$$\begin{aligned}\min_{x \in X, t} \left\{ t + \sum_i \lambda_i^* (f_i(x) - t) \right\} &= t_* + \sum_i \lambda_i^* (f_i(x_*) - t_*) \ (a) \\ \max_{\lambda \geq 0} \left\{ t_* + \sum_i \lambda_i (f_i(x_*) - t_*) \right\} &= t_* + \sum_i \lambda_i^* (f_i(x_*) - t_*) \ (b)\end{aligned}$$

(b) implies that $t_* + \sum\limits_i \lambda_i^* (f_i(x_*) - t_*) = t_*$ (o.w., the LHS is unbounded).

(a) implies that $\sum\limits_i \lambda_i^* = 1$ (o.w., the LHS is unbounded). Thus, $\lambda^* \geq 0, \sum_i \lambda_i^* = 1$ and

$$\begin{aligned}&\min_{x \in X} \sum_i \lambda_i^* f_i(x) = \min_{x \in X, t} \left\{ t + \sum_i \lambda_i^* (f_i(x) - t) \right\} \\ &= t_* + \sum_i \lambda_i^* (f_i(x_*) - t_*) = t_* = \min_{x \in X} \max_i f_i(x).\end{aligned}$$

■

We are now ready to prove Theorem 4.5.

Proof of Sion-Kakutani Theorem: We should prove that problems

$$\begin{array}{rcl}\mathrm{Opt}\,(P) &=& \inf\limits_{x\in X} \overbrace{\sup\limits_{\lambda\in\Lambda} F(x,\lambda)}^{\overline{F}(x)}\ (P)\\ \mathrm{Opt}\,(D) &=& \sup\limits_{\lambda\in\Lambda} \underbrace{\inf\limits_{x\in X} F(x,\lambda)}_{\underline{F}(\lambda)}\ (D)\end{array}$$

are solvable with equal optimal values.

1^0. Since X is compact and $F(x,\lambda)$ is continuous on $X\times\lambda$, the function $\underline{F}(\lambda)$ is continuous on Λ. Besides this, the sets

$$\Lambda^a = \{\lambda\in\Lambda : \underline{F}(\lambda)\geq a\}$$

are contained in the sets

$$\Lambda_a = \{\lambda\in\Lambda : F(\bar{x},\lambda)\geq a\}$$

and therefore are bounded. Finally, Λ is closed, so that the *continuous* function $\underline{F}(\cdot)$ with *bounded* level sets Λ^a attains it maximum on a *closed* set Λ. Thus, (D) is solvable; let λ^* be an optimal solution to (D).

2^0. Consider the sets

$$X(\lambda) = \{x\in X : F(x,\lambda)\leq \mathrm{Opt}\,(D)\}.$$

These are closed convex subsets of a compact set X. Let us prove that every finite collection of these sets has a nonempty intersection. Indeed, assume that

$$X(\lambda^1)\cap\ldots\cap X(\lambda^N) = \emptyset.$$

so that

$$\max_{j=1,\ldots,N} F(x,\lambda^j) > \mathrm{Opt}\,(D).$$

By MinMax Lemma, there exist weights $\mu_j\geq 0, \sum\limits_j \mu_j = 1$, such that

$$\min_{x\in X} \underbrace{\sum_j \mu_j F(x,\lambda^j)}_{\leq F(x,\underbrace{\sum_j \mu_j\lambda^j}_{\bar{\lambda}})} > \mathrm{Opt}\,(D)$$

which is impossible by the definition of $\mathrm{Opt}\,(D)$.

3^0. Since every finite collection of closed convex subsets $X(\lambda)$ of a compact set has a nonempty intersection, all those sets have a nonempty intersection:

$$\exists x_* \in X : F(x_*,\lambda)\leq \mathrm{Opt}\,(D)\ \forall\lambda.$$

Due to $\mathrm{Opt}(P) \geq \mathrm{Opt}(D)$, this is possible iff x_* is optimal for (P) and $\mathrm{Opt}(P) = \mathrm{Opt}(D)$.

4.8 First-order Optimality Conditions

We are given a Mathematical Programming problem

$$\min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \leq 0 \\ (h_1(x), ..., h_k(x)) = 0 \\ x \in X \end{array} \right\}.$$

Assume that we are given a feasible solution x_* to (P). What are the conditions (necessary, sufficient, necessary and sufficient) for x_* to be optimal?

Except for *convex programs*, there are no *verifiable local sufficient* conditions for *global* optimality. There exist, however,

a) verifiable local *necessary* conditions for *local* (and thus – for *global*) optimality.
b) verifiable local *sufficient* conditions for *local* optimality.

By definition, x^* being local optima means that there exists $r > 0$ such that for every feasible x with $\|x - x_*\| \leq r$ one has

$$f(x) \geq f(x_*).$$

Note that existing conditions for local optimality assume that $x_* \in \mathrm{int}X$, which, from the viewpoint of local optimality of x_*, is exactly the same as to say that $X = \mathbb{R}^n$. Therefore, from now on, we drop this domain constraint and consider

$$\min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \leq 0 \\ (h_1(x), ..., h_k(x)) = 0 \end{array} \right\}. \tag{4.8.1}$$

Moreover, we assume that the objective and all the constraints are continuously differentiable in a neighborhood of x_*. We will study *first-order optimality conditions* that are expressed via values and gradients of the objective and the constraints at x_*. Except for convex case, only *necessary* first-order conditions are known.

The basic idea for our development is to approximate (4.8.1) around x_* by a linear programming problem

$$\begin{array}{l} \min\limits_x f(x_*) + (x - x_*)^T f'(x_*) \\ \text{s.t.} \\ \overbrace{g_j(x_*)}^{0} + (x - x_*)^T g_j'(x_*) \leq 0,\ j \in J(x_*) \\ \underbrace{h_i(x_*)}_{0} + (x - x_*)^T h_i'(x_*) = 0,\ 1 \leq i \leq k \\ [J(x_*) = \{j : g_j(x_*) = 0\}] \end{array}$$

Since all $g_j(\cdot)$ are continuous at x_*, the *non-active at* x_* inequality constraints (those with $g_j(x_*) < 0$) do not affect local optimality of x_* and do not participate in the above problem. After removing some constant terms, we have

$$\min_x \left\{ (x-x_*)^T f'(x_*) : \begin{array}{l} (x-x_*)^T g_j'(x_*) \le 0, \\ j \in J(x_*) \\ (x-x_*)^T h_i'(x_*) = 0, \\ i = 1,...,k \end{array} \right\} \quad (LP)$$
$$J(x_*) = \{j : g_j(x_*) = 0\}$$

It is natural *to guess* that if x_* is locally optimal for (4.8.1), then x_* is locally optimal for (LP) as well. LP is a *convex* program with *affine* constraints, whence the KKT conditions are necessary and sufficient for optimality:

$$\begin{array}{l} \qquad\qquad x_* \text{ is optimal for } (LP) \\ \qquad\qquad\qquad \Updownarrow \\ \exists(\lambda_j^* \ge 0, j \in J(x_*), \mu_i): \\ f'(x_*) + \sum\limits_{j\in J(x_*)} \lambda_j^* g_j'(x_*) + \sum\limits_{i=1}^k \mu_i h_i'(x_*) = 0 \\ \qquad\qquad\qquad \Updownarrow \\ \exists(\lambda_j^* \ge 0, \mu_i^*): \\ \quad f'(x_*) + \sum\limits_j \lambda_j^* g_j'(x_*) + \sum\limits_i \mu_i^* h_i'(x_*) = 0 \\ \quad \lambda_j^* g_j(x_*) = 0,\ j = 1,...,m \end{array}$$

We then have the following intermediate result.

Proposition 4.2. *Let x_* be a locally optimal solution of (4.8.1). Assume that x_* remains locally optimal when passing from (4.8.1) to the linearized problem in (LP). Then at x_* the KKT condition holds:*

$$\begin{array}{l} \exists(\lambda_j^* \ge 0, \mu_i^*): \\ f'(x_*) + \sum\limits_j \lambda_j^* g_j'(x_*) + \sum\limits_i \mu_i^* h_i'(x_*) = 0 \\ \lambda_j^* g_j(x_*) = 0,\ j = 1,...,m. \end{array}$$

To make the above proposition useful, we need a verifiable sufficient condition for "x_* remains locally optimal when passing from (P) to (LP)". The most natural form of such a condition is *regularity*, meaning that the gradients, taken at x_*, of all constraints active at x_* are linearly independent. Of course, all equality constraints by definition are active at every feasible solution. The motivation of this regularity condition comes from the well-known implicit function theorem.

Recall a special form of Implicit Function Theorem as follows.

Theorem 4.6. *Let $x_* \in \mathbb{R}^n$ and let $p_\ell(x)$, $\ell = 1,...,L$, be real-valued functions such that*

a) p_ℓ are $\kappa \ge 1$ times continuously differentiable in a neighborhood of x_*;

b) $p_\ell(x_*) = 0$*;*
c) *vectors* $\nabla p_\ell(x_*)$*,* $\ell = 1,\dots,L$*, are linearly independent.*

Then there exists substitution of variables

$$y \mapsto x = \Phi(y)$$

defined in a neighborhood V *of the origin and mapping* V*, in a one-to-one manner, onto a neighborhood* B *of* x_**, such that*

a) $x_* = \Phi(0)$*;*
b) *both* $\Phi : V \to B$ *and its inverse mapping* $\Phi^{-1} : B \to V$ *are* κ *times continuously differentiable;*
c) *in coordinates* y*, the functions* p_ℓ *become just the coordinates:*

$$y \in V \Rightarrow p_\ell(\Phi(y)) \equiv y_\ell,\, \ell = 1,\dots,L.$$

We are now ready to state the first-order optimality condition.

Proposition 4.3. *Let* x_* *be a locally optimal* regular *solution of (4.8.1). Then* x_* *is optimal for* (LP) *and, consequently, the KKT conditions take place at* x_**.*

Proof. Let x_* be a regular locally optimal solution to (P). Assume, on the contrary to what should be proven, that x_* is not an optimal solution to (LP), and let us lead this to contradiction.

1^0. Since $x = x_*$ is not an optimal solution to (LP), there exists a feasible solution

$$x' = x_* + d$$

to the problem with

$$(x' - x_*)^T f'(x_*) = d^T f'(x_*) < 0,$$

so that

$$d^T f'(x_*) < 0, \underbrace{d^T h_i'(x_*) = 0}_{\forall i}, \underbrace{d^T g_j'(x_*) \le 0}_{\forall j \in J(x_*)}$$

2^0. W.l.o.g., assume that $J(x_*) = \{1,\dots,\ell\}$. By Theorem 4.6, there exist continuously differentiable local substitution of argument

$$x = \Phi(y) \qquad [\Phi(0) = x_*]$$

with a continuously differentiable in a neighborhood of x_* inverse $y = \Psi(x)$ such that in a neighborhood of origin one has

$$h_i(\Phi(y)) \equiv y_i,\ g_j(\Phi(y)) = y_{k+j},\ j = 1,\dots,\ell.$$

Since $\Psi(\Phi(y)) \equiv y$, we have

$$\Psi'(x_*)\Phi'(0) = I,$$

whence $\Phi'(0)$ is invertible and

$$\exists e : \Phi'(0)e = d.$$

3^0. Now we have found a smooth local substitution of argument $x = \Phi(y)$ ($y = 0$ corresponds to $x = x_*$) and a direction e such that in a neighborhood of $y = 0$ one has

$$\begin{array}{ll}
(a) & \quad h_i(\Phi(y)) \equiv y_i,\ i \le k \\
(b) & \quad g_j(\Phi(y)) \equiv y_{k+j},\ j \le \ell \\
 & \qquad\qquad\quad [J(x_*) = \{1,\dots,\ell\}] \\
(c) & [\Phi'(0)e]^T h_i'(x_*) = 0,\ i \le k \\
(d) & [\Phi'(0)e]^T g_j'(x_*) \le 0,\ j \le \ell \\
(e) & [\Phi'(0)e]^T f'(x_*) < 0
\end{array}$$

Consider the differentiable curve

$$x(t) = \Phi(te).$$

We have, by taking derivatives w.r.t. t,

$$\begin{array}{l}
te_i \equiv h_i(\Phi(te)) \Rightarrow e_i = [\Phi'(0)e]^T h_i'(x_*) = 0 \\
te_{k+j} \equiv g_j(\Phi(te)) \Rightarrow e_{k+j} = [\Phi'(0)e]^T g_j'(x_*) \le 0 \\
\Rightarrow \underbrace{h_i(x(t)) = te_i = 0}_{\forall i},\ \underbrace{g_j(x(t)) = te_{k+j} \le 0}_{\forall j \in J(x_*)}
\end{array}$$

Thus, $x(t)$ is feasible for all small $t \ge 0$. But:

$$\frac{d}{dt}\Big|_{t=0} f(x(t)) = [\Phi'(0)e]^T f'(x_*) < 0,$$

whence $f(x(t)) < f(x(0)) = f(x_*)$ for all small enough $t > 0$, which is a contradiction with local optimality of x_*. ■

Observe that the regularity of x_* is important for the KKT condition to be necessary for local optimality. For example, $x_* = 0$ is the only feasible solution to the problem $\min\{f(x) := x : h(x) := x^2 = 0\}$ and therefore is even globally optimal. The KKT condition would say that there exists μ^* such that $0 = \nabla f(x_*) + \mu^* \nabla h(x_*) = 1 + \mu^* \cdot 0$, which is impossible. The source of the difficulty is that $\nabla h(x_*) = 0$, that is, $x_* = 0$ is not a regular locally optimal solution.

4.9 Second-order Condition for Unconstrained Problems

In the case of unconstrained minimization problem

$$\min_x f(x) \tag{4.9.1}$$

with continuously differentiable objective, the KKT conditions reduce to the Fermat Rule: *If* x_* is locally optimal for (4.9.1), *then* $\nabla f(x_*) = 0$.

Fermat Rule is the "first order" part of *Second Order Necessary Optimality Condition* in unconstrained minimization: *If* x_* is locally optimal for (4.9.1) and f is twice differentiable in a neighborhood of x_*, *then*

$$\nabla f(x_*) = 0 \;\&\; \nabla^2 f(x_*) \succeq 0 \Leftrightarrow d^T \nabla^2 f(x_*) d \geq 0 \forall d$$

Indeed, let x_* be locally optimal for (4.9.1); then for appropriate $r_d > 0$

$$\begin{array}{l} 0 \leq t \leq r_d \\ \Rightarrow 0 \leq f(x_* + td) - f(x_*) \\ \quad = t \underbrace{d^T \nabla f(x_*)}_{=0} + \frac{1}{2} t^2 d^T \nabla^2 f(x_*) d + t^2 \underbrace{R_d(t)}_{\substack{\to 0, \\ t \to 0}} \\ \Rightarrow \frac{1}{2} d^T \nabla^2 f(x_*) d + R_d(t) \geq 0 \Rightarrow d^T \nabla^2 f(x_*) d \geq 0 \end{array}$$

The second-order *necessary* optimality condition can be strengthened to a second-order sufficient optimality condition for unconstrained minimization as shown below.

Theorem 4.7. *Let f be twice differentiable in a neighborhood of x_*. If*

$$\nabla f(x_*) = 0, \nabla^2 f(x_*) \succ 0 \Leftrightarrow d^T \nabla^2 f(x_*) d > 0 \forall d \neq 0$$

then x_* is locally optimal for (4.9.1).

Proof. Since $d^T \nabla^2 f(x_*) d > 0$ for all $d > 0$, then there exists $\alpha > 0$ such that $d^T \nabla^2 f(x_*) d \geq \alpha d^T d$ for all d. By differentiability, for every $\varepsilon > 0$ there exists $r_\varepsilon > 0$ such that

$$\begin{array}{l} \|d\|_2 \leq r_\varepsilon \\ \Rightarrow f(x_* + d) - f(x_*) \geq \underbrace{d^T \nabla f(x_*)}_{=0} + \frac{1}{2} \underbrace{d^T \nabla^2 f(x_*) d}_{\geq \alpha d^T d} - \frac{\varepsilon}{2} d^T d \\ \\ \Rightarrow f(x_* + d) - f(x_*) \geq \frac{1}{2} (\alpha - \varepsilon) d^T d \end{array}$$

Setting $\varepsilon = \frac{\alpha}{2}$, we see that x_* is a local minimizer of f. ■

4.10 Second-order Necessary Condition for Constrained Problems

We now consider a constrained mathematical programming problem

$$\min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), \ldots, g_m(x)) \leq 0 \\ (h_1(x), \ldots, h_k(x)) = 0 \end{array} \right\} \tag{4.10.1}$$

In the optimality conditions for a constrained problem, the role of $\nabla^2 f(x_*)$ is played by the Hessian of the *Lagrange function:*

$$L(x;\lambda,\mu) = f(x) + \sum_j \lambda_j g_j(x) + \sum_i \mu_i h_i(x).$$

We now establish the second-order necessary optimality condition for problem (4.10.1).

Theorem 4.8. *Let x_* be a* regular *feasible solution of (4.10.1) such that the functions f, g_j, h_i are twice continuously differentiable in a neighborhood of x_*. If x_* is locally optimal,* then

a) There exist uniquely defined Lagrange multipliers $\lambda_j^* \geq 0$, μ_i^* such that the KKT conditions hold:

$$\begin{array}{l} \nabla_x L(x_*;\lambda^*,\mu^*) = 0 \\ \lambda_j^* g_j(x_*) = 0,\ j = 1,...,m. \end{array}$$

b) For every d orthogonal to the gradients, taken at x_*, of all equality constraints and all active *at x_* inequality constraints, one has*

$$d^T \nabla_x^2 L(x_*;\lambda^*,\mu^*) d \geq 0.$$

Proof. **1^0.** Constraints which are non-active at x_* clearly do not affect neither local optimality of x_*, nor the conclusion to be proven. Removing these constraints, we reduce the situation to one where all constraints in the problem (4.10.1) are active at x_*.

2^0. Applying Implicit Function Theorem, we can find a local change of variables

$$x = \Phi(y) \Leftrightarrow y = \Psi(x) \qquad\qquad [\Phi(0) = x_*, \Psi(x_*) = 0]$$

with locally twice continuously differentiable Φ and Ψ such that

$$g_j(\Phi(y)) \equiv y_j,\ j \leq m, h_i(\Phi(y)) \equiv y_{m+i},\ i \leq k.$$

In variables y, problem (4.10.1) becomes

$$\min_y \left\{ \underbrace{f(\Phi(y))}_{\phi(y)} : y_j \leq 0,\ j \leq m, y_{m+i} = 0,\ i \leq k \right\}. \tag{4.10.2}$$

Our plan is as follows. Since Φ is a smooth one-to-one mapping of a neighborhood of x_* onto a neighborhood of $y_* = 0$, x_* is locally optimal for (4.10.1) iff $y_* = 0$ is locally optimal for (4.10.2). So we intend to build necessary/sufficient conditions for $y_* = 0$ to be locally optimal for (4.10.2). Then “translated” to x-variables, these conditions will imply necessary/sufficient conditions for local optimality of x_* for (4.10.1).

3^0. Since $x_* = \Phi(0)$ is locally optimal for (4.10.1), $y_* = 0$ is locally optimal for (4.10.2). In particular, if e_i is i-th basic orth, then for appropriate $\varepsilon > 0$:

$$\begin{array}{c} j \leq m \Rightarrow y(t) = -te_j \text{ is feasible for (4.10.2) when} \\ \varepsilon \geq t \geq 0 \Rightarrow -\frac{\partial \phi(0)}{\partial y_j} = \frac{d}{dt}\Big|_{t=0} \phi(y(t)) \geq 0 \\ \Rightarrow \lambda_j^* \equiv -\frac{\partial \phi(0)}{\partial y_j} \geq 0 \end{array}$$

and

$$\begin{array}{c} s > m+k \Rightarrow y(t) = te_s \text{ is feasible for (4.10.2) when} \\ \varepsilon \geq t \geq -\varepsilon \Rightarrow \frac{\partial \phi(0)}{\partial y_s} = \frac{d}{dt}\Big|_{t=0} \phi(y(t)) = 0 \\ \Rightarrow \mu_i^* \equiv -\frac{\partial \phi(0)}{\partial y_{m+i}} = 0, i = 1, \ldots, k. \end{array}$$

Therefore, $\exists \lambda^* \geq 0, \mu^*$:

$$0 = \frac{\partial M(0;\lambda^*,\mu^*)}{\partial y_\ell} \equiv \begin{cases} \frac{\partial \phi(0)}{\partial y_\ell} + \lambda_\ell^*, & \ell \leq m, (\text{active inequality constraints}) \\ \frac{\partial \phi(0)}{\partial y_\ell} + \mu_{\ell-m}^*, & m < \ell \leq m+k, (\text{equality constraints}) \\ \frac{\partial \phi(0)}{\partial y_\ell}, & \ell > m+k (\text{inactive inequality constraints}) \end{cases} \tag{KKT}$$

Note that the condition $\nabla_y M(0;\lambda^*,\mu^*) = 0$ defines λ^*, μ^* in a unique fashion.

4^0. We have seen that for (4.10.2), the first order part of the second-order necessary optimality condition holds true. Let us prove the second order part of the condition, which reads

$$\begin{array}{c} \forall (d : d^T \nabla_y y_\ell = 0, \ell \leq m+k): \\ d^T \nabla_y^2 M(0;\lambda^*,\mu^*) d \geq 0. \end{array} \tag{4.10.3}$$

This is evident since $M(y;\lambda^*,\mu^*) = \phi(y) + \sum\limits_{j=1}^{m} \lambda_j^* y_j + \sum\limits_{i=1}^{k} \mu_i^* y_{m+i}$, we have

$$\nabla_y^2 M(0;\lambda^*,\mu^*) = \nabla^2 \phi(0).$$

The claim in (4.10.3) therefore states that $d^T \nabla^2 \phi(0) d \geq 0$ for every vector d from the linear subspace $L = \{d : d_1 = \ldots = d_{m+k} = 0\}$. But this subspace is feasible for (4.10.2), so that ϕ, restricted onto L, should attain unconstrained local minimum at the origin. By the second-order necessary optimality condition for unconstrained minimization,

$$d^T \nabla^2 \phi(0) d \geq 0 \, \forall d \in L.$$

5^0. We have seen that if x_* is locally optimal for (4.10.1), then there exist uniquely defined $\lambda^* \geq 0$, μ^* such that

$$\nabla_y M(0;\lambda^*,\mu^*) = 0,$$

and one has

$$d^T\nabla_y y_\ell = 0,\, \ell \le m+k \Rightarrow d^T\nabla_y^2 M(0;\lambda^*,\mu^*)d \ge 0.$$

Let us prove that

$$\nabla_x L(x_*;\lambda^*,\mu^*) = 0 \tag{4.10.4}$$

and

$$\left.\begin{array}{l} e^Tg_j'(x_*) = 0,\, j\le m \\ e^Th_i'(x_*) = 0,\, i \le k \end{array}\right\} \Rightarrow e^T\nabla_x^2 L(x_*);\lambda^*,\mu^*)e \ge 0. \tag{4.10.5}$$

First, setting $\mathscr{L}(x) = L(x;\lambda^*,\mu^*)$, $\mathscr{M}(y) = M(y;\lambda^*,\mu^*)$, we have

$$\begin{array}{c} \mathscr{L}(x) = \mathscr{M}(\Psi(x)) \\ \Rightarrow\, \nabla_x\mathscr{L}(x_*) = [\Psi'(x_*)]^T\nabla_y\mathscr{M}(y_*) = 0, \end{array}$$

as required in (4.10.4).

Second, let e satisfy the premise in (4.10.5), and let $d = [\Phi'(0)]^{-1}e$. Then

$$\begin{array}{l}
\overbrace{\left.\frac{d}{dt}\right|_{t=0} g_j(\Phi(td))}^{\frac{d}{dt}\big|_{t=0} td_j} = [g_j'(x_*)]^T\,\overbrace{[\Phi'(0)]d}^{e} \\
\Rightarrow d_j = e^Tg_j'(x_*) = 0,\, j\le m \\
\underbrace{\left.\frac{d}{dt}\right|_{t=0} h_i(\Phi(td))}_{\frac{d}{dt}\big|_{t=0} td_{m+i}} = [h_i'(x_*)]^T\,\underbrace{[\Phi'(0)]d}_{e} \\
\Rightarrow d_{m+i} = e^Th_i'(x_*) = 0,\, i \le k.
\end{array}$$

We have

$$\begin{array}{l}
e^T\nabla^2\mathscr{L}(x_*)e = \left.\frac{d^2}{dt^2}\right|_{t=0}\mathscr{L}(x_*+te) = \left.\frac{d^2}{dt^2}\right|_{t=0}\mathscr{M}(\Psi(x_*+te)) \\
= \left.\frac{d}{dt}\right|_{t=0}\left[e^T[\Psi'(x_*+te)]^T\nabla\mathscr{M}(\Psi(x_*+te))\right] \\
= e^T[\Psi'(x_*)]^T\nabla^2\mathscr{M}(0)\;\overbrace{[\Psi'(x_*)e]}^{=[\Phi'(0)]^{-1}e=d} \\
\quad +e^T[\left.\frac{d}{dt}\right|_{t=0}\Psi'(x_*+te)]^T\,\underbrace{\nabla\mathscr{M}(0)}_{=0} \\
= d^T\nabla^2\mathscr{M}d \ge 0 \text{ due to } d_j = 0,\, 1\le j\le m+k.
\end{array}$$

Thus, whenever e is orthogonal to the gradients of all constraints active at x_*, we have $e^T\nabla^2\mathscr{L}e \ge 0$. ■

4.11 Second-order Sufficient Condition for Constrained Problems

Below we state the second-order sufficient conditions for achieving local optimality of (4.10.1).

Theorem 4.9. *Let x_* be a regular feasible solution of (4.10.1) such that the functions f, g_j, h_i are twice continuously differentiable in a neighborhood of x_*. If there exist Lagrange multipliers $\lambda_j^* \geq 0$, μ_i^* such that*

a) the KKT conditions hold:

$$\begin{array}{l}\nabla_x L(x_*;\lambda^*,\mu^*) = 0\\ \lambda_j^* g_j(x_*) = 0,\ j = 1,...,m.\end{array}$$

b) For every $d \neq 0$ orthogonal to the gradients, taken at x_*, of all equality constraints and those active at x_* inequality constraints for which $\lambda_j^* > 0$, one has

$$d^T \nabla_x^2 L(x_*;\lambda^*,\mu^*) d > 0,$$

then x_* is locally optimal for (4.10.1).

Proof. **1^0.** As in the case of Second Order Necessary Optimality Condition, we can reduce the situation to one where all inequality constraints are active at x_* and The problem is of the special form

$$\min_y \left\{ \phi(y) : \begin{array}{c} y_j \leq 0,\ j \leq m \\ y_{m+i} = 0,\ i \leq k \end{array} \right\}. \tag{4.11.1}$$

2^0. In the case of (4.11.1), the second-order sufficient condition reads: $\exists \lambda^* \geq 0, \mu^*$:

$$\begin{array}{c} \nabla_y \big|_{y=0} \left\{ \phi(y) + \sum\limits_{j=1}^m \lambda_j^* y_j + \sum\limits_{i=1}^k \mu_i^* y_{m+i} \right\} = 0 \\ d_j = 0, j \in J, d \neq 0 \Rightarrow d^T \nabla^2 \phi(0) d > 0 \\ \Big[J = \{ j \leq m : \lambda_j^* > 0 \} \cup \{ m+1, ..., m+k \} \Big] \end{array} \tag{4.11.2}$$

Assuming w.l.o.g. $\{j : \lambda_j^* > 0\} = \{1, ..., q\}$. Then (4.11.2) reads:

$$\begin{array}{rcl} \frac{\partial \phi(0)}{\partial y_\ell} & < & 0,\ \ell = 1,...,q (\text{inequality constraints with } \lambda_l^* > 0) \\ \frac{\partial \phi(0)}{\partial y_\ell} & = & 0,\ \ell = q+1,...,m (\text{inequality constraints with } \lambda_l^* = 0) \\ \frac{\partial \phi(0)}{\partial y_\ell} & = & 0,\ \ell = m+k+1,...,n (\text{inactive constraints}) \\ 0 \neq d & \in & T^+ = \{ d : d_\ell = 0,\ \ell \in \{1,...,q, m+1,...,m+k\} \} : \\ & \Rightarrow & d^T \nabla^2 \phi(0) d > 0 \end{array} \tag{4.11.3}$$

Our goal is to derive from this assumption the local optimality of $y_* = 0$ for (4.11.1). Note that the feasible set of (4.11.1) is the closed cone

$$K = \{d : d_\ell \leq 0, \ell = 1, ..., m, d_\ell = 0, \ell = m+1, ..., m+k\} \tag{4.11.4}$$

We need to use the following result.

Claim. *For* $0 \neq d \in K$ *one has* $d^T \nabla \phi(0) \geq 0$. *Moreover, if* $d^T \nabla \phi(0) = 0$, *then* $d^T \nabla^2 \phi(0) d > 0$.
Proof of the Claim: For $d \in K$, we have

$$d^T \nabla \phi(0) = \sum_{\ell=1}^{n} \frac{\partial \phi(0)}{\partial y_\ell} d_\ell$$

By (4.11.3) and (4.11.4), the first q terms in this sum are nonnegative, and the remaining are 0. Thus, the sum always is ≥ 0. For $d \neq 0$, the only possibility for the sum to vanish is to have $d \in T^+$, and in this case $d^T \phi''(0) d > 0$.
3^0. To summarize the situation, problem (4.11.1) is equivalent to

$$\min_{y \in K} \phi(y), \tag{4.11.5}$$

where K is a closed cone, ϕ is twice continuously differentiable in a neighborhood of the origin and is such that

$$\begin{gathered} d \in K \Rightarrow d^T \nabla \phi(0) \geq 0 \\ d \in K \backslash \{0\}, d^T \nabla \phi(0) = 0 \Rightarrow d^T \nabla^2 \phi(0) d > 0 \end{gathered}$$

Claim: 0 *is a locally optimal solution to (4.11.5).*
Proof of the Claim: Let $M = \{d \in K : \|d\|_2 = 1\}$, and let $M_0 = \{d \in M : d^T \nabla \phi(0) = 0\}$. Since K is closed, both M and M_0 are compact sets.
We know that $d^T \nabla^2 \phi(0) d > 0$ for $d \in M_0$. Since M_0 is a compact set, there exists a neighborhood V of M_0 and $\alpha > 0$ such that

$$d \in V \Rightarrow d^T \nabla^2 \phi(0) d \geq \alpha.$$

The set $V_1 = M \backslash V$ is compact and $d^T \nabla \phi(0) > 0$ when $d \in V_1$; thus, there exists $\beta > 0$ such that

$$d \in V_1 \Rightarrow d^T \nabla \phi(0) \geq \beta.$$

Note that K is a cone, and the set $M = \{d \in K : \|d\|_2 = 1\}$ is partitioned into two subsets $V_0 = V \cap M$ and V_1 in such a way that

$$\begin{aligned} &d \in V_0 \Rightarrow d^T \nabla \phi(0) \geq 0, d^T \nabla^2 \phi(0) d \geq \alpha > 0 \\ &d \in V_1 \to d^T \nabla \phi(0) \geq \beta > 0 \end{aligned}$$

Our goal is to prove that 0 is local minimizer of ϕ on K, or, which is the same, that

$$\begin{aligned} &\exists r > 0 : \\ &\quad \phi(0) \leq \phi(td) \, \forall (d \in M, 0 \leq t \leq r). \end{aligned}$$

Let $d \in M$, $t \geq 0$. When $d \in V_0$, we have

$$\begin{aligned} \phi(td) - \phi(0) &\geq td^T \nabla\phi(0) + \tfrac{1}{2}t^2 d^T \nabla^2\phi(0)d - t^2 \underbrace{R(t)}_{\substack{\to 0,\\ t\to +0}} \\ &\geq \tfrac{1}{2}t^2(\alpha - 2R(t)) \\ \Rightarrow \exists r_0 > 0: \quad & \phi(td) - \phi(0) \geq \tfrac{1}{4}t^2\alpha \geq 0 \forall t \leq r_0. \end{aligned}$$

When $d \in V_1$, we have

$$\begin{aligned} \phi(td) - \phi(0) &\geq td^T \nabla\phi(0) + \tfrac{1}{2}t^2 d^T \nabla^2\phi(0)d - t^2 \underbrace{R(t)}_{\substack{\to 0,\\ t\to +0}} \\ &\geq \beta t - Ct^2 - t^2 R(t) \\ \Rightarrow \exists r_1 > 0: \quad & \phi(td) - \phi(0) \geq \tfrac{\beta}{2}t \geq 0 \forall t \leq r_1. \end{aligned}$$

Thus, $\phi(td) - \phi(0) \geq 0$ for all $t \leq \min[r_0, r_1]$, $d \in M$.

■

Note that the difference between sufficient and necessary optimality conditions is in their "second order" parts and has twofold.

The first one is a minor difference. The necessary condition states positive *semi*definiteness of $\nabla_x^2 L(x_*; \lambda^*, \mu^*)$ along linear subspace:

$$\begin{aligned} \forall d \in T = \{d : \overbrace{d^T h_i'(x_*) = 0}^{\forall i \leq k}, \overbrace{d^T g_j'(x_*) = 0}^{\forall j \in J(x_*)}\} : \\ d^T \nabla_x^2 L(x_*; \lambda^*, \mu^*) d \geq 0 \end{aligned}$$

while Sufficient condition requires positive *definiteness* of $\nabla_x^2 L(x_*; \lambda^*, \mu^*)$ along linear subspace:

$$\begin{aligned} \forall 0 \neq d \in T^+ = \{d : \overbrace{d^T h_i'(x_*) = 0}^{\forall i \leq k}, \overbrace{d^T g_j'(x_*) = 0}^{\forall j : \lambda_j^* > 0}\} : \\ d^T \nabla_x^2 L(x_*; \lambda^*, \mu^*) d > 0 \end{aligned}$$

The second one is a major difference. The linear subspaces in question are different, and $T \subset T^+$; the subspaces are equal to each other iff *all* active at x_* inequality constraints have positive Lagrange multipliers λ_j^*. Observe that this "gap" is essential, as is shown by example

$$\min_{x_1, x_2} \left\{ f(x) = x_2^2 - x_1^2 : g_1(x) \equiv x_1 \leq 0 \right\} \qquad [x_* = (0,0)^T]$$

Here the second-order necessary optimality condition is satisfied "strictly":

$$L(x; \lambda) = x_2^2 - x_1^2 + \lambda x_1,$$

whence

$$\begin{array}{c}\lambda^* = 0 \Rightarrow \nabla_x L(x_*;\lambda^*) = 0,\\ T = \{d : d^T g_1'(0) = 0\} = \{d : d_1 = 0\},\\ 0 \neq d \in T \Rightarrow d^T \nabla_x^2 L(x_*;\lambda^*) d = d_2^2 > 0\}\end{array}$$

while x_* is *not* a local solution.

4.12 Applications: Eigenvalue Decomposition and S-Lemma

In this section, we discuss two important applications of optimality conditions.

Example 4.2. Consider optimization problem

$$\mathrm{Opt} = \min_{x\in\mathbb{R}^n} \left\{ f(x) = x^T Ax : h(x) = 1 - x^T x = 0 \right\} \qquad (P)$$

where $A = A^T$ is an $n \times n$ matrix. The problem clearly is solvable. Let x_* be its optimal solution.

What can we say about x_*? We first observe that x_* is a regular solution. Indeed, we should prove that the gradients of active at x_* constraints are linearly independent. There is only one constraint, and its gradient at the feasible set is nonzero.

Since x_* is a regular globally (and therefore locally) optimal solution, at x_* the second-order necessary optimality condition should hold: $\exists \mu^*$:

$$\begin{array}{c}\nabla_x \overbrace{\left[x^T Ax + \mu^*(1 - x^T x)\right]}^{L(x;\mu^*)} = 0 \Leftrightarrow 2(A - \mu^* I)x_* = 0\\ \underbrace{d^T \nabla_x h(x_*) = 0}_{\Leftrightarrow d^T x_* = 0} \Rightarrow \underbrace{d^T \nabla_x^2 L(x_*;\mu^*) d \geq 0}_{\Leftrightarrow d^T (A - \mu^* I) d \geq 0}\end{array}$$

Therefore, if x_* is optimal, then $\exists \mu^*$:

$$\begin{array}{rr} Ax_* = \mu^* x_* & (A)\\ d^T x_* = 0 \Rightarrow d^T (A - \mu^* I) d \geq 0 & (B)\end{array}$$

- (A) says that $x_* \neq 0$ is an eigenvector of A with eigenvalue μ^*; in particular, *we see that a symmetric matrix always has a real eigenvector*
- (B) along with (A) says that $y^T(A - \mu^* I)y \geq 0$ *for all* y. Indeed, every $y \in \mathbb{R}^n$ can be represented as $y = tx_* + d$ with $d^T x_* = 0$. We now have

$$\begin{array}{l} y^T[A - \mu^* I]y = (tx_* + d)^T [A - \mu^* I](tx_* + d)\\ = t^2 x_*^T \underbrace{[A - \mu^* I]x_*}_{=0} + 2td^T \underbrace{d^T [A - \mu^* I]x_*}_{=0}\\ \quad + \underbrace{d^T [A - \mu^* I]d}_{\geq 0} \geq 0\end{array}$$

Observe that in the case in question, the second-order necessary optimality conditions can be rewritten equivalently as $\exists \mu^*$:

$$\begin{array}{c}[A-\mu^* I]x_* = 0\\ y^T[A-\mu^* I]y \geq 0 \forall y.\end{array} \qquad (*)$$

In fact, these conditions are not only necessary, but also sufficient for feasible solution x_* to be globally optimal. To prove sufficiency, let x_* be feasible, and μ^* be such that $(*)$ holds true. For every feasible solution x, one has

$$0 \leq x^T[A-\mu^* I]x = x^T Ax - \mu^* x^T x = x^T Ax - \mu^*,$$

whence $x^T Ax \geq \mu^*$. For $x = x_*$, we have

$$0 = x_*^T[A-\mu^* I]x_* = x_*^T Ax_* - \mu^* x_*^T x_* = x_*^T Ax_* - \mu^*,$$

whence $x_*^T Ax_* = \mu^*$. Thus, x_* is globally optimal for (P), and μ^* is the optimal value in (P).

Example 4.3. **Extension: *S*-Lemma.** Let A, B be symmetric matrices, and let B be such that

$$\exists \bar{x} : \bar{x}^T B\bar{x} > 0. \qquad (*)$$

Then the inequality

$$x^T Ax \geq 0 \qquad (A)$$

is a consequence of the inequality

$$x^T Bx \geq 0 \qquad (B)$$

if and only if (A) is a “linear consequence” of (B): there exists $\lambda \geq 0$ such that

$$x^T[A-\lambda B]x \geq 0 \forall x \qquad (C)$$

that is, (A) is a weighted sum of (B) (weight $\lambda \geq 0$) and identically true inequality (C).

We provide a sketch of the proof of this important result. Note that the only nontrivial statement is that “*If* (A) is a consequence of (B), *then* t e exists $\lambda \geq 0$ such that ...”. To prove this statement, assume that (A) is a consequence of (B), i.e.,

$$\exists \bar{x} : \bar{x}^T B\bar{x} > 0; \underbrace{x^T Bx \geq 0}_{(B)} \Rightarrow \underbrace{x^T Ax \geq 0}_{(A)}$$

Consider optimization problem

$$\mathrm{Opt} = \min_x \left\{x^T Ax : h(x) \equiv 1 - x^T Bx = 0\right\}.$$

This problem is feasible by $(*)$, and $\mathrm{Opt} \geq 0$. Assume that an optimal solution x_* exists. Then, same as above, x_* is regular, and at x_* the Second Order Necessary condition holds true: $\exists\mu^*$:

$$\begin{array}{c}\nabla_x\big|_{x=x_*}\left[x^TAx+\mu^*[1-x^TBx]\right]=0\Leftrightarrow[A-\mu^*B]x_*=0\\ \underbrace{d^T\nabla_x\big|_{x=x_*}h(x)=0}_{\Leftrightarrow d^TBx_*=0}\Rightarrow d^T[A-\mu^*B]d\geq 0\end{array}$$

We have $0=x_*^T[A-\mu^*B]x_*$, that is, $\mu_*=\text{Opt}\geq 0$. Representing $y\in\mathbb{R}^n$ as tx_*+d with $d^TBx_*=0$ (that is, $t=x_*^TBy$), we get

$$\begin{array}{l}y^T[A-\mu^*B]y=t^2x_*^T\underbrace{[A-\mu^*B]x_*}_{=0}\\ +2td^T\underbrace{[A-\mu^*B]x_*}_{=0}+\underbrace{d^T[A-\mu^*B]d}_{\geq 0}\geq 0,\end{array}$$

Thus, $\mu^*\geq 0$ and $y^T[A-\mu^*B]y\geq 0$ for all y,

4.13 Exercises

1. Consider the optimization problem

$$\begin{array}{ll}\text{minimize} & f_0(x_1,x_2)\\ \text{subject to} & 2x_1+x_2\geq 1\\ & x_1+3x_2\geq 1\\ & x_1\geq 0,\quad x_2\geq 0.\end{array}$$

Make a sketch of the feasible set. For each of the following objective functions, give the optimal set and the optimal value.

(a) $f_0(x_1,x_2)=x_1+x_2$.
(b) $f_0(x_1,x_2)=-x_1-x_2$.
(c) $f_0(x_1,x_2)=x_1$.
(d) $f_0(x_1,x_2)=\max\{x_1,x_2\}$.
(e) $f_0(x_1,x_2)=x_1^2+9x_2^2$.

2. Prove that $x^\star=(1,1/2,-1)$ is optimal for the optimization problem

$$\begin{array}{ll}\text{minimize} & (1/2)x^TPx+q^Tx+r\\ \text{subject to} & -1\leq x_i\leq 1,\quad i=1,2,3,\end{array}$$

where

$$P=\begin{bmatrix}13&12&-2\\12&17&6\\-2&6&12\end{bmatrix},\quad q=\begin{bmatrix}-22.0\\-14.5\\13.0\end{bmatrix},\quad r=1.$$

3. A simple example. Consider the optimization problem

$$\begin{array}{ll} \text{minimize} & x^2+1 \\ \text{subject to} & (x-2)(x-4) \leq 0 \end{array}$$

with variable $x \in \mathbf{R}$.

(a) Analysis of primal problem. Give the feasible set, the optimal value, and the optimal solution.
(b) Lagrangian and dual function. Plot the objective x^2+1 versus x. On the same plot, show the feasible set, optimal point and value, and plot the Lagrangian $L(x,\lambda)$ versus x for a few positive values of λ. Verify the lower bound property ($p^\star \geq \inf_x L(x,\lambda)$ for $\lambda \geq 0$). Derive and sketch the Lagrange dual function g.
(c) Lagrange dual problem. State the dual problem, and verify that it is a concave maximization problem. Find the dual optimal value and dual optimal solution $\lambda^\star$. Does strong duality hold?
(d) Sensitivity analysis. Let $p^\star(u)$ denote the optimal value of the problem

$$\begin{array}{ll} \text{minimize} & x^2+1 \\ \text{subject to} & (x-2)(x-4) \leq u, \end{array}$$

as a function of the parameter u. Plot $p^\star(u)$. Verify that $dp^\star(0)/du = -\lambda^\star$.

4. Weak duality for unbounded and infeasible problems. The weak duality inequality, $d^\star \leq p^\star$, clearly holds when $d^\star = -\infty$ or $p^\star = \infty$. Show that it holds in the other two cases as well: If $p^\star = -\infty$, then we must have $d^\star = -\infty$, and also, if $d^\star = \infty$, then we must have $p^\star = \infty$.

5. Problems with one inequality constraint. Express the dual problem of

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & f(x) \leq 0 \end{array}$$

with $c \neq 0$, in terms of the conjugate f^*. Explain why the problem you give is convex. We do not assume f is convex.

6. Interpretation of LP dual via relaxed problems. Consider the inequality form LP

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

with $A \in \mathbf{R}^{m\times n}, b \in \mathbf{R}^m$. In this exercise we develop a simple geometric interpretation of the dual LP (5.22).

Let $w \in \mathbf{R}_+^m$. If x is feasible for the LP, i.e., satisfies $Ax \preceq b$, then it also satisfies the inequality

$$w^T Ax \leq w^T b.$$

Geometrically, for any $w \succeq 0$, the halfspace $H_w = \left\{x \mid w^T Ax \leq w^T b\right\}$ contains the feasible set for the LP. Therefore if we minimize the objective $c^T x$ over the halfspace H_w we get a lower bound on $p^\star$.

(a) Derive an expression for the minimum value of $c^T x$ over the halfspace H_w (which will depend on the choice of $w \succeq 0$).
(b) Formulate the problem of finding the best such bound, by maximizing the lower bound over $w \succeq 0$.
(c) Relate the results of (a) and (b) to the Lagrange dual of the LP.

7. Dual of general *LP*. Find the dual function of the LP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Gx \preceq h \\ & Ax = b. \end{array}$$

Give the dual problem, and make the implicit equality constraints explicit.

8. Piecewise-linear minimization. We consider the convex piecewise-linear minimization problem

$$\text{minimize} \max_{i=1,\cdots,m} \left(a_i^T x + b_i\right) \qquad (1)$$

with variable $x \in \mathbf{R}^n$.

(a) Derive a dual problem, based on the Lagrange dual of the equivalent problem

$$\begin{array}{ll} \text{minimize} & \max_{i=1,\cdots,m} y_i \\ \text{subject to} & a_i^T x + b_i = y_i, \quad i = 1, \cdots, m, \end{array}$$

with variables $x \in \mathbf{R}^n, y \in \mathbf{R}^m$.
(b) Formulate the piecewise-linear minimization problem (1) as an LP, and form the dual of the LP. Relate the LP dual to the dual obtained in part (a).
(c) Suppose we approximate the objective function in (1) by the smooth function

$$f_0(x) = \log\left(\sum_{i=1}^m \exp\left(a_i^T x + b_i\right)\right),$$

and solve the unconstrained geometric program

$$\text{minimize } \log\left(\sum_{i=1}^m \exp\left(a_i^T x + b_i\right)\right). \qquad (2)$$

A dual of this problem is given by (2). Let $p^\star_{\mathrm{pwl}}$ and $p^\star_{\mathrm{gp}}$ be the optimal values of (1) and (2), respectively. Show that

$$0 \le p^\star_{\mathrm{gp}} - p^\star_{\mathrm{pwl}} \le \log m$$

(d) Derive similar bounds for the difference between $p^\star_{\mathrm{pwl}}$ and the optimal value of

$$\mathrm{minimize} (1/\gamma) \log \left(\sum_{i=1}^{m} \exp \left(\gamma \left(a_i^T x + b_i \right) \right) \right)$$

where $\gamma > 0$ is a parameter. What happens as we increase γ ?

9. Derive a dual problem for

$$\mathrm{minimize} \sum_{i=1}^{N} \|A_i x + b_i\|_2 + (1/2) \|x - x_0\|_2^2 .$$

The problem data are $A_i \in \mathbf{R}^{m_i \times n}, b_i \in \mathbf{R}^{m_i}$, and $x_0 \in \mathbf{R}^n$. First introduce new variables $y_i \in \mathbf{R}^{m_i}$ and equality constraints $y_i = A_i x + b_i$.

10. Analytic centering. Derive a dual problem for

$$\mathrm{minimize} \ -\sum_{i=1}^{m} \log \left(b_i - a_i^T x \right)$$

with domain $\left\{ x \mid a_i^T x < b_i, i = 1, \cdots, m \right\}$. First introduce new variables y_i and equality constraints $y_i = b_i - a_i^T x$.

11. A penalty method for equality constraints. We consider the problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \qquad (1) \\ \text{subject to} & Ax = b \end{array}$$

where $f_0 : \mathbf{R}^n \to \mathbf{R}$ is convex and differentiable, and $A \in \mathbf{R}^{m \times n}$ with $\mathrm{rank} A = m$. In a quadratic penalty method, we form an auxiliary function

$$\phi(x) = f(x) + \alpha \|Ax - b\|_2^2,$$

where $\alpha > 0$ is a parameter. This auxiliary function consists of the objective plus the penalty term $\alpha \|Ax - b\|_2^2$. The idea is that a minimizer of the auxiliary function, $\tilde{x}$, should be an approximate solution of the original problem. Intuition suggests that the larger the penalty weight α, the better the approximation $\tilde{x}$ to a solution of the original problem. Suppose $\tilde{x}$ is a minimizer of ϕ. Show how to find, from $\tilde{x}$, a dual feasible point for (1). Find the corresponding lower bound on the optimal value of (1).

12. Consider the problem

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

where the functions $f_i : \mathbf{R}^n \to \mathbf{R}$ are differentiable and convex. Let $h_1, \cdots, h_m : \mathbf{R} \to \mathbf{R}$ be increasing differentiable convex functions. Show that

$$\phi(x) = f_0(x) + \sum_{i=1}^{m} h_i\left(f_i(x)\right)$$

is convex. Suppose $\tilde{x}$ minimizes ϕ. Show how to find from $\tilde{x}$ a feasible point for the dual of (1). Find the corresponding lower bound on the optimal value of (1).

13. A convex problem in which strong duality fails. Consider the optimization problem

$$\begin{array}{ll} \text{minimize} & e^{-x} \\ \text{subject to} & x^2/y \leq 0 \end{array}$$

with variables x and y, and domain $\mathscr{D} = \{(x, y) \mid y > 0\}$.

(a) Verify that this is a convex optimization problem. Find the optimal value.
(b) Give the Lagrange dual problem, and find the optimal solution $\lambda^\star$ and optimal value $d^\star$ of the dual problem. What is the optimal duality gap?
(c) Does Slater's condition hold for this problem?
(d) What is the optimal value $p^\star(u)$ of the perturbed problem

$$\begin{array}{ll} \text{minimize} & e^{-x} \\ \text{subject to} & x^2/y \leq u \end{array}$$

as a function of u ? Verify that the global sensitivity inequality

$$p^\star(u) \geq p^\star(0) - \lambda^\star u$$

does not hold.

14. Convex-concave functions and the saddle-point property. We derive conditions under which the saddle-point property

$$\sup_{z \in Z} \inf_{w \in W} f(w, z) = \inf_{w \in W} \sup_{z \in Z} f(w, z) \qquad (1)$$

holds, where $f : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}, W \times Z \subseteq \operatorname{dom} f$, and W and Z are nonempty. We will assume that the function

$$g_z(w) = \begin{cases} f(w, z) & w \in W \\ \infty & \text{otherwise} \end{cases}$$

is closed and convex for all $z \in Z$, and the function

$$h_w(z) = \begin{cases} -f(w,z) & z \in Z \\ \infty & \text{otherwise} \end{cases}$$

is closed and convex for all $w \in W$.

(a) The righthand side of (1) can be expressed as $p(0)$, where

$$p(u) = \inf_{w \in W} \sup_{z \in Z} \left(f(w,z) + u^T z \right).$$

Show that p is a convex function.

(b) Show that the conjugate of p is given by

$$p^*(v) = \begin{cases} -\inf_{w \in W} f(w,v) & v \in Z \\ \infty & \text{otherwise.} \end{cases}$$

(c) Show that the conjugate of p^* is given by

$$p^{**}(u) = \sup_{z \in Z} \inf_{w \in W} \left(f(w,z) + u^T z \right).$$

Combining this with (a), we can express the max-min equality (1) as $p^{**}(0) = p(0)$

(d) We know that $p^{**}(0) = p(0)$ if $0 \in \textbf{int dom} p$. Conclude that this is the case if W and Z are bounded.

(e) As another consequence of exercises 3.28 and 3.39, we have $p^{**}(0) = p(0)$ if $0 \in \textbf{dom}\ p$ and p is closed. Show that p is closed if the sublevel sets of g_z are bounded.

15. Consider the QCQP

$$\begin{array}{ll} \text{minimize} & x_1^2 + x_2^2 \\ \text{subject to} & (x_1-1)^2 + (x_2-1)^2 \le 1 \\ & (x_1-1)^2 + (x_2+1)^2 \le 1 \end{array}$$

with variable $x \in \mathbf{R}^2$.

(a) Sketch the feasible set and level sets of the objective. Find the optimal point $x^\star$ and optimal value $p^\star$.

(b) Give the KKT conditions. Do there exist Lagrange multipliers $\lambda_1^\star$ and $\lambda_2^\star$ that prove that $x^\star$ is optimal?

(c) Derive and solve the Lagrange dual problem. Does strong duality hold?

16. Find the minimizer of a linear function

$$f(x) = c^T x$$

on the set

$$V_p = \left\{ x \in \mathbf{R}^n : \sum_{i=1}^{n} |x_i|^p \leq 1 \right\};$$

here $p, 1 < p < \infty$, is a parameter. What happens with the solution when the parameter becomes 0.5 ?

17. Let $a_1, \cdots, a_n > 0, \alpha, \beta > 0$. Solve the optimization problem

$$\min_x \left\{ \sum_{i=1}^{n} \frac{a_i}{x_i^\alpha} : x > 0, \sum_i x_i^\beta \leq 1 \right\}$$

18. Consider the optimization problem

$$\max_{x,t} \left\{ \xi^T x + \tau t + \ln\left(t^2 - x^T x\right) : (t,x) \in X = \left\{ t > \sqrt{x^T x} \right\} \right\}$$

where $\xi \in \mathbf{R}^n, \tau \in \mathbf{R}$ are parameters. Is the problem convex $^{5)}$? What is the domain in space of parameters where the problem is solvable? What is the optimal value? Is it convex in the parameters?

19. Consider the optimization problem

$$\max_{x,y} \{ f(x,y) = ax + by + \ln(\ln y - x) + \ln(y) : (x,y) \in X = \{ y > \exp\{x\} \} \},$$

where $a, b \in \mathbf{R}$ are parameters. Is the problem convex? What is the domain in space of parameters where the problem is solvable? What is the optimal value? Is it convex in the parameters?

20. Consider the problem of minimizing the linear form

$$f(x) = x_2 + 0.1 x_1$$

on the $2D$ plane over the triangle with the vertices $(1,0), (0,1), (0,1/2)$ (draw the picture!).
1) Verify that the problem has unique optimal solution $x_* = (1,0)$ Solution: clear from the picture, or from computing the values of the objective at the 3 vertices of the triangle (a linear function on a polytope attains its extrema at vertices).
2) Verify that the problem can be posed as the LP program

$$\min_x \{x_2 + 0.1x_1 : x_1 + x_2 \le 1, x_1 + 2x_2 \ge 1, x_1, x_2 \ge 0\}.$$

21. Consider the following elementary problem:

$$\min_x \left\{ f(x_1, x_2) = x_1^2 - x_2 : h(x) \equiv x_2 = 0 \right\}$$

with the evident unique optimal solution $(0,0)$. Is the KKT condition satisfied at this solution? Solution: $x_* = (0,0)$ is regular and locally optimal, the data are smooth, so that the KKT is satisfied at x_*.
Rewrite the problem equivalently as

$$\min_x \left\{ f(x_1, x_2) = x_1^2 - x_2 : h(x) \equiv x_2^2 = 0 \right\}.$$

What about the KKT condition in this equivalent problem?

22. Consider an inequality constrained optimization problem

$$\min_x \{ f(x) : g_i(x) \le 0, i = 1, \cdots, m \}.$$

Assume that x_* is locally optimal solution, f, g_i are continuously differentiable in a neighbourhood of x_* and the constraints g_i are concave in this neighbourhood. Prove that x_* is locally optimal solution to the linearized problem

$$\min_x \left\{ f(x_*) + (x - x_*)^T \nabla f(x_*) : (x - x_*)^T \nabla g_j(x_*) = 0, j \in J(x_*) = \left\{ j : g_j(x_*) = 0 \right\} \right\}.$$

Is x_* a *KKT* point of the problem?

23. Let $a_1, \cdots, a_n$ be positive reals, and let $0 < s < r$ be two reals. Find maximum and minimum of the function

$$\sum_{i=1}^{n} a_i |x_i|^r$$

on the surface

$$\sum_{i=1}^{n} |x_i|^s = 1$$

24. Recall S-Lemma: If A, B are two symmetric $n \times n$ matrices such that

$$\bar{x}^T B \bar{x} > 0$$

for certain $\bar{x}$, then the implication

$$x^T B x \ge 0 \Rightarrow x^T A x \ge 0 \qquad (*)$$

holds true iff there exists $\lambda \geq 0$ such that $A - \lambda B \succeq 0$
($P \succeq 0$ means that P is symmetric positive semidefinite: $P = P^T$ and $x^T Px \geq 0$ for all $x \in \mathbf{R}^n$) The proof given in Lecture on optimality conditions (see Transparencies) was incomplete - it was taken for granted that certain optimization problem has an optimal solution. Fill the gap in the proof or find an alternative proof.
Fact: Let $A = A^T$ be a symmetric $n \times n$ matrix, and let B, C be $m \times n$ matrices, $C \neq 0$. Then all matrices

$$A + B^T \Delta C + C^T \Delta^T B$$

corresponding to $m \times m$ matrices Δ with $\|\Delta\| \leq 1^{6)}$ are positive semidefinite iff there exists $\lambda \geq 0$ such that the matrix

$$\left[\begin{array}{c|c} A - \lambda C^T C & -B^T \\ \hline -B & \lambda I_m \end{array}\right]$$

is positive semidefinite.

25. An Example of Lagrangian Duality
Consider the problem

$$\begin{array}{ll} \text{minimize} & f(x) \qquad (1) \\ \text{subject to} & x \in X, \quad e_i' x = d_i, \quad i = 1, \cdots, m, \end{array}$$

where $f : \Re^n \mapsto \Re$ is a convex function, X is a nonempty convex set, and e_i and d_i are given vectors and scalars, respectively. Consider the min common/max crossing framework where M is the subset of $\Re^{m+1}$ given by

$$M = \left\{ \left(e_1' x - d_1, \cdots, e_m' x - d_m, f(x) \right) \mid x \in X \right\}$$

and assume that $w^* < \infty$.

(a) Show that w^* is equal to the optimal value of problem (1), and that the max crossing problem is to maximize $q(\mu)$ given by

$$q(\mu) = \inf_{x \in X} \left\{ f(x) + \sum_{i=1}^{m} \mu_i \left(e_i' x - d_i \right) \right\}$$

(b) Show that the corresponding set $\bar{M}$ is convex.
(c) Show that if X is compact, then $q^* = w^*$.
(d) Show that if there exists a vector $\bar{x} \in \mathrm{ri}(X)$ such that $e_i' \bar{x} = d_i$ for all $i = 1, \cdots, m$, then $q^* = w^*$ and the max crossing problem has an optimal solution.

26. Lagrangian Duality and Compactness of the Constraint Set

Consider the problem of Exercise 25, but assume that f is linear and X is compact (instead of f and X being convex). Show that q^* is equal to the minimal value of $f(x)$ subject to $x \in \text{conv}(X)$ and $e_i'x = d_i, i = 1, \cdots, m$. Hint: Show that

$$\text{conv}(M) = \left\{ \left(e_1'x - d_1, \cdots, e_m'x - d_m, f(x)\right) \mid x \in \text{conv}(X) \right\}$$

and use Exercise 25(c)

27. Monotone transformation of the objective. Consider the optimization problem

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

where $f_i : \mathbf{R}^n \to \mathbf{R}$ for $i = 0, 1, \cdots, m$ are convex. Suppose $\phi : \mathbf{R} \to \mathbf{R}$ is increasing and convex. Then the problem

$$\begin{array}{ll} \text{minimize} & \tilde{f}_0(x) = \phi\left(f_0(x)\right) \qquad (2) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \cdots, m \end{array}$$

is convex and equivalent to it; in fact, it has the same optimal set as (1).
In this problem we explore the connections between the duals of the two problems (1) and (2). We assume f_i are differentiable, and to make things specific, we take $\phi(a) = \exp a$.

a. Suppose λ is feasible for the dual of (1), and $\bar{x}$ minimizes

$$f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x).$$

Show that $\bar{x}$ also minimizes

$$\exp f_0(x) + \sum_{i=1}^{m} \tilde{\lambda}_i f_i(x)$$

for appropriate choice of $\tilde{\lambda}$. Thus, $\tilde{\lambda}$ is dual feasible for (2).

b. Let $p^\star$ denote the optimal value of (1) (so the optimal value of (2) is $\exp p^\star$). From λ we obtain the bound

$$p^\star \geq g(\lambda),$$

where g is the dual function for (1). From $\tilde{\lambda}$ we obtain the bound $\exp p^\star \geq \tilde{g}(\tilde{\lambda})$, where $\tilde{g}$ is the dual function for (2). This can be expressed as

$$p^\star \geq \log \tilde{g}(\tilde{\lambda}).$$

How do these bounds compare? Are they the same, or is one better than the other?

Chapter 5
Optimization Methods

5.1 Introduction

Our goal is to find approximate numerically solutions to Mathematical Programming (MP) problems

$$\min_x \left\{ f(x) : \begin{array}{l} g_j(x) \leq 0,\ j = 1, ..., m \\ h_i(x) = 0,\ i = 1, ..., k \end{array} \right\} \tag{5.1.1}$$

Most MP algorithms to be considered in the Chapter do not assume the analytic structure of (5.1.1) to be known in advance (and do not know how to use the structure when it is known). These algorithms are black-box-oriented: when solving (5.1.1), method generates a sequence of iterates x_1, x_2,... in such a way that x_{t+1} depends solely on local information of (5.1.1) gathered along the preceding iterates $x_1, ..., x_t$. Information on (5.1.1) obtained at x_t usually is comprised of the values and the first and the second derivatives of the objective and the constraints at x_t.

In some cases, local information, available to black-box-oriented algorithms, is really poor, so that approximating global solution to the problem becomes seeking needle in multidimensional haystack. Let us look at a 3D haystack with 2 m edges, and let a needle be a cylinder of height 20 mm and radius of cross-section 1 mm (see Figure 5.1).

So how difficult it is to find the needle in the haystack? In the optimization setting: We want to minimize a smooth function f which is zero "outside of the needle" and negative inside it. When only local information on the function is available, we get trivial information unless the sequence of iterates we are generating hits the needle. As a result, it is easy to show that the number of iterations needed to hit the needle with a reasonable confidence cannot be much smaller than when generating the iterates at random. In this case, the probability for an iterate to hit a needle is as small as 7.8×10^{-9}, that is, to find the needle with a reasonable confidence, we need to generate *hundreds of millions* of iterates. Moreover, as the dimension of the problem grows, the indicated difficulties are dramatically amplified. For example, preserving the linear sizes of the haystack and the needle and increasing the dimension of the

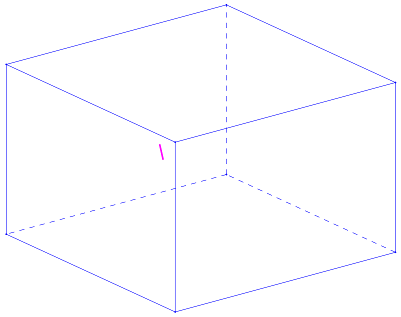


Fig. 5.1: Haystack and the needle

haystack from 3 to 20, the probability for an iterate to hit the needle becomes as small as 8.9×10^{-67} !

In the "needle in the haystack" problem it is easy to find a *locally optimal* solution. However, slightly modifying the problem, we can make the latter task disastrously difficult as well. In unconstrained minimization, it is not too difficult to find a point where the *gradient* of the objective becomes small, i.e., where the first-order necessary optimality condition is "nearly" satisfied. On the other hand, in constrained minimization, it could be disastrously difficult to find just a feasible solution. However, the classical algorithms of continuous optimization, while providing no meaningful guarantees in the worst case, are capable to process quite efficiently typical optimization problems arising in applications.

In optimization, there exist algorithms which do exploit problem's structure and allow to approximate the global solution in a reasonable time. Traditional methods of this type – Simplex method and its variations – do not go beyond Linear Programming and Linearly Constrained Convex Quadratic Programming. In 1990's, new efficient ways to exploit problem's structure were discovered (Interior Point methods). The resulting algorithms, however, do not go beyond Convex Programming. Except for very specific and relatively simple problem classes, like Linear Programming or Linearly Constrained Quadratic Programming, optimization algorithms cannot guarantee finding exact solution – local or global – in finite time. The best we can expect from these algorithms is *convergence* of approximate solutions generated by algorithms to the exact solutions. Even in the case when "finite" solution methods do exist (Simplex method in Linear Programming), no reasonable complexity bounds for these methods are known, therefore in reality the ability of a method to generate the exact solution in finitely many steps is neither necessary, nor sufficient to justify the method.

Aside of convex programming, traditional optimization methods are unable to guarantee convergence to a globally optimal solution. Indeed, in the non-convex case there is no way to conclude from local information whether a given point is/is not globally optimal. In order to guarantee approximating global solution, it seems unavoidable to "scan" a dense set of the values of x in order to be sure that the globally optimal solution is not missed. Theoretically, such a possibility exists; however, the complexity of "exhaustive search" methods blows up exponentially with the dimension of the decision vector, which makes these methods completely impractical.

Traditional optimization methods do not incorporate exhaustive search and, as a result, cannot guarantee convergence to a global solution. A typical theoretical result on a traditional optimization method as applied to a general (not necessary convex) problem sounds like:

> *Assume that problem (5.1.1) possesses the following properties:*
> *...*
> *...*
> *Then the sequence of approximate solutions generated by method X is bounded, and all its limiting points are KKT points of the problem.*

or

> *Assume that x_* is a nondegenerate local solution to (5.1.1). Then method X, started close enough to x_*, converges to x_*.*

There are two major traditional classifications of MP algorithms. The first one is to classify algorithms by application fields primarily into

a) algorithms for unconstrained optimization,
b) algorithms for constrained optimization.

The second one is to classify algorithms by information used, primarily into

a) zeroth-order methods which use only the values of the objective and the constraints,
b) first-order methods (use both values and first order derivatives),
c) second order methods (use values, first- and second order derivatives).

5.2 Rate of Convergence

There is a necessity to quantify the convergence properties of MP algorithms. Traditionally, this is done via *asymptotical rate of convergence* defined as follows.

First, we introduce an appropriate error measure – a nonnegative function $\mathrm{Error}_P(x)$ of approximate solution and of the problem we are solving which is zero exactly at the set X_* of solutions to (5.1.1) we intend to approximate. Below are a few examples of error measure.

a) Distance to the set X_*:

$$\mathrm{Error}_P(x) = \inf_{x_* \in X_*} \|x - x_*\|_2$$

.

b) Residual in terms of the objective and the constraints

$$\mathrm{Error}_P(x) = \max \left[f(x) - \mathrm{Opt}(P), \right.$$
$$[g_1(x)]_+, ..., [g_m(x)]_+,$$
$$\left. |h_1(x)|, ..., |h_k(x)| \right]$$

Second, assume that we have established convergence of our method, that is, we know that if x_t^* are approximate solutions generated by method as applied to a problem (5.1.1) from a given family, then

$$\mathrm{Error}_P(t) \equiv \mathrm{Error}_P(x_t^*) \to 0, t \to \infty$$

We then roughly quantify the rate at which the sequence $\mathrm{Error}_P(t)$ of nonnegative reals converges to 0. Specifically, we say that

a) the method converges sublinearly, if the error goes to zero less rapidly than a geometric progression, e.g., as $1/t$ or $1/t^2$.
b) the method converges linearly, if there exist $C < \infty$ and $q \in (0,1)$ such that

$$\mathrm{Error}_{(P)}(t) \leq Cq^t.$$

Here q is called the convergence ratio. E.g.,

$$\mathrm{Error}_P(t) \asymp e^{-at}$$

exhibits linear convergence with ratio e^{-a}. A sufficient condition for linear convergence with ratio $q \in (0,1)$ is that

$$\overline{\lim_{t\to\infty}} \frac{\mathrm{Error}_P(t+1)}{\mathrm{Error}_P(t)} < q$$

c) the method converges superlinearly, if the sequence of errors converges to 0 faster than every geometric progression:

$$\forall q \in (0,1) \exists C : \mathrm{Error}_P(t) \leq Cq^t$$

For example,

$$\mathrm{Error}_P(t) \asymp e^{-at^2}$$

corresponds to superlinear convergence. A sufficient condition for superlinear convergence is

$$\lim_{t\to\infty} \frac{\mathrm{Error}_P(t+1)}{\mathrm{Error}_P(t)} = 0.$$

d) the method exhibits convergence of order $p > 1$, if

$$\exists C : \mathrm{Error}_P(t+1) \leq C\left(\mathrm{Error}_P(t)\right)^p$$

Convergence of order 2 is called quadratic. For example,

$$\mathrm{Error}_P(t) = e^{-ap^t}$$

converges to 0 with order p.

We can provide informal explanation for the above convergence rates. When the method converges, $\mathrm{Error}_P(t)$ goes to 0 as $t \to \infty$, that is, eventually the decimal representation of $\mathrm{Error}_P(t)$ has zero before the decimal dot and more and more zeros after the dot; the number of zeros following the decimal dot is called the number of accuracy digits in the corresponding approximate solution. Traditional classification of rates of convergence is based on how many steps, asymptotically, is required to add a new accuracy digit to the existing ones.

a) With sublinear convergence, the "price" of accuracy digit grows with the position of the digit. For example, with rate of convergence $O(1/t)$ every new accuracy digit is 10 times more expensive, in terms of # of steps, than its predecessor.
b) With linear convergence, every accuracy digit has the same price, proportional to $\ln^{-1}(1/\text{convergence ratio})$. Equivalently: every step of the method adds a fixed number r of accuracy digits (for q not too close to 0, $r \approx 1-q$).
c) With superlinear convergence, every subsequent accuracy digit eventually becomes cheaper than its predecessor – the price of accuracy digit goes to 0 as the position of the digit grows. Equivalently, every additional step adds more and more accuracy digits.
d) With convergence of order $p > 1$, the price of accuracy digit not only goes to 0 as the position k of the digit grows, but does it rapidly enough – in a geometric progression. Equivalently, eventually every additional step of the method multiplies by p the number of accuracy digits.

With the traditional approach, the convergence properties of a method are the better the higher is the "rank" of the method in the above classification. Given a family of problems, traditionally it is thought that linearly converging on every problem of the family method is faster than a sublinearly converging, superlinearly converging method is faster than a linearly converging one, etc.

Observe that usually we are able to prove existence of parameters C and q quantifying linear convergence:

$$\mathrm{Error}_P(t) \le Cq^t$$

or convergence of order $p > 1$:

$$\mathrm{Error}_P(t+1) \le C(\mathrm{Error}_P(t))^p,$$

but are unable to find numerical values of these parameters – they may depend on "unobservable" characteristics of a particular problem we are solving. As a result, traditional "quantification" of convergence properties is qualitative and asymptotical.

We have seen that as applied to general MP programs, optimization methods have a number of severe theoretical limitations, including the following major ones:

a) Unless exhaustive search (completely unrealistic in high-dimensional optimization) is used, there are no guarantees of approaching global solution.
c) Quantification of convergence properties is of asymptotical and qualitative character. As a result, the most natural questions like:

> *We should solve problems of such and such structure with such and such sizes and the data varying in such and such ranges. How many steps of method X are sufficient to solve problems within such and such accuracy?*

usually do not admit theoretically valid answers. This latter question is called complexity analysis in modern terms.

In spite of their theoretical limitations, in reality traditional MP algorithms allow to solve many, if not all, MP problems of real-world origin, including those with many thousands variables and constraints. Moreover, there exists a "solvable case" when practical efficiency admits solid theoretical guarantees – the complexity analysis of convex programming. More recently, the complexity analysis has been generalized to finding stationary points for nonconvex optimization.

5.3 Gradient Descent Method

Smooth Functions

We first introduce some notation for differentiable functions. Let $Q \subseteq \mathbb{R}^n$. We say $f \in C^k(Q)$ if $f : Q \in \mathbb{R}$ is k times continuously differentiable and $f \in C_L^{k,p}(Q)$ if $f : Q \in \mathbb{R}$ is k times continuously differentiable on Q and its p^{th} derivative satisfies

$$\|f^{(p)}(x) - f^{(p)}(y)\| \le L\|x-y\|, \quad \forall x, y \in Q.$$

The most important class is $C_L^{1,1}(\mathbb{R}^n)$ and for $f \in C_L^{1,1}(\mathbb{R}^n)$, we have

$$\|f'(x) - f'(y)\| \le L\|x-y\|, \quad \forall x, y \in \mathbb{R}^n.$$

Some examples are given by (a) $f(x) = a_0 + \langle a, x \rangle$, (b) $f(x) = a_0 + \langle a, x \rangle + \frac{1}{2}\langle Ax, x \rangle$, and (c) $f(x) = \sqrt{1+x^2}$.

Below are some important properties of $C_L^{1,1}(\mathbb{R}^n)$.

Lemma 5.1. *Let $f \in C^2(\mathbb{R}^n)$. $f \in C_L^{(1,1)}$ iff $\|f''(x)\| \le L$, $\forall x \in \mathbb{R}^n$.*

Proof. For any $x, y \in \mathbb{R}^n$, we have $f'(y) = f'(x) + \int_0^1 f''(x + \tau(y-x))(y-x)d\tau$. Therefore,

$$\begin{aligned}\|f'(y) - f'(x)\| &= \|\textstyle\int_0^1 f''(x+\tau(y-x))(y-x)d\tau\|\\ &\le \|\textstyle\int_0^1 f''(x+\tau(y-x))d\tau\|\|y-x\|\\ &\le \textstyle\int_0^1 \|f''(x+\tau(y-x))\|d\tau\|y-x\| \le L\|y-x\|.\end{aligned}$$

On the other hand, if $\|f'(y)-f'(x)\| \le L\|y-x\|, \forall x,y \in \mathbb{R}^n$,

$$\| \left(\int_0^{\alpha} f''(x+\tau s)d\tau\right) s\| = \|f'(x+\alpha s)-f'(x)\| \quad \le aL\|s\|.$$

Dividing both sides by α and tending $\alpha \to 0$, we have $\|f''(x)\| \le L$. ■

Lemma 5.2. *Let $f \in C_L^{1,1}(\mathbb{R}^n)$. Then $\forall x,y \in \mathbb{R}^n$, $|f(y)-f(x)-\langle f'(x),y-x\rangle| \le \frac{L}{2}\|y-x\|^2$.*

Proof. For any $x,y \in \mathbb{R}^n$, we have

$$\begin{aligned} f(y) &= f(x)+\textstyle\int_0^1 \langle f'(x+\tau(y-x),y-x\rangle d\tau \\ &= f(x)+\langle f'(x),y-x\rangle+\textstyle\int_0^1 \langle f'(x+\tau(y-x))-f'(x),y-x\rangle d\tau. \end{aligned}$$

Hence,

$$\begin{aligned} &|f(y)-f(x)-\langle f'(x),y-x\rangle| \\ &\le |\textstyle\int_0^1 \langle f'(x+\tau(y-x))-f'(x),y-x\rangle d\tau| \\ &\le \textstyle\int_0^1 |\langle f'(x+\tau(y-x))-f'(x),y-x\rangle| d\tau \\ &\le \textstyle\int_0^1 L\tau\|y-x\|^2 = \frac{L}{2}\|y-x\|^2. \end{aligned}$$

■

Similarly, we have the following results.

Lemma 5.3. *Let $f \in C_L^{2,2}(\mathbb{R}^n)$.*

$$\begin{aligned} &\|f'(y)-f'(x)-f''(x)(y-x)\| \le \tfrac{L}{2}\|y-x\|^2 \\ &|f(y)-f(x)-\langle f'(x),y-x\rangle-\tfrac{1}{2}\langle f''(x)(y-x),y-x\rangle| \le \tfrac{L}{6}\|y-x\|^3. \end{aligned}$$

Proof. Indeed,

$$\begin{aligned} \|f'(y)-f'(x)-f''(x)(y-x)\| &= \| \int_0^1 [f''(x+\tau(y-x))-f''(x)](y-x)d\tau\| \\ &\le L\|y-x\|^2 \int_0^1 \tau d\tau = \tfrac{1}{2}L\|y-x\|^2. \end{aligned}$$

Therefore,

$$\begin{aligned} &|f(y)-f(x)-\langle f'(x),y-x\rangle-\tfrac{1}{2}\langle f''(x)(y-x),y-x\rangle| \\ &= |\int_0^1 \langle f'(x+\lambda(y-x))-f'(x)-\lambda f''(x)(y-x),y-x\rangle d\lambda| \\ &\le \tfrac{1}{2}L\|y-x\|^3 \int_0^1 \lambda^2 d\lambda = \tfrac{L}{6}\|y-x\|^3. \end{aligned}$$

■

Corollary 5.1. *Let $f \in C_M^{2,2}(\mathbb{R}^n)$ and $\|y-x\| = r$. Then $f''(x)-MrI_n \preceq f''(y) \preceq f''(x)+MrI_n$, where I_n is the identity matrix.*

Proof. Denote $G = f''(y)-f''(x)$. Since $f \in C_M^{2,2}(\mathbb{R}^n)$, $\|G\| \le Mr$, i.e., $|\lambda_i(G)| \le Mr, i = 1,\dots,n$. Hence $-MrI_n \preceq G \preceq MrI_n$. ■

Gradient Descent for Nonconvex Problems

In this section, our goal is to find an approximate solution to

$$\min_{x\in\mathbb{R}^n} f(x), \tag{5.3.1}$$

where $f\in C_L^{1,1}(\mathbb{R}^n)$. First observe that the direction $-f'(\bar{x})$ (antigradient) is the direction of the fastest local decrease of f at point $\bar{x}$. Indeed, let $s\in\mathbb{R}^n$ and $\|s\|=1$. Consider the local decrease of $f(x)$ along s:

$$\Delta(s) = \lim_{\alpha\to 0}\tfrac{1}{\alpha}\left[f(\bar{x}+\alpha s)-f(\bar{x})\right] = \langle f'(\bar{x}), s\rangle.$$

We have $\langle f'(\bar{x}),s\rangle \geq -\|f'(\bar{x})\|\|s\| = -\|f'(\bar{x})\|$ and the equality holds when $s = -f'(\bar{x})/\|f'(x)\|$. In view of this discussion, we can define the gradient descent method as follows.

Gradient Descent (GD): Choose $x_0\in\mathbb{R}^n$ and set

$$x_{k+1} = x_k - h_k f'(x_k), k=0,1,\ldots.$$

The key issue in GD is how to choose stepsize h_k. There exist a few different ways to specify this algorithmic parameter.

a) Constant: $h_k = h > 0, k = 0, 1.$
b) Full relaxation: $h_k = \text{argmin}_{h\geq 0} f(x_k - hf'(x_k))$.
c) Armijo Rule: Choose h_k s.t.

$$\begin{aligned}\alpha\langle f'(x_k), x_k - x_{k+1}\rangle &\leq f(x_k) - f(x_{k+1})\\ \beta\langle f'(x_k), x_k - x_{k+1}\rangle &\geq f(x_k) - f(x_{k+1}),\end{aligned}$$

where $0<\alpha<\beta<1$ are some fixed parameters.

Theorem 5.1. *For any one of the stepsize policy, each iteration of the GD method satisfies*

$$f(x_{k+1}) \leq f(x_k) - \tfrac{w}{2}\|f'(x_k)\|^2, \tag{5.3.2}$$

for some $w>0$. As a consequence, we have $\lim_{k\to+\infty}\|f'(x_k)\|\to 0$ *and*

$$\min_{0\leq k\leq N}\|f'(x_k)\| \leq \tfrac{1}{\sqrt{N+1}}\left[\tfrac{L}{w}(f(x_0)-f^*)\right]^{1/2}. \tag{5.3.3}$$

Proof. Let $y = x - hf'(x)$.

$$\begin{aligned}f(y) &\leq f(x) + \langle f'(x), y-x\rangle + \tfrac{L}{2}\|y-x\|^2\\ &\leq f(x) - h\|f'(x)\|^2 + \tfrac{Lh^2}{2}\|f'(x)\|^2\\ &\leq f(x) - h(1-\tfrac{Lh}{2})\|f'(x)\|^2.\end{aligned}$$

We can now consider the following cases.

a) Constant stepsize. $x_{k+1} = x_k - hf'(x_k)$,

$$f(x_{k+1}) \leq f(x_k) - h(1 - \tfrac{Lh}{2})\|f'(x_k)\|^2.$$

Need $h < 2/L$ to guarantee the descreasing of f. Let $h = 2\alpha/L$, $\alpha \in (0,1)$,

$$f(x_{k+1}) \leq f(x_k) - \tfrac{2}{L}\alpha(1-\alpha)\|f'(x_k)\|^2.$$

The optimal choices: $\alpha = \frac{1}{2}$, $h_k = \frac{1}{L}$.

b) Full relaxation. $h_k = \mathrm{argmin}_{h\geq 0} f(x_k - hf'(x_k))$. We must have

$$f(x_{k+1}) \leq f(x_k) - \tfrac{1}{2L}\|f'(x_k)\|^2. \quad (Why?)$$

c) Armijo rule.

$$f(x_k) + \beta\langle f'(x_k), x_{k+1} - x_k\rangle \leq f(x_{k+1}) \leq f(x_k) + \alpha\langle f'(x_k), x_{k+1} - x_k\rangle$$

Or equivalently,

$$f(x_k) - \beta h_k\|f'(x_k)\|^2 \leq f(x_{k+1}) \leq f(x_k) - \alpha h_k\|f'(x_k)\|^2$$

In view of $f(x_{k+1}) \leq f(x_k) - h_k(1 - \frac{h_k}{2}L)\|f'(x_k)\|^2$, we have $-\beta h_k \leq -h_k(1 - \frac{h_k}{2}L)$ and thus $h_k \geq \frac{2}{L}(1-\beta)$. Moreover,

$$f(x_{k+1}) \leq f(x_k) - \alpha h_k\|f'(x_k)\|^2 \leq f(x_k) - \tfrac{2}{L}\alpha(1-\beta)\|f'(x_k)\|^2.$$

Therefore, we show that (5.3.2) holds for all these stepsize policies. Summing up (5.3.2) for $k = 0, 1, \ldots, N$,

$$f(x_{N+1}) \leq f(x_0) - \tfrac{w}{L}\textstyle\sum_{k=0}^N \|f'(x_k)\|^2.$$

Since $f(x_{N+1} \geq f^*$,

$$\tfrac{w}{L}\textstyle\sum_{k=0}^N \|f'(x_k)\|^2 \leq f(x_0) - f(x_{N+1}) \leq f(x_0) - f^*,$$

from which the results immediately follow. ■

In view of Theorem 5.1, for solving the problem $\min_x f(x)$ with $f \in C_L^{1,1}(\mathbb{R}^n)$ and bounded below. Using the first-order black box oracle and the termination criterion $\|f'(\bar{x})\| \leq \varepsilon$, the number of iterations (or iteration complexity) of the GD method can be bounded by $\mathscr{O}(\frac{L}{w\varepsilon^2}(f(x_0) - f^*)$.

Example 5.1. This example shows that GD converges to a stationary point instead of a local minimum. Let $f(x) = \frac{1}{2}(x_1)^2 + \frac{1}{4}(x_2)^4 - \frac{1}{2}(x_2)^2$. One can easily see that $f'(x) = (x_1; x_2^3 - x_2)$ and $f''(x) = \mathrm{Diag}(1, 3x_2^2 - 1)$. Hence $(0,0)$ is a stationary point, and $(0,-1), (0,1)$ are local minimum. Starting from $(1,0)$, any point along the trajectory will have $x_2 = 0$. Hence the sequence must converge to a stationary point.

In the sequel, we study the local convergence behavior of GD. To this end, we make the following assumptions: (a) $f \in C_M^{2,2}(\mathbb{R}^n)$, (b) $\exists$ nondegernate local minimum x^* ($f''(x^*) \succ 0$), and (c) $\exists 0 < l \leq L < \infty$: $lI_n \preceq f''(x^*) \preceq LI_n$. Moreover, we assume that x_0 close enough to x^*. We first state a simple technical result.

Lemma 5.4. *If $p_{k+1} \leq (1-q)p_k + p_k^2$, then $\frac{q}{p_{k+1}} - 1 \geq (1+q)(\frac{q}{p_k} - 1)$.*

Proof. Noting that

$$p_{k+1} \leq (1-q)p_k + p_k^2 = p_k[1+(p_k-q)] = \frac{p_k[1-(p_k-q)^2]}{1-(p_k-q)} \leq \frac{p_k}{1+q-p_k},$$

we have $\frac{1}{p_{k+1}} \geq \frac{1+q}{p_k} - 1$ or $\frac{q}{p_{k+1}} - 1 \geq \frac{q(1+q)}{p_k} - q - 1 = (1+q)(\frac{q}{p_k} - 1)$. ■

We now establish the locally linear rate of convergence for GD.

Theorem 5.2. *Let x_0 be close enough to a local minimum, i.e., $r_0 = \|x_0 - x^*\| < \bar{r} = \frac{2l}{M}$. Then the GD method satisfies*

$$\|x_k - x^*\| \leq \frac{2lr_0}{M(\bar{r}-r_0)}(\frac{1}{1+q})^k,$$

where $q = 2l/(L+l)$.

Proof. Denote $G_k = \int_0^1 f''(x^* + \tau(x_k - x^*))d\tau$. We have $f'(x_k) = f'(x_k) - f'(x^*) = G_k(x_k - x^*)$. Hence,

$$x_{k+1} - x^* = x_k - x^* - h_k G_k(x_k - x^*) = (I - h_k G_k)(x_k - x^*),$$

which implies that $\|x_{k+1} - x^*\| \leq \|I - h_k G_k\| \|x_k - x^*\|$. We intend to bound $\|I - h_k G_k\|$ Denote $r_k = \|x_k - x^*\|$. It follows from Corollary 5.1 that

$$f''(x^*) - \tau M r_k I_n \preceq f''(x^* + \tau(x_k - x^*)) \preceq f''(x^*) + \tau M r_k I_n,$$

implying that $(l - \frac{r_k}{2}M)I_n \preceq G_k \preceq (L + \frac{r_k}{2}M)I_n$ and hence that

$$(1 - h_k(L + \tfrac{r_k}{2}M))I_n \preceq I_n - h_k G_k \preceq (1 - h_k(l - \tfrac{r_k}{2}M))I_n.$$

Therefore, we obtain $\|I_n - h_k G_k\| \leq \max\{a_k(h_k), b_k(h_k)\}$, with

$$a_k(h) := 1 - h(l - \tfrac{r_k}{2}M), b_k(h) := h(L + \tfrac{r_k}{2}M) - 1.$$

Observe that if $r_k \leq \bar{r} \equiv \frac{2l}{M}$, we can ensure $\|I_k - h_k G_k\| \leq 1$ for small enough h_k. Moreover, an “Optimal” selection of h_k is given by

$$\min_h \max\{a_k(h), b_k(h)\}.$$

In particular, setting $a_k(h) = b_k(h)$, we obtain the optimal selection $h_k^* = \frac{2}{L+l}$. Hence, we have

$$a_k(h_k^*) = 1 - h_k^*(l - \tfrac{r_k}{2}M) = \tfrac{L-l}{L+l} + \tfrac{r_k M}{L+l}.$$

which, in view of $r_{k+1} \le a_k(h_k^*) r_k$, implies $r_{k+1} \le \frac{L-l}{L+l} r_k + \frac{M}{L+l} r_k^2$ or equivalently,

$$\tfrac{M r_{k+1}}{L+l} \le \tfrac{L-l}{L+l} \tfrac{M r_k}{L+l} + (\tfrac{M r_k}{L+l})^2.$$

Applying Lemma 5.4 with $p_k = \frac{M r_k}{L+l}$ and $q = \frac{2l}{L+l}$, we have

$$\tfrac{q}{p_k} - 1 \ge (1+q)^k (\tfrac{q}{p_0} - 1) = (1+q)^k (\tfrac{2l}{M r_0} - 1),$$

which, in view of $\bar{r} = 2l/M$, implies that

$$p_k \le \frac{q}{1+(1+q)^k(\frac{2l}{M r_0}-1)} \le \tfrac{q r_0}{\bar{r} - r_0} (\tfrac{1}{1+q})^k.$$

■

Gradient Descent for Convex Problems

We now consider the case when $f \in C_L^{1,1}(\mathbb{R}^n)$ is convex, and denote this class of functions by $\mathscr{F}_L^{1,1}(\mathbb{R}^n)$. We also consider a restriction of $\mathscr{F}_L^{1,1}(\mathbb{R}^n)$, strongly convex functions denoted by $\mathscr{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$ s.t. $\exists \mu > 0$ s.t. for any $x, y \in X$, $f(y) \ge f(x) + \langle f'(x), y - x \rangle + \frac{1}{2}\mu \|y - x\|^2$.

Lemma 5.5. *Let $f \in \mathscr{F}_L^{1,1}(\mathbb{R}^n)$. For any $x, y \in \mathbb{R}^n$, we have*

$$\begin{aligned}
&(a) f(x) + \langle f'(x), y - x \rangle + \tfrac{1}{2L}\|f'(x) - f'(y)\|^2 \le f(y).\\
&(b) \tfrac{1}{L}\|f'(x) - f'(y)\|^2 \le \langle f'(x) - f'(y), x - y \rangle.
\end{aligned}$$

Proof. Let $\phi(y) = f(y) - \langle f'(x), y \rangle$ for a given $x \in \mathbb{R}^n$. Clearly $\phi \in \mathscr{F}_L^{1,1}(\mathbb{R}^n)$, $\phi'(x) = 0$, and hence x is a minimizer of $\phi(y)$.

$$\begin{aligned}
\phi(x) &\le \phi(y - \tfrac{1}{L}\phi'(y))\\
&\le \phi(y) + \langle \phi'(y), -\tfrac{1}{L}\phi'(y) \rangle + \tfrac{L}{2}\|\tfrac{1}{L}\phi'(y)\|^2\\
&= \phi(y) - \tfrac{1}{2L}\|\phi'(y)\|^2.
\end{aligned}$$

Therefore, we have $f(x) - \langle f'(x), x \rangle \le f(y) - \langle f'(x), y \rangle - \frac{1}{2L}\|f'(y) - f'(x)\|^2$, implying that

$$f(x) + \langle f'(x), y - x \rangle + \tfrac{1}{2L}\|f'(y) - f'(x)\|^2 \le f(y).$$

Similarly,

$$f(y) + \langle f'(y), x - y \rangle + \tfrac{1}{2L}\|f'(y) - f'(x)\|^2 \le f(x).$$

Adding up these two inequalities, we obtain the result in (a). (b) follows from (a) by adding two inequalities with x and y interchanged. ■

We analyze the simplest variant of GD with $h_k = h > 0$. Denote by x^* the optimal solution of our problem and $f^* = f(x^*)$.

Lemma 5.6. *The sequence generated by the gradient descent method with $h \in (0, \frac{2}{L})$ satisfies $f(x_{k+1}) \le f(x_k)$.*

Proof. We have

$$\begin{aligned}
&f(x_{k+1}) \le f(x_k) + \langle f'(x_k), x_{k+1} - x_k\rangle + \tfrac{L}{2}\|x_{k+1} - x_k\|^2\\
&\le f(x_k) - h\|f'(x_k)\|^2 + \tfrac{L}{2}h^2\|f'(x_k)\|^2\\
&\le f(x_k) - h(1 - \tfrac{L}{2}h)\|f'(x_k)\|^2 \le f(x_k).
\end{aligned}$$

■

Theorem 5.3. *Let $f \in \mathscr{F}_L^{1,1}(\mathbb{R}^n)$ and $h \in (0, \frac{2}{L}]$. Then the sequence $\{x_k\}$ generated by the gradient descent method satisfies*

$$f(x_k) - f^* \le \tfrac{\|x_0 - x^*\|^2}{kh(2-Lh)}.$$

Proof. Denote $r_k = \|x_k - x^*\|$. Then

$$\begin{aligned}
&r_{k+1}^2 = \|x_k - x^* - hf'(x_k)\|^2 = r_k^2 - 2h\langle f'(x_k), x_k - x^*\rangle + h^2\|f'(x_k)\|^2\\
&\le r_k^2 - 2h\langle f'(x_k), x_k - x^*\rangle + Lh^2\langle f'(x_k), x_k - x^*\rangle\\
&= r_k^2 - h(2 - Lh)\langle f'(x_k), x_k - x^*\rangle.
\end{aligned}$$

Noting that $f(x_k) - f^* \le \langle f'(x_k), x_k - x^*\rangle$, we have

$$h(2 - Lh)[f(x_k) - f^*] \le r_k^2 - r_{k+1}^2.$$

Summing up these inequalities, we have $h(2-Lh)\sum_{t=1}^k [f(x_t) - f^*] \le r_0^2 - r_{k+1}^2$. Using $f(x_0) \ge f(x_1) \ge \ldots \ge f(x_k)$, we have

$$h(2 - Lh)k[f(x_k) - f^*] \le r_0^2 - r_{k+1}^2.$$

■

Note that the optimal selection of stepsize is given by of $h = 1/L$.

We now study the convergence of GD for strongly convex problems.

Lemma 5.7. *If $f \in S_{\mu,L}^{1,1}(\mathbb{R}^n)$, then for any $x, y \in \mathbb{R}^n$, we have*

$$\langle f'(y) - f'(x), y - x\rangle \ge \tfrac{\mu L}{\mu + L}\|x - y\|^2 + \tfrac{1}{\mu + L}\|f'(x) - f'(y)\|^2.$$

Proof. Denote $\phi(x) = f(x) - \frac{1}{2}\mu\|x\|^2$. Then $\phi'(x) = f'(x) - \mu x$. Hence $\phi(x) \in \mathscr{F}_{L-\mu}^{1,1}(\mathbb{R}^n)$ (It can be easily checked that $\langle \phi'(x) - \phi'(y), x - y\rangle \le (L - \mu)\|x - y\|^2$). If $\mu = L$, then the result follows from the facts that

$$\langle f'(y) - f'(x), y - x\rangle \ge \mu\|x - y\|^2$$

and

$$\langle f'(y)-f'(x), y-x\rangle \geq \tfrac{1}{L}\|f'(x)-f'(y)\|^2.$$

If $\mu < L$, we have

$$\langle \phi'(y)-\phi'(x), y-x\rangle \geq \tfrac{1}{L-\mu}\|\phi'(y)-\phi'(x)\|^2$$

which is exactly the result. ■

Theorem 5.4. *If $f \in \mathscr{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$ and $h \in (0, \frac{2}{\mu+L}]$, then the gradient descent method generates a sequence $\{x_k\}$ s.t.*

$$\|x_k - x^*\|^2 \leq (1-\tfrac{2h\mu L}{\mu+L})^k \|x_0 - x^*\|^2.$$

If $h = \frac{2}{\mu+L}$, then

$$\|x_k - x^*\| \leq (\tfrac{Q_f-1}{Q_f+1})^k \|x_0 - x^*\|, \tag{5.3.4}$$

$$f(x_k) - f^* \leq \tfrac{L}{2}(\tfrac{Q_f-1}{Q_f+1})^{2k} \|x_0 - x^*\|^2, \tag{5.3.5}$$

where $Q_f = L/\mu$.

Proof. Denote $r_k = \|x_k - x^*\|$. Then

$$\begin{aligned} r_{k+1}^2 &= \|x_k - x^* - hf'(x_k)\|^2 = r_k^2 - 2h\langle f'(x_k), x_k - x^*\rangle + h^2\|f'(x_k)\|^2 \\ &= (1-\tfrac{2h\mu L}{\mu+L})r_k^2 + h(h-\tfrac{2}{L+\mu})\|f'(x_k)\|^2. \end{aligned}$$

■

We now derive the complexity of GD for strongly convex problems. Using the fact $\ln(1-t) \leq t$ for $t \in (0,1)$,

$$f(x_k) - f^* \leq \tfrac{L}{2}(1-\tfrac{2}{Q_f+1})^{2k}\|x_0-x^*\|^2 \leq \tfrac{L}{2}\exp(-\tfrac{4k}{Q_f+1})\|x_0-x^*\|^2.$$

If

$$k \geq \tfrac{Q_f+1}{4}\log\tfrac{L\|x_0-x^*\|^2}{2\varepsilon},$$

then $f(x_k) - f^* \leq \varepsilon$.

5.4 Conjugate Gradient Method

Conjugate Gradient Method was initially designed for quadratic programming:

$$\min f(x) = \tfrac{1}{2}x^T Ax + b^T x, \tag{5.4.1}$$

where $A = A^T \succ 0$. Clearly, the optimal solution is $x^* = -A^{-1}b$ and hence

$$f(x) = \frac{1}{2}x^T Ax - (x^*)^T Ax = \frac{1}{2}(x-x^*)^T A(x-x^*) - \frac{1}{2}(x^*)^T Ax^*,$$
$$f'(x) = A(x-x^*), f^* = -\frac{1}{2}\langle Ax^*, x^*\rangle.$$

Given a starting point x_0, consider the linear Krylov subspaces

$$L_k = \text{lin}\{A(x_0 - x^*), \ldots, A^k(x_0 - x^*)\}, k \geq 1.$$

Observe that $L_1 \subseteq L_2 \subseteq L_3 \subseteq \ldots$. Moreover, let t be the largest value of k such that $A(x_0 - x^*), \ldots, A^t(x_0 - x^*)$ are linearly independent, then

$$L_k \subset L_{k+1}, \forall k < t-1$$
$$L_k = L_{k+1}, \forall k \geq t.$$

The classic CG method computes the iterate x_k according to

$$x_k = \text{argmin}_{x \in x_0 + L_k} f(x), k \geq 1.$$

Note that this is for theoretical analysis only, and we will see a more algorithmic form later.

Lemma 5.8. *If the algorithm does not terminates, for any $k \geq 1$ we have*

$$L_k = \text{lin}\{f'(x_0), \ldots, f'(x_{k-1})\}.$$

Moreover, for any $k, i \geq 0$, $k \neq i$, we have

$$f'(x_k)^T f'(x_i) = 0.$$

Proof. For $k = 1$, the statements are true since $f'(x_0) = A(x_0 - x^*)$. Suppose they are true for some $k \geq 1$. Since $x_k \in x_0 + L_k$, $x_k = x_0 + \sum_{i=1}^k \lambda_i A^i (x_0 - x^*)$ for some $\lambda_i \in \mathbb{R}$. Therefore,

$$\begin{aligned} f'(x_k) &= A(x_k - x^*) = A(x_0 - x^*) + \textstyle\sum_{i=1}^k \lambda_i A^{i+1}(x_0 - x^*) \\ &= \underbrace{A(x_0 - x^*) + \textstyle\sum_{i=1}^{k-1} \lambda_i A^{i+1}(x_0 - x^*)}_{y \in L_k} + \lambda_k A^{k+1}(x_0 - x^*) \\ &= y + \lambda_k A^{k+1}(x_0 - x^*). \end{aligned}$$

Thus,

$$L_{k+1} = \text{lin}\{L_k, A^{k+1}(x_0 - x^*)\} \supseteq \text{lin}\{L_k, f'(x_k)\} = \text{lin}\{f'(x_0), \ldots, f'(x_k)\}. \tag{5.4.2}$$

Now let $k > i$ consider the function $\phi(x) = f(x_0 + \sum_{j=1}^k \mu_j f'(x_{j-1})$. By inductive hypothesis ($L_k = \text{lin}\{f'(x_0), \ldots, f'(x_{k-1})\}$), we have

$$x_k = x_0 + \textstyle\sum_{j=1}^k \mu_j^* f'(x_{j-1}).$$

Moreover, by definition, $x_k = \text{argmin}_{x \in x_0 + L_k} f(x)$. Therefore $\phi'(\mu^*) = 0$ and

$$\frac{\partial\phi(\mu^*)}{\partial\mu_i} = f'(x_i)^T f'(x_k) = 0.$$

Hence the dimension of $\{L_k, f'(x_k)\}$ is $k+1$, which implies that (5.4.2) holds with equality. ■

Corollary 5.2. *The sequence generated by the CG method for QP is finite.*

Proof. The number of orthogonal directions in $\mathbb{R}^n$ cannot exceed n. ■

Lemma 5.9. *Denote $\delta_i \equiv x_{i+1} - x_i$, $i \geq 0$. If the algorithm does not terminate, then*

$$L_k = \text{lin}\{\delta_0, \dots, \delta_{k-1}\}$$
$$\langle A\delta_k, \delta_k\rangle = 0, \forall i < k.$$

Proof. By definition of x_1, we have $x_1 - x_0 = \mu_0 f'(x_0)$ for some $\mu_0 \in \mathbb{R}$. Now assume the results hold for some $k \geq 0$. Note that by definition of x_{k+1}, we have $x_{k+1} - x_0 \in L_{k+1}$. Since

$$\delta_k = x_{k+1} - x_k = x_{k+1} - x_0 - (x_k - x_0) = x_{k+1} - x_0 + \textstyle\sum_{i=0}^{k-1}\delta_i$$

and $\delta_i \in L_k \subseteq L_{k+1}$ for $i \leq k-1$ by inductive hypothesis. We must have $\delta_k \in L_{k+1}$ and

$$\text{lin}\{\delta_0, \dots, \delta_k\} \subseteq L_{k+1}. \tag{5.4.3}$$

Note that, since $\delta_i = x_{i+1} - x_i \in L_{i+1} \subseteq L_k$,

$$\langle A\delta_k, \delta_i\rangle = \langle A(x_{k+1} - x_k), \delta_i\rangle$$
$$= \langle f'(x_{k+1}) - f'(x_k), \delta_i\rangle = 0,$$

Therefore the vectors $\delta_0, \dots, \delta_k$ are A-orthogonal and thus are linearly independent [1], which implies (5.4.3) holds with equality.

■

We now state the conjugate gradient method in an algorithmic form. Let $L_{k+1} = \text{lin}\{L_k, f'(x_k)\}$ and $L_k = \text{lin}\{\delta_0, \dots, \delta_{k-1}\}$ nad $\delta_k = x_{k+1} - x_k \in L_{k+1}$. We can write x_{k+1} as

$$x_{k+1} = x_k - h_k f'(x_k) + \textstyle\sum_{j=0}^{k-1}\lambda_j\delta_j$$

or

$$\delta_k = -h_k f'(x_k) + \textstyle\sum_{j=0}^{k-1}\lambda_j\delta_j. \tag{5.4.4}$$

To compute $h_k, \lambda_0, \dots, \lambda_{k-1}$, multiplying (5.4.4) by A and δ_i, $0 \leq i \leq k-1$, we have

$$\begin{aligned} 0 = \langle A\delta_k, \delta_i\rangle &= -h_k\langle Af'(x_k), \delta_i\rangle + \textstyle\sum_{j=0}^{k-1}\lambda_j\langle A\delta_j, \delta_i\rangle \\ &= -h_k\langle Af'(x_k), \delta_i\rangle + \lambda_i\langle A\delta_i, \delta_i\rangle \\ &= -h_k\langle f'(x_k), A\delta_i\rangle + \lambda_i\langle A\delta_i, \delta_i\rangle \\ &= -h_k\langle f'(x_k), f'(x_{i+1}) - f'(x_i)\rangle + \lambda_i\langle A\delta_i, \delta_i\rangle. \end{aligned}$$

[1] We need to show $\sum_{i=1}^k \lambda_i\delta_i = 0 \Rightarrow \lambda_i = 0$. Indeed, multiplying these equations by $\delta_i^T A$ and using $\delta_i^T A\delta_j = 0, \forall i \neq j$, we see that $\sum_{i=1}^k \lambda_i\delta_i^T A\delta_i = 0$, implying $\lambda_i = 0$ due to $A \succ 0$.

Hence, for $i < k-1$, we must have $\lambda_i = 0$. For $i = k-1$, we have

$$\lambda_{k-1} = \frac{h_k \|f'(x_k)\|^2}{\langle A\delta_{k-1}, \delta_{k-1}} = \frac{h_k \|f'(x_k)\|^2}{\langle f'(x_k) - f'(x_{k-1}), \delta_{k-1}\rangle}.$$

Thus $x_{k+1} = x_k - h_k p_k$ (i.e., $\delta_k = -h_k p_k$), where

$$\begin{aligned} p_k &= f'(x_k) - \frac{\|f'(x_k)\|^2 \delta_{k-1}}{\langle f'(x_k) - f'(x_{k-1}, \delta_{k-1}\rangle} \\ &= f'(x_k) - \frac{\|f'(x_k)\|^2 p_{k-1}}{\langle f'(x_k) - f'(x_{k-1}, p_{k-1}\rangle} \end{aligned}$$

and h_k can be obtained using line search.

One remaining question about the conjugate gradient method is rate of convergence. Will it be faster than gradient descent method? This problem has been well-under stood for the convex quadratic case. We state this result without proof and direct interesting readers to the original development by Nemirovksi and Yudin. However, we will show a similar result obtained by the accelerated gradient method later.

Theorem 5.5. *One has*

$$f(x_k) - \min_x f(x) \le 4 \left[\frac{\sqrt{Q_f}-1}{\sqrt{Q_f}+1}\right]^{2k} [f(x_0) - \min_x f(x)], \tag{5.4.5}$$

where Q_f is the condition number of f, i.e., the ratio the ratio of the largest and the smallest eigenvalues of A.

In view of the above result, every $\sqrt{Q_f}$ new iterations decrease the right hand side in (5.4.5) by absolute constant factor. For Steepest decent similar improvement requires Q_f new iterations.

We now generalize the conjugate Gradient method for general NLP as follows. Let $x_0 \in \mathbb{R}^n$. Compute $f(x_0)$, $f'(x_0)$. Set $p_0 = f'(x_0)$. In the k^{th} iteration ($k \ge 0$), we perform the following steps.

1. Find $x_{k+1} = x_k - h_k p_k$ (h_k by line search).
2. Compute $f(x_{k+1})$ and $f'(x_{k+1})$.
3. Compute the coefficient β_k.
4. Set $p_{k+1} = f'(x_{k+1}) + \beta_k p_k$.

We have not specified β_k. In fact, there are many different formulas.

$$\beta_k = \frac{\|f'(x_{k+1})\|^2}{\langle f'(x_{k+1}) - f'(x_k), p_k\rangle}, \tag{5.4.6}$$

$$\beta_k = -\frac{\|f'(x_{k+1}\|^2}{\|f'(x_k)\|^2}, \tag{5.4.7}$$

$$\beta_k = -\frac{\langle f'(x_{k+1}, f'(x_{k+1}) - f'(x_k)\rangle}{\|f'(x_k)\|^2}. \tag{5.4.8}$$

All of them give the same result on quadratic functions. But in a general NLP they generate different results. Recall in the quadratic case, CG terminates in n iterations or less ($p_{n+1} = 0$). However, in a nonlinear case that is not true. After n iterations, this direction lose any interpretation. In practice, there exists a restarting strategy, usually after every n iterations set $\beta_k = 0$. This ensures global convergence (since we have a usual gradient descent step just after the restart and all other iterations decrease the function value). In a neighborhood of a strict minimum, CG has a local n-step quadratic convergence

$$\|x_{n+1} - x^*\| \leq C\|x_0 - x^*\|^2$$

5.5 Projected gradient method

Gradient descent applies only to unconstrained problems. For the constrained case when $X \neq \mathbb{R}^n$, the search point x_{t+1} defined by $x_{t+1} = x_t - \gamma_t \nabla f(x_t), t = 1,2,\ldots$, may fall outside the feasible set X. Hence, it is necessary to "push" x_{t+1} back to X by using projection. Incorporating these enhancements, we update x_t according to

$$x_{t+1} := \mathrm{argmin}_{x\in X} \|x - (x_t - \gamma_t g(x_t))\|_2, t = 1,2,\ldots, \tag{5.5.1}$$

for some $g(x_t) \in \nabla f(x_t)$ and $\gamma_t > 0$.

The projected gradient iteration in (5.5.1) admits some natural explanation from the proximity control point of view. Indeed, (5.5.1) can be written equivalently as

$$\begin{aligned} x_{t+1} &= \mathrm{argmin}_{y\in X} \tfrac{1}{2}\|x - (x_t - \gamma_t g(x_t))\|_2^2 \\ &= \mathrm{argmin}_{x\in X} \gamma_t \langle g(x_t), x - x_t\rangle + \tfrac{1}{2}\|x - x_t\|_2^2 \\ &= \mathrm{argmin}_{x\in X} \gamma_t \left[f(x_t) + \langle g(x_t), x - x_t\rangle\right] + \tfrac{1}{2}\|x - x_t\|_2^2 \\ &= \mathrm{argmin}_{x\in X} \gamma_t \langle g(x_t), x\rangle + \tfrac{1}{2}\|x - x_t\|_2^2. \end{aligned} \tag{5.5.2}$$

This implies that we would like to minimize the linear approximation $f(x_t) + \langle g(x_t), x - x_t\rangle$ of $f(x)$ over X, without moving too far away from x_t so as to have $\|x - x_t\|_2^2$ small. The parameter $\gamma_t > 0$ balances these two terms, and its selection will depend on the properties of the objective function f.

The following lemma provides an important characterization for x_{t+1} by using the representation in (5.5.2).

Lemma 5.10. *Let x_{t+1} be defined in (5.5.1). For any $y \in X$, we have*

$$\gamma_t \langle g(x_t), x_{t+1} - x\rangle + \tfrac{1}{2}\|x_{t+1} - x_t\|_2^2 \leq \tfrac{1}{2}\|x - x_t\|_2^2 - \tfrac{1}{2}\|x - x_{t+1}\|_2^2.$$

Proof. Denote $\phi(x) = \gamma_t \langle g(x_t), x\rangle + \frac{1}{2}\|x - x_t\|_2^2$. By the strong convexity of ϕ, we have

$$\phi(x) \geq \phi(x_{t+1}) + \langle \phi'(x_{t+1}), x - x_{t+1}\rangle + \tfrac{1}{2}\|x - x_{t+1}\|_2^2.$$

Moreover, by the first-order optimality condition of (5.5.2), we have $\langle \phi'(x_{t+1}), x - x_{t+1}\rangle \geq 0$ for any $x \in X$. The result immediately follows by combining these two inequalities. ■

Our next result shows that the function values at the iterates x_t, $t \geq 1$, are monotonically non-increasing.

Lemma 5.11. *Let $\{x_t\}$ be generated by (5.5.1). If*

$$\gamma_t \leq \tfrac{2}{L}, \tag{5.5.3}$$

then

$$f(x_{t+1}) \leq f(x_t),\ \forall t \geq 1.$$

Proof. By the optimality condition of (5.5.1), we have

$$\langle \gamma_t g(x_t) + x_{t+1} - x_t, x - x_{t+1}\rangle \geq 0,\ \forall x \in X.$$

Letting $x = x_t$ in the above relation, we obtain

$$\gamma_t \langle g(x_t), x_{t+1} - x_t\rangle \leq -\|x_{t+1} - x_t\|_2^2. \tag{5.5.4}$$

It then follows from the smoothness of f and the above relation that

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) + \langle g(x_t), x_{t+1} - x_t\rangle + \tfrac{L}{2}\|x_{t+1} - x_t\|_2^2 \\ &\leq f(x_t) - \left(\tfrac{1}{\gamma_t} - \tfrac{L}{2}\right)\|x_{t+1} - x_t\|_2^2 \leq f(x_t). \end{aligned}$$

■

We are now ready to establish the main convergence properties for the projected gradient method applied to smooth convex optimization problems.

Theorem 5.6. *Let $\{x_t\}$ be generated by (5.5.1). If*

$$\gamma_t = \gamma \leq \tfrac{1}{L}, \forall t \geq 1, \tag{5.5.5}$$

then

$$f(x_{k+1}) - f(x) \leq \tfrac{1}{2\gamma k}\|x - x_1\|_2^2,\ \forall x \in X.$$

Proof. By the smoothness of f, we have

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) + \langle g(x_t), x_{t+1} - x_t\rangle + \tfrac{L}{2}\|x_{t+1} - x_t\|_2^2 \\ &\leq f(x_t) + \langle g(x_t), x - x_t\rangle + \langle g(x_t), x_{t+1} - x\rangle + \tfrac{L}{2}\|x_{t+1} - x_t\|_2^2. \end{aligned} \tag{5.5.6}$$

It then follows from the above inequality, the convexity of f and Lemma 5.10 that

$$\begin{aligned} f(x_{t+1}) &\leq f(x) + \tfrac{1}{2\gamma_t}\left(\|x - x_t\|_2^2 - \|x - x_{t+1}\|_2^2 - \|x_t - x_{t+1}\|_2^2\right) + \tfrac{L}{2}\|x_{t+1} - x_t\|_2^2 \\ &\leq f(x) + \tfrac{1}{2\gamma}\left(\|x - x_t\|_2^2 - \|x - x_{t+1}\|_2^2\right), \end{aligned}$$

where the last inequality follows from (5.5.5). Summing up the above inequalities from $t = 1$ to k, and using Lemma 5.11, we have

$$k[f(x_{k+1}) - f(x)] \le \sum_{t=1}^{k} [f(x_{t+1}) - f(x)] \le \tfrac{1}{2\gamma}\|x - x_1\|_2^2.$$

■

In view of Theorem 5.6, one may choose $\gamma = 1/L$ and then the rate of convergence of the projected gradient method becomes $f(x_{k+1}) - f^* \le L/(2k)$.

We now discuss the convergence properties of the projected gradient method when the objective function f is smooth and strongly convex.

Theorem 5.7. *Let $\{x_t\}$ be generated by (5.5.1). Suppose that f is smooth with L-Lipschitz gradients and strongly convex with modulus μ. If $\gamma_t = \gamma = 1/L$, $t = 1, \ldots, k$, then*

$$\|x - x_{k+1}\|_2^2 \le (1 - \tfrac{\mu}{L})^k \|x - x_1\|_2^2. \tag{5.5.7}$$

Proof. It follows from (5.5.6), the strong convexity of f and Lemma 5.10 that

$$\begin{aligned} f(x_{t+1}) &\le f(x) - \tfrac{\mu}{2}\|x - x_t\|_2^2 + \tfrac{1}{2\gamma_t}\left(\|x - x_t\|_2^2 - \|x - x_{t+1}\|_2^2 - \|x_t - x_{t+1}\|_2^2\right) \\ &\quad + \tfrac{L}{2}\|x_{t+1} - x_t\|_2^2 \\ &\le f(x) + \tfrac{1-\mu\gamma}{2\gamma}\|x - x_t\|_2^2 - \tfrac{1}{2\gamma}\|x - x_{t+1}\|_2^2. \end{aligned}$$

Using the above relation, the facts $\gamma = 1/L$ and $f(x_t) - f(x^*) \ge 0$, we have

$$\|x_{t+1} - x^*\|_2^2 \le (1 - \tfrac{\mu}{L})\|x_t - x^*\|_2^2,$$

which clearly implies (5.5.7). ■

In order to find a solution $\bar{x} \in X$ such that $\|\bar{x} - x^*\|^2 \le \varepsilon$, it suffices to have

$$\begin{aligned} (1 - \tfrac{\mu}{L})^k \|x - x_1\|_2^2 \le \varepsilon &\Longleftrightarrow k \log(1 - \tfrac{\mu}{L}) \le \log \tfrac{\varepsilon}{\|x - x_1\|_2^2} \\ &\Longleftrightarrow k \ge \tfrac{1}{-\log(1 - \frac{\mu}{L})} \log \tfrac{\|x - x_1\|_2^2}{\varepsilon} \\ &\Longleftarrow k \ge \tfrac{L}{\mu} \log \tfrac{\|x - x_1\|_2^2}{\varepsilon}, \end{aligned} \tag{5.5.8}$$

where the last inequality follows from the fact that $-\log(1 - \alpha) \ge \alpha$ for any $\alpha \in [0, 1)$.

We can also extend the analysis of the projected gradient method to the nonconvex setting. For that purpose, we need to define a termination criterion: $P_X(x, g, \gamma) = \frac{1}{\gamma}(x - x^+)$, where $x^+ = \arg\min_{u \in X} \left\{ \langle g, u \rangle + \frac{1}{\gamma} V(u, x) + h(u) \right\}$. It is not difficult to show that one can find a point $\bar{x} \in X$ s.t. $\|P_X(\bar{x}, \nabla f(\bar{x}), \gamma)\|^2 \le \varepsilon$ in $\mathcal{O}(1/\varepsilon)$ iterations with $\gamma \in (0, 1/L]$.

5.6 Accelerated Gradient Descent Method

In this section, we consider the problem of $\min_{x \in X} f(x)$, where $X \subseteq \mathbb{R}^n$ is closed and convex, and $f \in \mathscr{F}_L^{1,1}(\mathbb{R}^n)$. The basic version of the accelerated gradient descent (AGD) method is stated as follows.

0) Choose $\bar{x}_0 = x_0 \in X$.
1) Set $\underline{x}_k = (1-\alpha_k)\bar{x}_{k-1} + \alpha_k x_{k-1}$.
2) Compute $f'(\underline{x}_k)$ and set
$x_k = \operatorname{argmin}_{x \in X}\{\alpha_k \langle f'(\underline{x}_k), x\rangle$
$\quad + \frac{\beta_k}{2}\|x - x_{k-1}\|^2\}$,
$\bar{x}_k = (1-\alpha_k)\bar{x}_{k-1} + \alpha_k x_k$.
3) Set $k \leftarrow k+1$ and go to step 1).

Note that we have not specified α_k and β_k yet. We will show how to select them after establishing some general convergence properties of AGD.

Lemma 5.12. *Let $v : X \to \mathbb{R}$ be a convex function, $c > 0$ and $\bar{x} \in X$ be given. Also let $\hat{x} = \operatorname{argmin}_{x \in X}\{v(x) + \frac{c}{2}\|x - \bar{x}\|^2\}$. Then for any $x \in X$,*

$$v(\hat{x}) + \tfrac{c}{2}\|\hat{x} - \bar{x}\|^2 \le v(x) + \tfrac{c}{2}\|x - \bar{x}\|^2 - \tfrac{c}{2}\|x - \hat{x}\|^2.$$

Proof. Define $\phi(x) = v(x) + \frac{c}{2}\|x - \bar{x}\|^2$. ϕ is strongly convex with modulus $\ge c$. Hence,

$$\begin{aligned}\phi(x) &\ge \phi(\hat{x}) + \langle \phi'(\hat{x}), x - \hat{x}\rangle + \tfrac{c}{2}\|x - \hat{x}\|^2\\ &\ge \phi(\hat{x}) + \tfrac{c}{2}\|x - \hat{x}\|^2,\end{aligned}$$

where the last inequality follows from $\langle \phi'(\hat{x}), x - \hat{x}\rangle \ge 0, \forall x \in X$. ■

Lemma 5.13. *Let $(\underline{x}_k, x_k, \bar{x}_k)$ be generated by the AGD method. If $\beta_k \ge L\alpha_k^2$, then $\forall x \in X$,*

$$f(\bar{x}_k) - f(x) + \tfrac{\beta_k}{2}\|x - x_k\|^2 \le (1-\alpha_k)[f(\bar{x}_{k-1}) - f(x)] + \tfrac{\beta_k}{2}\|x - x_{k-1}\|^2.$$

Proof. Note that $\bar{x}_k - \underline{x}_k = \alpha_k(x_k - x_{k-1})$.

$$\begin{aligned}&f(\bar{x}_k) \le f(\underline{x}_k) + \langle f'(\underline{x}_k), \bar{x}_k - \underline{x}_k\rangle + \tfrac{L}{2}\|\bar{x}_k - \underline{x}_k\|^2\\ &= (1-\alpha_k)[f(\underline{x}_k) + \langle f'(\underline{x}_k), \bar{x}_{k-1} - \underline{x}_k\rangle] + \alpha_k[f(\underline{x}_k) + \langle f'(\underline{x}_k), x_k - \underline{x}_k\rangle] + \tfrac{L\alpha_k^2}{2}\|x_k - x_{k-1}\|^2\\ &\le (1-\alpha_k)[f(\underline{x}_k) + \langle f'(\underline{x}_k), \bar{x}_{k-1} - \underline{x}_k\rangle] + \alpha_k[f(\underline{x}_k) + \langle f'(\underline{x}_k), x_k - \underline{x}_k\rangle] + \tfrac{\beta_k}{2}\|x_k - x_{k-1}\|^2.\end{aligned}$$

Noting that

$$\alpha_k\langle f'(\underline{x}_k), x_k - \underline{x}_k\rangle + \tfrac{\beta_k}{2}\|x_k - x_{k-1}\|^2 \le \alpha_k\langle f'(\underline{x}_k), x - \underline{x}_k\rangle + \tfrac{\beta_k}{2}\|x - x_k\|^2 - \tfrac{\beta_k}{2}\|x - x_k\|^2,$$

we obtain

$$f(\bar{x}_k) \le (1-\alpha_k)f(\bar{x}_{k-1}) + \alpha_k[f(\underline{x}_k) + \langle f'(\underline{x}_k), x - \underline{x}_k\rangle] + \tfrac{\beta_k}{2}\|x - x_{k-1}\|^2 - \tfrac{\beta_k}{2}\|x - x_k\|^2.$$

Subtracting $f(x)$ from both sides and re-arranging the terms, we obtain the result. ∎

We are now ready to establish the convergence of the AGD method.

Theorem 5.8. *If $\beta_k \geq L\alpha_k^2$ and $\beta_k = (1-\alpha_k)\beta_{k-1}$ for $k \geq 1$, then*

$$\begin{aligned} f(\bar{x}_k) - f(x) \leq & \tfrac{1-\alpha_1}{\beta_1}[f(\bar{x}_0) - f(x)] \\ & + \tfrac{\beta_k}{2}\|x - x_0\|^2, \ \forall x \in X. \end{aligned}$$

Proof. By Lemma 5.13,

$$\tfrac{1}{\beta_k}[f(\bar{x}_k) - f(x)] + \tfrac{1}{2}\|x - x_k\|^2 \leq \tfrac{1}{\beta_{k-1}}[f(\bar{x}_{k-1}) - f(x)] + \tfrac{1}{2}\|x - x_{k-1}\|^2$$

for any $k \geq 2$. Moreover, for $k = 1$, we have

$$\tfrac{1}{\beta_1}[f(\bar{x}_1) - f(x)] + \tfrac{1}{2}\|x - x_1\|^2 \leq \tfrac{1-\alpha_1}{\beta_1}[f(\bar{x}_0) - f(x)] + \tfrac{1}{2}\|x - x_0\|^2.$$

The result follows by summing up these inequalities. ∎

The following corollary shows the complexity of AGD

Corollary 5.3. *If $\alpha_k = 2/(k+1)$ and $\beta_k = 4L/[k(k+1)]$, then*

$$f(\bar{x}_k) - f(x^*) \leq \tfrac{2L}{k(k+1)}\|x_0 - x^*\|^2.$$

Proof. The result follows from the previous theorem (with $x = x^*$) by using the parameters α_k and β_k stated in the premise.

In order to fine an ϵ-solution, i.e., a point $\bar{x} \in X$ s.t. $f(\bar{x}) - f^* \leq \epsilon$, the number of iterations is bounded by $\sqrt{\frac{2L\|x_0 - x^*\|^2}{\epsilon}}$. ∎

We now discuss AGD for strongly convex problems. Suppose $f \in \mathscr{S}^{1,1}_{\mu,L}(\mathbb{R}^n$. We describe a mulit-phase algorithm by restarting the AGD method every

$$N \equiv \left\lceil \sqrt{\tfrac{8L}{\mu}} \right\rceil$$

iterations.

Input: $p_0 \in X$.
Phase $t = 1, 2, \ldots$:
Set $p_t = \bar{x}_N$, where $\bar{x}_N$ is obtained from the AGD method with $x_0 = p_{t-1}$.

We are now ready to

Theorem 5.9. *For any $t \geq 1$, we have*

$$\|p_t - x^*\|^2 \leq (\tfrac{1}{2})^t \|p_0 - x^*\|^2.$$

Proof. Note that $f(p_t) - f^* \leq 2L\|p_{t-1} - x^*\|^2/N^2$. Using the strong convexity and the definition of N, we have

$$\|p_t - x^*\|^2 \leq \frac{4L}{\mu N^2}\|p_{t-1} - x^*\|^2 \leq \frac{1}{2}\|p_{t-1} - x^*\|^2.$$

■

To have $\|p_t - x^*\|^2 \leq \varepsilon$, the total number of iterations is bounded by

$$\left\lceil \sqrt{\frac{8L}{\mu}} \right\rceil \log \frac{\|p_0 - x^*\|^2}{\varepsilon}.$$

5.7 Lower Complexity Bound

In this section, we first establish the Lower complexity bound for $\mathscr{F}_L^{1,1}(\mathbb{R}^n)$.

We make the following assumption about the algorithms.

Assumption. An iterative method $\mathscr{M}$ generates a sequence of test points $\{x_k\}$ such that

$$x_k \in x_0 + \operatorname{lin}\{f'(x_0), \ldots, f'(x_{k-1})\}, k \geq 1.$$

This assumption is not absolutely necessary and it can be avoided by a more sophisticated reasoning in the original development by Nemirovski and Yudin. We want to point out the "worst function in the world ($\mathscr{F}_L^{1,1}(\mathbb{R}^n)$)". This function appears to be difficult for all iterative methods satisfying this assumption.

Let us fix some constant $L > 0$. Consider the family of quadratic functions

$$f_k(x) = \frac{L}{4}\left\{\frac{1}{2}\left[(x^{(1)})^2 + \textstyle\sum_{i=1}^{k-1}(x^{(i)} - x^{(i+1)})^2 + (x^{(k)})^2\right] - x^{(1)}\right\}.$$

It can be seen that

$$f(x) = \frac{L}{8}x^T A x - \frac{L}{4}e_1$$

where

$$A_k = \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 & 0 & \vert 0 \\ -1 & 2 & -1 & \ldots & 0 & 0 & \vert 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 2 & -1 & \vert 0 \\ 0 & 0 & 0 & \ldots & -1 & 2 & \vert 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & \vert 0 \end{pmatrix}$$

and $e_1 = (1, 0, \ldots, 0)^T$. Hence, $f_k'(x) = \frac{L}{4}[A_k x - e_1]$, $f_k''(x) = \frac{L}{4}A_k$, and $0 \preceq f_k''(x) \preceq LI_n$. (using $(x-y)^2 \leq 2x^2 + 2y^2$).

Setting $f'(\bar{x}) = L(A_k\bar{x} - e_1)/4 = 0$, we have the following unique solution of f_k:

$$\bar{x}_k^{(i)} = \begin{cases} 1 - \frac{i}{k+1}, & i = 1, \ldots, k, \\ 0, & k+1 \leq i \leq n. \end{cases}$$

Hence the optimal value is given by

$$f_k^* = \frac{L}{4}\left[\frac{1}{2}\langle A_k\bar{x}_k, \bar{x}_k\rangle - \langle e_1, \bar{x}_k\rangle\right] = -\frac{L}{8}\langle e_1, \bar{x}_k\rangle = \frac{L}{8}(-1 + \frac{1}{k+1}).$$

Therefore,

$$\begin{aligned}
&\|\bar{x}_k\|^2 = \textstyle\sum_{i=1}^n (\bar{x}_k^{(i)})^2 = \sum_{i=1}^k (1 - \frac{i}{k+1})^2 \\
&= k - \tfrac{2}{k+1} \textstyle\sum_{i=1}^k i + \frac{1}{(k+1)^2} \sum_{i=1}^k i^2 \\
&\leq k - \tfrac{2}{k+1} \tfrac{k(k+1)}{2} + \tfrac{1}{(k+1)^2} \tfrac{(k+1)^3}{3} \\
&= \tfrac{1}{3}(k+1).
\end{aligned}$$

Denote $\mathbb{R}^{k,n} \equiv \{x \in \mathbb{R}^n | x^{(i)} = 0, k+1 \leq i \leq n\}$. Only the first k components can differ from 0. From the analytical form of f_k, it is easy to see that for all $x \in \mathbb{R}^{k,n}$, we have

$$f_p(x) = f_k(x), p = k, \ldots, n, \quad \forall x \in \mathbb{R}^{k,n}.$$

Let us fix some p, $1 \leq p \leq n$.

Lemma 5.14. *Let $x_0 = 0$. Then for any sequence $\{x_k\}_{k=0}^p$ satisfying*

$$x_k \in L_k \equiv \operatorname{lin}\{f'_p(x_0), \ldots, f'_p(x_{k-1})\},$$

we have $L_k \subseteq \mathbb{R}^{k,n}$.

Proof. Since $x_0 = 0$, we have $f'_p(x_0) = -L/4e_1 \in \mathbb{R}^{1,n}$. Thus $L_1 \equiv \mathbb{R}^{1,n}$. Let $L_k \subseteq \mathbb{R}^{k,n}$ for some $k < p$. Since A_p is tri-diagonal, for any $x \in \mathbb{R}^{k,n}$, we have $f'_p(x) \in \mathbb{R}^{k+1,n}$. Therefore $L_{k+1} \subseteq \mathbb{R}^{k+1,n}$ and we can complete the proof by induction. ■

Corollary 5.4. *For any sequence $\{x_k\}_{k=0}^p$ s.t. $x_0 = 0$ and $x_k \in L_k$, we have $f_p(x_k) \geq f_k^*$.*

Proof. $x_k \in L_k \subseteq \mathbb{R}^{k,n}$ and therefore $f_p(x_k) = f_k(x_k) \geq f_k^*$. ■

Theorem 5.10. *For any k, $1 \leq k \leq \frac{1}{2}(n-1)$ and any $x_0 \in \mathbb{R}^n$, there exists a function $f \in \mathscr{F}_L^{1,1}(\mathbb{R}^n)$ s.t. for any first-order method M satisfying our assumption, we have*

$$f(x_k) - f^* \geq \tfrac{3L\|x_0 - x^*\|^2}{32(k+1)^2},$$

where x^* is the minimum of $f(x)$ and $f^* = f(x^*)$.

Proof. It is clear that the methods are invariant w.r.t. a shift of variables. The sequence of iterates of $f(x)$ starting from x_0 is just a shift of the sequence generated for $\bar{f}(x) = f(x + x_0)$ starting from the origin. Therefore we can assume $x_0 = 0$.

Now fix k and apply M to minimize $f(x) = f_{2k+1}(x)$. Then $x^* = \bar{x}_{2k+1}$ and $f^* = f_{2k+1}^*$. Using Corollary 5.4, $f(x_k) = f_{2k+1}(x_k) = f_k(x_k) \geq f_k^*$. Hence, by using the expressions for f_k^* and f_{2k+1}^*, the bound on $\|\bar{x}_k\|^2$, the fact $x_0 = 0$, we have

$$\tfrac{f(x_k) - f^*}{\|x_0 - x^*\|^2} \geq \frac{\frac{L}{8}\left(-1 + \frac{1}{k+1} + 1 - \frac{1}{2k+2}\right)}{\frac{1}{3}(2k+2)} = \tfrac{3}{8} L \tfrac{1}{4(k+1)^2}.$$

■

We add a few remarks on the above lower complexity bound. First, the above theorem is valid only under the assumption $k \le \frac{1}{2}(n-1)$. Hence, they are valid for very large problems, in which we cannot wait even for n iterates of the methods. Second, for problems with a moderate dimension, it describes the performance of numerical methods on the initial stage. Third, it warns us that without a direct use of finite dimensional arguments, we cannot get better complexity.

Observe that the result states a lower bound on the convergence in terms of function values, which is rather optimistic. However, the result on the behavior of $\{x_k\}$ is quite disappointing. It can be shown that $\|x_k - x^*\|^2 \ge \frac{1}{8}\|x_0 - x^*\|^2$ for any $k \le \frac{1}{2}(n-1)$. Hence, the convergence to the optimal solution can be arbitrarily slow. The only thing we can do is to try to find problem classes in which the situation could be better, e.g., strongly convex problems.

We now establish a Lower complexity for $\mathscr{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$. Note that we do not say anything about the dimension n and n could be ∞. We are going to give an example of some bad functions defined in infinite-dimensional space. We could do that also in a finite dimension, but the corresponding reasoning is more complicated.

Consider $\mathbb{R}^\infty \equiv \mathscr{L}_2$, the space of all sequences $x = \{x^{(i)}\}_{i=1}^\infty$ with finite norm $\|x\|^2 = \sum_{i=1}^\infty (x^{(i)})^2 < \infty$. Let us choose some parameters $\mu > 0$ and $Q_f > 1$ and define

$$f_{\mu,Q_f}(x) = \frac{\mu(Q_f-1)}{8}\left\{(x^{(1)})^2 + \sum_{i=1}^\infty (x^{(i)} - x^{(i+1)})^2 - 2x^{(1)}\right\} + \frac{\mu}{2}\|x\|^2.$$

Denote

$$A = \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \dots & 0 \\ 0 & -1 & 2 & \dots & 0 \\ 0 & 0 & 0 & \dots & \dots \end{pmatrix}$$

and $e_1 = (1,0,\dots,0)^T$. Hence, $f'_{\mu,Q_f}(x) = \frac{\mu(Q_f-1)}{4}Ax + \mu x - \frac{\mu(Q_f-1)}{4}e_1$, $f''_{\mu,Q_f}(x) = \frac{\mu(Q_f-1)}{4}A + \mu I$, and $\mu I \preceq f''_k(x) \preceq (\mu(Q_f-1)+\mu)I = \mu Q_f I$ by using $(x-y)^2 \le 2x^2 + 2y^2$.

Setting $f'_{\mu,Q_f}(x) = 0$ we have $(A + \frac{4}{Q_f-1})x = e_1$. The coordinate form is given by

$$\begin{aligned} &2\frac{Q_f+1}{Q_f-1}x^{(1)} - x^{(2)} = 1 \\ &x^{(k+1)} - 2\frac{Q_f+1}{Q_f-1}x^{(k)} + x^{(k-1)} = 0, k = 2,\dots \end{aligned}$$

Let q be the smallest root of the equation

$$q^2 - 2\frac{Q_f+1}{Q_f-1}q + 1 = 0.$$

That is $q = \frac{\sqrt{Q_f}-1}{\sqrt{Q_f}+1}$. Then the sequence $(x^*)^{(k)} = q^k$, $k = 1,2,\dots$, satisfies the system.

Theorem 5.11. *For any x_0, there exists a function $f \in \mathscr{S}^{1,1}_{\mu,\mu Q_f}(\mathbb{R}^\infty)$ s.t. for any first-order methods*

$$\|x_k - x^*\|^2 \geq \left(\frac{\sqrt{Q_f}-1}{\sqrt{Q_f}+1}\right)^{2k} \|x_0 - x^*\|^2,$$
$$f(x_k) - f^* \geq \tfrac{\mu}{2}\left(\frac{\sqrt{Q_f}-1}{\sqrt{Q_f}+1}\right)^{2k} \|x_0 - x^*\|^2,$$

where x^* is the minimum of f and $f^* = f(x^*)$.

Proof. Choose $x_0 = 0$. Bound $\|x_0 - x^*\|^2$ and $\|x_k - x^*\|^2$.

$$\|x_0 - x^*\|^2 = \|x^*\|^2 = \textstyle\sum_{i=1}^\infty q^{2i} = \frac{q^2}{1-q^2}.$$

Note that $x^k \in \mathbb{R}^{n,k}$ hence

$$\|x_k - x^*\|^2 \geq \textstyle\sum_{i=k+1}^\infty q^{2i} = \frac{q^{2(k+1)}}{1-q^2} = q^{2k}\|x_0 - x^*\|^2.$$

The second bound follow from the first one and $f(x_k) - f^* \geq \mu\|x_k - x^*\|^2/2$. ■

It then follows that

$$\|x_k - x^*\|^2 \leq \left(\frac{\sqrt{Q_f}-1}{\sqrt{Q_f}+1}\right)^{2k} \|x_0 - x^*\|^2$$
$$= \left(1 - \frac{2}{\sqrt{Q_f}+1}\right)^{2k} \|x_0 - x^*\|^2 \leq \exp\left(-\frac{4k}{\sqrt{Q_f}+1}\right) \|x_0 - x^*\|^2.$$

Hence, to have $\|x_k - x*\|^2 \leq \varepsilon$, the number of iterations will be at least $k \geq \frac{\sqrt{Q_f}+1}{4} \log \frac{\|x_0 - x^*\|^2}{\varepsilon}$.

5.8 Conditional gradient method

In this section, we consider the problem of

$$\min_{x\in X} f(x), \tag{5.8.1}$$

where $X \subseteq \mathbb{R}^n$ is a closed convex set and $f \in C^{1,1}$.

The classic CndG method is one of the earliest iterative algorithms to solve problem (5.8.1). The basic scheme of this algorithm is stated as follows.

Algorithm 5.1 The Conditional Gradient (CndG) Method

```
Let $x_0 \in X$ be given. Set $y_0 = x_0$.
for $k = 1, \ldots$ do
    Compute $x_k \in \mathrm{Argmin}_{x \in X} \langle f'(y_{k-1}), x \rangle$.
    Set $y_k = (1 - \alpha_k) y_{k-1} + \alpha_k x_k$ for some $\alpha_k \in [0,1]$.
end for
```

In order to guarantee the convergence of the classic CndG method, we need to properly specify the stepsizes α_k used in the definition of y_k. There are two popular options for selecting α_k: one is to set

$$\alpha_k = \tfrac{2}{k+1}, \quad k = 1, 2, \ldots, \tag{5.8.2}$$

and the other is to compute α_k by solving a one-dimensional minimization problem:

$$\alpha_k = \mathrm{argmin}_{\alpha \in [0,1]} f((1-\alpha) y_{k-1} + \alpha x_k), \quad k = 1, 2, \ldots. \tag{5.8.3}$$

We now formally describe the convergence properties of the above classic CndG method. Observe that we state explicitly in Theorem 5.12 how the rate of convergence associated with this algorithm depends on distance between the previous iterate y_{k-1} and x_k, i.e., $\|x_k - y_{k-1}\|$. Also observe that, given a candidate solution $\bar{x} \in X$, we use the function optimality gap $f(\bar{x}) - f^*$ as a termination criterion for the algorithm. It is also possible to show that the CndG method also exhibit the same rate of convergence in terms of a stronger termination criterion, i.e., the Wolfe gap given by $\max_{x \in X} \langle f'(\bar{x}), \bar{x} - x \rangle$. The following quanitity will be used our convergence analysis.

$$\Gamma_k := \begin{cases} 1, & k = 1, \\ (1 - \gamma_k)\Gamma_{k-1}, & k \geq 2. \end{cases} \tag{5.8.4}$$

Theorem 5.12. *Let $\{x_k\}$ be the sequence generated by the classic CndG method applied to problem (5.8.1) with the stepsize policy in (5.8.2) or (5.8.3). Then for any $k = 1, 2, \ldots$,*

$$f(y_k) - f^* \leq \tfrac{2L}{k(k+1)} \textstyle\sum_{i=1}^k \|x_i - y_{i-1}\|^2. \tag{5.8.5}$$

Proof. Let Γ_k be defined in (5.8.4) with

$$\gamma_k := \tfrac{2}{k+1}. \tag{5.8.6}$$

It is easy to check that

$$\Gamma_k = \tfrac{2}{k(k+1)} \quad \text{and} \quad \tfrac{\gamma_k^2}{\Gamma_k} \leq 2, \quad k = 1, 2, \ldots. \tag{5.8.7}$$

Denoting $\tilde{y}_k = (1 - \gamma_k) y_{k-1} + \gamma_k x_k$, we conclude from from (5.8.2) (or (5.8.3)) and the definition of y_k in Algorithm 5.1 that $f(y_k) \leq f(\tilde{y}_k)$. It also follows from the definition of $\tilde{y}_k$ that $\tilde{y}_k - y_{k-1} = \gamma_k (x_k - y_{k-1})$. Letting $l_f(x; y) = f(x) + \langle \nabla f(x), y - x \rangle$ and using these two observations, the smoothness of f, the definition of x_k and the convexity of $f(\cdot)$, we have

$$\begin{aligned} f(y_k) &\le f(\tilde{y}_k) \le l_f(y_{k-1};\tilde{y}_k) + \tfrac{L}{2}\|y_k - y_{k-1}\|^2 \\ &= (1-\gamma_k)f(y_{k-1}) + \gamma_k l_f(y_{k-1};x_k) + \tfrac{L}{2}\gamma_k^2\|x_k - y_{k-1}\|^2 \\ &\le (1-\gamma_k)f(y_{k-1}) + \gamma_k l_f(y_{k-1};x) + \tfrac{L}{2}\gamma_k^2\|x_k - y_{k-1}\|^2, \\ &\le (1-\gamma_k)f(y_{k-1}) + \gamma_k f(x) + \tfrac{L}{2}\gamma_k^2\|x_k - y_{k-1}\|^2, \ \forall x \in X. \end{aligned} \tag{5.8.8}$$

Subtracting $f(x)$ from both sides of the above inequality, we obtain

$$f(y_k) - f(x) \le (1-\gamma_k)[f(y_{k-1}) - f(x)] + \tfrac{L}{2}\gamma_k^2\|x_k - y_{k-1}\|^2, \tag{5.8.9}$$

which then implies that

$$\begin{aligned} f(y_k) - f(x) &\le \Gamma_k(1-\gamma_1)[f(y_0) - f(x)] + \tfrac{\Gamma_k L}{2}\textstyle\sum_{i=1}^k \tfrac{\gamma_i^2}{\Gamma_i}\|x_i - y_{i-1}\|^2 \\ &\le \tfrac{2L}{k(k+1)}\textstyle\sum_{i=1}^k \|x_i - y_{i-1}\|^2, \ \ k = 1,2,\ldots, \end{aligned} \tag{5.8.10}$$

where the last inequality follows from the fact that $\gamma_1 = 1$ and (5.8.7). ■

We now add a few remarks about the results obtained in Theorem 5.12. Let us denote

$$\bar{D}_X \equiv \bar{D}_{X,\|\cdot\|} := \max_{x,y\in X} \|x - y\|. \tag{5.8.11}$$

Firstly, note that by (5.8.5) and (5.8.11), we have, for any $k = 1,\ldots,$

$$f(y_k) - f^* \le \tfrac{2L}{k+1}\bar{D}_X^2.$$

Hence, the number of iterations required by the classic CndG method to find an ε-solution of problem (5.8.1) is bounded by

$$\mathcal{O}(1)\tfrac{L\bar{D}_X^2}{\varepsilon}. \tag{5.8.12}$$

Secondly, although the CndG method does not require the selection of the norm $\|\cdot\|$, the iteration complexity of this algorithm, as stated in (5.8.12), does depend on $\|\cdot\|$ as the two constants, i.e., $L \equiv L_{\|\cdot\|}$ and $\bar{D}_X \equiv \bar{D}_{X,\|\cdot\|}$, depend on $\|\cdot\|$. However, since the result in (5.8.12) holds for an arbitrary $\|\cdot\|$, the iteration complexity of the classic CndG method to solve problem (5.8.1) can actually be bounded by

$$\mathcal{O}(1)\inf_{\|\cdot\|}\left\{\tfrac{L_{\|\cdot\|}\bar{D}_{X,\|\cdot\|}^2}{\varepsilon}\right\}. \tag{5.8.13}$$

For example, if X is a simplex, a widely-accepted strategy to accelerate gradient type methods is to set $\|\cdot\| = \|\cdot\|_1$ and $d(x) = \sum_{i=1}^n x_i \log x_i$ as the distance generating function, in order to obtain (nearly) dimension-independent complexity results, which only grow mildly with the increase of the dimension of the problem. On the other hand, the classic CndG method does automatically adjust to the geometry of the feasible set X in order to obtain such scalability to high-dimensional problems.

We can also extend the analysis of the CndG method to the nonconvex setting. For this purpose, we need to define a new termination criterion given by $Q(\bar{x}) := \max_{x\in X}\langle \nabla f(\bar{x}), \bar{x}-x\rangle$. We often call $Q(\bar{x})$ the Wolfe-gap at $\bar{x}$. We can show that the number of iterations required by the CndG method to find a solution $\bar{x}\in X$ s.t. $Q(\bar{x}) \le \varepsilon$ can be bounded by $\mathscr{O}(1/\varepsilon^2)$ for the nonconvex setting when f is smooth but not necessarily convex.

5.9 Ellipsoid methods for convex optimization

The basic idea is to use approximate localization sets.

Let H be a positive definite symmetric $n\times n$ matrix. Consider the ellipsoid

$$E(H,\bar{x}) = \{x\in\mathbb{R}^n | \langle H^{-1}(x-\bar{x}), x-\bar{x}\rangle \le 1\}.$$

Choose a direction $g\in\mathbb{R}^n$ and define

$$E_+ = \{x\in E(H,\bar{x}) | \langle g, \bar{x}-x\rangle \ge 0\}.$$

We need to show that E_+ belongs to another ellipsoid, whose volume is strictly smaller than that of $E(H,\bar{x})$.

Lemma 5.15. *Denote*

$$\begin{aligned}\bar{x}_+ &= \bar{x} - \tfrac{1}{n+1}\tfrac{Hg}{\langle Hg,g\rangle^{\frac{1}{2}}}\\ H_+ &= \tfrac{n^2}{n^2-1}\left(H - \tfrac{2}{n+1}\tfrac{Hgg^TH}{\langle Hg,g\rangle}\right).\end{aligned}$$

Then $E_+ \subset E(H_+,\bar{x}_+)$ *and*

$$\mathrm{Vol}_n E(H_+,\bar{x}_+) \le \left(1 - \tfrac{1}{(n+1)^2}\right)^{\frac{n}{2}} \mathrm{Vol}_n E(H,\bar{x}).$$

Proof. Denote $G = H^{-1}$ and $G_+ = H_+^{-1}$. It is clear that

$$G_+ = \tfrac{n^2-1}{n^2}\left(G + \tfrac{2}{n-1}\tfrac{gg^T}{\langle Hg,g\rangle}\right).$$

W.L.O.G., we assume $\bar{x}=0$ and $\langle Hg,g\rangle = 1$. Suppose that $x\in E_+$. Note that $\bar{x}_+ = -\frac{1}{n+1}Hg$. Therefore,

$$\begin{aligned}
&\|x-\bar{x}_+\|^2_{G_+} \\
&\quad = \tfrac{n^2-1}{n^2}\left(\|x-\bar{x}_+\|^2_G + \tfrac{2}{n-1}\langle g, x-\bar{x}_+\rangle^2\right), \\
&\|x-\bar{x}_+\|^2_G = \|x\|^2_G + \tfrac{2}{n+1}\langle g,x\rangle + \tfrac{1}{(n+1)^2}, \\
&\langle g, x-\bar{x}_+\rangle^2 = \langle g,x\rangle^2 + \tfrac{2}{n+1}\langle g,x\rangle + \tfrac{1}{(n+1)^2}.
\end{aligned}$$

Putting all these terms together, we obtain

$$\begin{aligned}
\|x-\bar{x}_+\|^2_{G+} = \tfrac{n^2-1}{n^2}\Big(&\|x\|^2_G + \\
&\tfrac{2}{n-1}\langle g,x\rangle^2 + \tfrac{2}{n-1}\langle g,x\rangle + \tfrac{1}{n^2-1}\Big).
\end{aligned}$$

Note that $\langle g,x\rangle \leq 0$ and $\|x\|_G \leq 1$. Hence, $|\langle g,x\rangle| \leq \|g\|_H\|x\|_G \leq 1$ and

$$\langle g,x\rangle^2 + \langle g,x\rangle = \langle g,x\rangle(1+\langle g,x\rangle) \leq 0,$$

which implies that

$$\|x-\bar{x}_+\|^2_{G+} \leq \tfrac{n^2-1}{n^2}\left(\|x\|^2_G + \tfrac{1}{n^2-1}\right) \leq 1.$$

Thus, we have $E_+ \subset E(H_+,\bar{x}_+)$. Using the matrix determinatnt lemma that $\det(A+UV^T) = (1+V^TA^{-1}U)\det(A)$, we can estimate the volume of $E(H_+,\bar{x}_+)$:

$$\begin{aligned}
\tfrac{\mathrm{Vol}_n E(H_+,\bar{x}_+)}{\mathrm{Vol}_n E(H,\bar{x})} &= \left(\tfrac{\det H_+}{\det H}\right)^{\frac{1}{2}} = \left[\left(\tfrac{n^2}{n^2-1}\right)^n \tfrac{n-1}{n+1}\right]^{\frac{1}{2}} \\
&= \left[\tfrac{n^2}{n^2-1}\left(1-\tfrac{2}{n+1}\right)^{\frac{1}{n}}\right]^{\frac{n}{2}} \\
&\leq \left[\tfrac{n^2}{n^2-1}\left(1-\tfrac{2}{n(n+1)}\right)\right]^{\frac{n}{2}} \\
&= \left[\tfrac{n^2(n^2+n-2)}{n(n-1)(n+1)^2}\right]^{\frac{n}{2}} \\
&= \left[1-\tfrac{1}{(n+1)^2}\right]^{\frac{n}{2}},
\end{aligned}$$

where the inequality follows from the concavity of $(1-x)^{\frac{1}{n}}$. ■

Geometrically, $E(H_+,\bar{x}_+)$ is the ellipsoid of the minimal volume containing the half of the initial ellipsoid E_+. We now formally describe the Ellipsoid method.

Algorithm 5.2 The Ellipsoid Method

```
Choose $y_0 \in \mathbb{R}^n$ and $R > 0$ such that $B_2(y_0,R) \supseteq Q$. Set $H_0 = R^2 I_n$.
for $k = 1,\dots$ do
    $g_k = \begin{cases} g(y_k), & \text{if} y_k \in Q, \\ \bar{g}(y_k), & \text{if} y_k \notin Q. \end{cases}$
    $y_{k+1} = y_k - \frac{1}{n+1} \frac{H_k g_k}{\langle H_k g_k, g_k \rangle^{\frac{1}{2}}}$
    $H_{k+1} = \frac{n^2}{n^2-1} \left( H_k - \frac{2}{n+1} \frac{H_k g_k g_k^T H_k}{\langle H_k g_k, g_k \rangle} \right).$
end for
```

This method is a particular implementation of the general cutting plane scheme by setting

$$E_k = \{x \in \mathbb{R}^n | \langle H_k^{-1}(x - y_k), x - y_k \rangle \le 1\}$$

and y_k being the center of this ellipsoid.

Denote $Y = \{y_k\}_{k=0}^{\infty}$, $X = Y \cap Q$, and $f_k^* = \min_{0 \le j \le k} f(x_j)$.

Theorem 5.13. *Let f be Lipschitz continuous on $B_2(x^*,R)$ with constant M. Then for $i(k) > 0$, we have*

$$f_{i(k)}^* - f^* \le MR \left(1 - \frac{1}{(n+1)^2}\right)^{\frac{k}{2}} \left[\frac{\mathrm{Vol}_n B_2(x_0,R)}{\mathrm{Vol}_n Q}\right]^{\frac{1}{n}}.$$

We now discuss the complexity of the Ellipsoid method, We need some additional assumption to guarantee $X \neq \emptyset$.

Assume $\exists \rho > 0$ and $\bar{x} \in Q$ s.t. $B_2(\bar{x},\rho) \subseteq Q$. Then

$$\begin{aligned} \left(\frac{\mathrm{Vol}_n E_k}{\mathrm{Vol}_n Q}\right)^{\frac{1}{n}} &\le \left(1 - \frac{1}{(n+1)^2}\right)^{\frac{k}{2}} \left(\frac{\mathrm{Vol}_n B_2(x_0,R)}{\mathrm{Vol}_n Q}\right)^{\frac{1}{n}} \\ &\le \frac{1}{\rho} e^{-\frac{k}{2(n+1)^2}} R. \end{aligned}$$

This implies $i(k) > 0$ for all $k \ge 2(n+1)^2 \ln \frac{R}{\rho}$. If $i(k) > 0$, then $f_{i(k)}^* - f^* \le \frac{1}{\rho} MR^2 e^{-\frac{k}{2(n+1)^2}}$. This implies that the complexity of the Ellipsoid method is bounded by $2(n+1)^2 \ln \frac{MR^2}{\rho \varepsilon}$.

Polynomial dependence on $\ln \frac{1}{\varepsilon}$ and a polynomial dependence on logarithms of the class parameters M, R and ρ. Several methods that work with localization sets in the form of a polytope:

- $E_k = \{x \in \mathbb{R}^n | \langle a_j, x \rangle \le b_j, j = 1, \dots, m_k\}$.
- Inscribed Ellipsoid method. The point y_k is chosen as $y_k =$ center of the maximal ellipsoid $W_k \subset E_k$.
- Analytic center method. y_k is chosen as the minimum of the analytic barrier:

$$F_k(x) = -\sum_{j=1}^{m_k} \ln(b_j - \langle a_j, x \rangle).$$

- Volumetric center method. This is also a barrier-type scheme. y_k is chosen as the minimum of the volumetric barrier $V_k(x) = \ln \det F_k''(x)$, where $F_k(x)$ is the analytic barrier of E_k.

All these methods are in $\mathscr{O}(n \ln^p \frac{1}{\varepsilon})$ with $p = 1$ or 2, but higher cost per iteration.

5.10 Newton's Method

Newton method intends to find the root of a function $\phi(t) : \mathbb{R} \to \mathbb{R}$ s.t.

$$\phi(t^*) = 0.$$

Assume t is close to t^* and consider

$$\phi(t + \Delta t) = \phi(t) + \phi'(t)\Delta t + o(|\Delta t|).$$

Setting $\phi(t+\Delta t) = 0$, or approximately $\phi(t) + \phi'(t)\Delta t = 0$, we have $\Delta_t = -\phi(t)/\phi'(t)$. Hence, we can derive the iterative scheme as

$$t_{k+1} = t_k - \tfrac{1}{\phi'(t_k)}\phi(t_k).$$

In order solve a system of nonlinear equations: $F(x) = 0$ with $F : \mathbb{R}^n \to \mathbb{R}^n$, we approximate it by $F(x) + F'(x)\Delta x = 0$, or $\Delta x = -[F'(x)]^{-1}F(x)$, and define the iterative scheme:

$$x_{k+1} = x_k - [F'(x_k)]^{-1}F(x_k).$$

For solving unconstrained minimization:

$$\min_{x\in\mathbb{R}^n} f(x),$$

we apply Newton's method to solve $f'(x) = 0$ and obtain

$$x_{k+1} = x_k - [f''(x_k)]^{-1}f'(x_k).$$

There also exists a different way to derive the iterative scheme. Set the gradient of

$$\phi(x) = f(x_k) + \langle f'(x_k), x - x_k\rangle + \tfrac{1}{2}\langle f''(x_k)(x - x_k), x - x_k\rangle. \tag{5.10.1}$$

to be zero, we have

$$\phi'(x_{k+1}) = f'(x_k) + f''(x_k)(x_{k+1} - x_k) = 0.$$

Newton's method can converge very fast in a neighborhood of a strict local minima. However, it can break down if $f''(x_k)$ is degenerate and can can diverge if x_0 is far away from x^*.

Example 5.2. Find the root of $\phi(t) = \frac{t}{\sqrt{1+t^2}}$ $(t^* = 0)$.

$$\begin{aligned} \phi'(t) &= \tfrac{1}{(1+t^2)^{3/2}}. \\ t_{k+1} &= t_k - \tfrac{\phi(t)}{\phi'(t)} \\ &= t_k - (1+t_k^2)t_k = -t_k^3. \end{aligned}$$

If $|t_0| < 1$, convergence is extremely fast. If $|t_0| = 1$, oscillation. If $|t_0| > 1$, diverge.

In order to address this divergence issue, one often resort to the *Damped Newton's method.*

$$x_{k+1} = x_k - h_k[f''(x_k)]^{-1}f'(x_k),$$

where $h_k > 0$ is a stepsize ($h_k < 1$ in the beginning, $h_k = 1$ in the final stage).

We now establish the locally quadratic convergence for Newton's method under the following assumptions.

a) $f \in C_M^{2,2}(\mathbb{R}^n)$.
b) $f''(x^*) \succeq lI_n$, $l > 0$.
c) Starting point x_0 is close to x^*.

Theorem 5.14. *If* $\|x_0 - x^*\| \le 2l/3M$, *then*

$$\|x_{k+1} - x^*\| \le \tfrac{M\|x_k - x^*\|^2}{2(l - M\|x_k - x^*\|)} \le \tfrac{3M\|x_k - x^*\|^2}{2l}.$$

Proof. By $x_{k+1} = x_k - [f''(x_k)]^{-1}f'(x_k)$,

$$\begin{aligned} x_{k+1} - x^* &= x_k - x^* - [f''(x_k)]^{-1}\textstyle\int_0^1 f''(x^* + \tau(x_k - x^*)(x_k - x^*)d\tau \\ &= [f''(x_k)]^{-1}\textstyle\int_0^1 [f''(x_k) - f''(x^* + \tau(x_k - x^*)](x_k - x^*)d\tau \\ &= [f''(x_k)]^{-1}G_k(x_k - x^*), \end{aligned}$$

where $G_k = \int_0^1 [f''(x_k) - f''(x^* + \tau(x_k - x^*)]d\tau$. Denote $r_k = \|x_k - x^*\|$. We have

$$\|G_k\| \le \int_0^1 \tau r_k M d\tau = \tfrac{1}{2}Mr_k.$$

Moreover,

$$f''(x_k) \succeq f''(x^*) - Mr_kI_n \succeq (l - Mr_k)I_n.$$

Thus if $l/M > r_k$, then $f''(x_k) \succ 0$ and $\|[f''(x_k)]^{-1}\| \le (l - Mr_k)^{-1}$. Hence $r_{k+1} \le \frac{Mr_k^2}{2(l - Mr_k)}$. It can be shown inductively that if $r_k \le 2l/3M$ then $r_{k+1} \le r_k \le \frac{2l}{3M}$. The result now follows by combining these two inequalities. ∎

5.11 Quasi-Newton (Variable Metric) method

By choosing different objective functions in (5.10.1), we can develop different optimization methods.

- $\phi_1(x) = f(\bar{x}) + \langle f'(\bar{x}), x-\bar{x}\rangle + \frac{1}{2h}\|x-\bar{x}\|^2$ yields gradient descent method.
- $\phi_2(x) = f(\bar{x}) + \langle f'(\bar{x}), x-\bar{x}\rangle + \frac{1}{2}\langle f''(\bar{x})(x-\bar{x}), x-\bar{x}\rangle$ yields Newton's method.

Quasi-Newton method is something in between these two methods by seting

$$\phi_G(x) = f(\bar{x}) + \langle f'(\bar{x}), x-\bar{x}\rangle + \tfrac{1}{2}\langle G(x-\bar{x}), x-\bar{x}\rangle$$

for some $G \succ 0$. $\phi_G'(x_G^*) = 0$ yields $x_G^* = \bar{x} - G^{-1}f'(\bar{x})$. The basic idea of Quasi-Newton method is to form a sequence of matrices $\{G_k\} : G_k \to f''(x^*)$, or $\{H_k = G_k^{-1}\} \to [f''(x^*)]^{-1}$. These methods are also called variable metric methods.

The variable metric method is named for the following reasons. Note that the gradient and Hessian are defined w.r.t. a standard Euclidean inner product $\langle x,y\rangle = \sum_{i=1}^n x^{(i)}y^{(i)}$ and $\|x\| = \langle x,x\rangle^{\frac{1}{2}}$. Now consider a new inner product: let $A \succ 0$ be given and define $\langle x,y\rangle_A = \langle Ax,y\rangle$ and $\|x\|_A = \langle Ax,x\rangle^{\frac{1}{2}}$. Topologically, the new metric is equivalent to the old one:

$$\lambda_n(A)^{\frac{1}{2}}\|x\| \le \|x\|_A \le \lambda_1(A)^{\frac{1}{2}}\|x\|.$$

However, the gradient and Hessian will change:

$$\begin{aligned} f(x+h) &= f(x) + \langle f'(x), h\rangle + \tfrac{1}{2}\langle f''(x)h, h\rangle + o(\|h\|) \\ &= f(x) + \langle A^{-1}f'(x), h\rangle_A + \tfrac{1}{2}\langle A^{-1}f''(x)h, h\rangle_A + o(\|h\|_A). \end{aligned}$$

With this new gradient: $A^{-1}f'(x)$ and new Hessian: $A^{-1}f''(x)$, Newton's method is equivalent to gradient method using $A = f''(x)$.

We now state the algorithmic form of Quasi-Newton Method

```
0. Choose $x_0$, set $H_0 = I_n$ and compute $f(x_0)$, $f'(x_0)$.
1. $k^{th}$ iteration:
    a) Set $p_k = H_k f'(x_k)$.
    b) Find $x_{k+1} = x_k - h_k p_k$.
    c) Compute $f(x_{k+1})$ and $f'(x_{k+1})$.
    d) Update $H_k : H_k \to H_{k+1}$.
```

One important question for Quasi-Newton method is how to update H_k. Consider the quadratic function $f(x) = \alpha + \langle a,x\rangle + \frac{1}{2}\langle Ax,x\rangle$. $f'(x) = Ax + \alpha$ and $f'(x) - f'(y) = A(x-y)$. The classic selection rule of Quasi-Newton method is set H_{k+1} s.t.

$$H_{k+1}[f'(x_{k+1}) - f'(x_k)] = x_{k+1} - x_k.$$

Denote $\Delta H_k = H_{k+1} - H_k$, $r_k = f'(x_{k+1}) - f'(x_k)$, $\delta_k = x_{k+1} - x_k$. We list the following widely used variants of Quasi-Newton methods.

- Rank-one correction
$$\Delta H_k = \frac{(\delta_k - H_k r_k)(\delta_k - H_k r_k)^T}{\langle \delta_k - H_k r_k, r_k \rangle}.$$
- Davidon-Fletcher-Powell Scheme (DFP)
$$\Delta H_k = \frac{\delta_k \delta_k^T}{\langle r_k, \delta_k \rangle} - \frac{H_k r_k r_k^T H_k}{\langle H_k r_k, r_k \rangle}.$$
- Broyden-Fletcher-Goldfarb-Shanno Scheme (BFGS)
$$\Delta H_k = \frac{H_k r_k \delta_k^T + \delta_k r_k^T H_k}{\langle H_k r_k, r_k \rangle} - \beta_k \frac{H_k r_k r_k^T H_k}{\langle H_k r_k, r_k \rangle},$$
where $\beta_k = 1 + \langle r_k, \delta_k \rangle / \langle H_k r_k, r_k \rangle$.

Quasi-Newton method usually terminates in n iterations. It also exhibits locally superlinear convergence rate, i.e., $\exists\, N$ s.t. for all $k \geq N$,

$$\|x_{k+1} - x^*\| \leq C \|x_k - x^*\| \|x_{k-n} - x^*\|.$$

However, its global convergence is not better than the gradient method.

5.12 Cubic Regularization*

In cubic-regularized Newton's method, we start from an arbitrary initial point x_0 and update the iteration x_k, $k = 0, 1, \ldots$ according to

$$s_{k+1} = \operatorname{argmin}_{s \in \mathbb{R}^n} \nabla f(x_k)^T s + \tfrac{1}{2} s^T \nabla^2 f(x_k) s + \tfrac{M}{6} \|s\|^3, \tag{5.12.1}$$
$$x_{k+1} = x_k + s_{k+1}, \tag{5.12.2}$$

where $M > 0$. Our main goal is to show that this method converges to a second-order stationary point x s.t. $\nabla f(x) = 0$ and $\nabla^2 f(x) \succeq 0$, and establish its rate of convergence. We will also show how to solve the subproblem in (5.12.1).

We assume that the objective function f satisfies the following assumptions.

a) f is twice-continuously differentiable and bounded below, i.e., $f^* := \inf_{x \in bbr^n} f(x) > -\infty$;
b) The Hessian $\nabla^2 f$ is L-Lipschitz continuous.

We need the following technical result. We use $\mathscr{S}^{n \times n}$ to denote the set of symmetric matrices.

Lemma 5.16. *Let* $M > 0$, $g \in \mathbb{R}^n$, $H \in \mathscr{S}^{n \times n}$, *and*

$$s = \operatorname{argmin}_{u \in \mathbb{R}^n} g^T u + \tfrac{1}{2} u^T H u + \tfrac{M}{6} \|u\|^3. \tag{5.12.3}$$

Then the following statements hold:

$$g + Hs + \tfrac{M}{2}\|s\|s = 0, \tag{5.12.4}$$

$$H + \tfrac{M}{2}\|s\|I \succeq 0, \tag{5.12.5}$$

$$g^T s + \tfrac{1}{2}s^T Hs + \tfrac{M}{6}\|s\|^3 \le -\tfrac{M}{12}\|s\|^3. \tag{5.12.6}$$

Proof. First (5.12.4) follows from the first-order necessary optimality condition of (5.12.3). The proof (5.12.5) is given by Proposition 1 of Nesterov and Polyak 06. We now (5.12.6).

$$\begin{aligned}
&g^T s + \tfrac{1}{2}s^T Hs + \tfrac{M}{6}\|s\|^3\\
&= (-Hs - \tfrac{M}{2}\|s\|s)^T s + \tfrac{1}{2}s^T Hs + \tfrac{M}{6}\|s\|^3\\
&= -\tfrac{1}{2}s^T(H + \tfrac{M}{2}\|s\|I)s - \tfrac{M}{12}\|s\|^3\\
&\le -\tfrac{M}{12}\|s\|^3,
\end{aligned}$$

where the first identity follows from (5.12.4) and the last inequality follows from (5.12.5). ■

To further explain, (5.12.4) corresponds to the first-order necessary optimality condition, (5.12.5) corresponds to the second-order necessary optimality condition but with a tighter form due to the specific form of this optimization problem, and (5.12.5) guarantees a sufficient decrease at this minimizer.

Theorem 5.15. *After k iterations, the sequence $\{x_i\}_{i\ge 1}$ generated by the cubic regularization method contains a point $\tilde{x}$ such that*

$$\|\nabla f(\tilde{x})\| \le \tfrac{C_1}{(k-1)^{2/3}} \ \textit{and}\ \nabla^2 f(\tilde{x}) \succeq -\tfrac{C_2}{(k-1)^{1/3}},$$

where $k > 1$, and C_1 and C_2 are universal constants.

Proof. Observe that

$$\begin{aligned}
f(x_{k+1}) - f(x_k) &\le \nabla f(x_k)^T s_{k+1} + \tfrac{1}{2}s_{k+1}^T \nabla f(x_k) s_{k+1} + \tfrac{L}{6}\|s_{k+1}\|^3\\
&\le -\tfrac{3M-2L}{12}\|s_{k+1}\|^3.
\end{aligned}$$

Summing up the above relation from 0 to $k-1$, we have

$$f(x_k) - f(x_0) \le -\tfrac{3M-2L}{12}\textstyle\sum_{i=1}^k \|s_i\|^3.$$

Let $m = \operatorname{argmin}_{i\in\{1,\dots,k\}}\|s_i\|^3$. We have

$$\|s_m\|^3 \le \tfrac{12(f(x_0)-f^*)}{(3M-2L)k}.$$

Observe that

$$\begin{aligned}\|\nabla f(x_m)\| &= \|\nabla f(x_m) - (\nabla f(x_{m-1}) + \nabla^2 f(x_{m-1})s_m + \tfrac{M}{2}\|s_m\|s_m\| \\ &\leq \|\nabla f(x_m) - (\nabla f(x_{m-1}) + \nabla^2 f(x_{m-1})s_m\| + \tfrac{M}{2}\|s_m\|^2 \\ &\leq \tfrac{L+M}{2}\|s_m\|^2.\end{aligned}$$

Moreover, by (5.12.5),

$$\nabla^2 f(x_m) \succeq -\tfrac{M}{2}\|s_m\|I.$$

The results then follow by combining these three observations. ■

5.13 Methods for function constraints

In this section, we provide an overview of different types of methods for solving nonlinear optimization problems given by

$$\begin{aligned}&\min f(x)\\ &\text{s.t. } g_j(x) \leq 0, j = 1,\ldots,m\\ &\quad h_i(x) = 0, i = 1,\ldots,k.\end{aligned} \tag{5.13.1}$$

Primal methods

These methods mimics unconstrained approaches, traveling along the feasible set in a way to ensure progress in objective at every step. The complexity of these methods have been studied for convex optimization. In particular, the following two directions in primal methods have been explored.

a) Polyak's alternative projection method. This algorithm alternatively moves along either the (sub)gradient direction of the objective or constraints, and project back to the feasible set X. This algorithm has been designed for general nonsmooth convex optimization and its compexity is bounded by $\mathscr{O}(1/\varepsilon^2)$.
b) Lemaréchal, Nesterov and Nemirovski's root-finding technique. For a given level estimate $l \in \mathbb{R}$, let us define

$$\begin{aligned}\phi(l) &:= \min_{x\in X} \max\{f(x) - l, g_1(x), \ldots, g_m(x)\}\\ &= \min_{x\in X} \max_{(\gamma,z)\in Z} \gamma[f(x) - l] + \sum_{i=1}^m z_i g_i(x).\end{aligned} \tag{5.13.2}$$

Here $Z := \{(\gamma, z) \in \mathbb{R}^{m+1} : \gamma + \sum_{i=1}^m z_i = 1, \gamma, z_i \geq 0\}$ denotes the standard simplex. We can easily verify that: (a) $\phi(l)$ is monotonically non-increasing and convex w.r.t. l; (b) $\phi(f^*) = 0$;

Lagrange Multiplier methods

These methods utilizes dual information of (5.13.1). Some dual methods (e.g., augmented Lagrangian method) reduce (5.13.1) to a sequence of unconstrained problems. Other primal-dual methods update the primal and dual variables at each iteration. Recently accelerated primal-dual methods that can achieve fast convergence for function constrained optimization have been studied by Digvijay, Deng and Lan (2019) and Zhang and Lan (2022). Some of these methods, especially those based on primal-dual approaches have been extended to nonconvex optimization with convex function constraints or even nonconvex optimization with nonconvex function constraints.

Penalty/Barrier methods

These methods reduce constrained minimization to a sequence of unconstrained problems

Sequential quadratic programming

These methods directly solves the KKT system associated with (P) by a kind of Newton method.

5.14 Excercises

Exercise 5.1. Suppose f is strongly convex with $mI \preceq \nabla^2 f(x) \preceq MI$. Let d be a descent direction at x. Show that the Inexact line search condition

$$f(x+td) \le f(x) + \alpha t \nabla f(x)^T d, \quad \alpha \in (0, 0.5),$$

holds for some

$$0 < t \le -\frac{\nabla f(x)^T d}{M\|d\|^2}.$$

Exercise 5.2. Let $f \in C_L^{2,2}(\mathbb{R}^n)$. Show that

$$\begin{aligned}&\|f'(y) - f'(x) - f''(x)(y-x)\| \le \tfrac{M}{2}\|y-x\|^2\\ &|f(y) - f(x) - \langle f'(x), y-x\rangle - \tfrac{1}{2}\langle f''(x)(y-x), y-x\rangle| \le \tfrac{M}{6}\|y-x\|^3.\end{aligned}$$

Exercise 5.3. Consider the optimization problem

$$\max_{x \in \mathbb{R}^n} f(x),$$

where f is differentiable and its gradient satisfies

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|^{\nu}, \ \ \forall x, y \in \mathbb{R}^n$$

for some $L > 0$ and $\nu \in (0,1]$. Given $x_0 \in \mathbb{R}^n$, we intend to solve this problem by using the gradient descent method

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k), \quad k \geq 0.$$

Please specify an appropriate selection of the stepsizes $\{\gamma_k\}$ which can guarantee the convergence of this algorithm to a stationary point. Also please show the rate of convergence of this method.

Exercise 5.4. Explain how to find a steepest descent direction in the l_∞-norm, and give a simple interpretation.

Exercise 5.5. The pure Newton's method with fixed stepsize 1 can diverge if the initial point is not close to x^*. In this problem we consider two examples.

a) $f(x) = \log(e^x + e^{-x})$ has a unique minimizer $x^* = 0$. Run the pure Newton method starting at $x_0 = 1$ and at $x_0 = 1.1$.
b) $f(x) = -\log x + x$ has a unique minimizer $x^* = 1$. Run the pure Newton method, starting at $x_0 = 3$.

Plot f and ∇f, and show the first few iterates.

Exercise 5.6. Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is increasing and convex, and $f : \mathbb{R}^n \to \mathbb{R}$ is convex, so $g(x) = \phi(f(x))$ is convex. (We assume that f and g are twice differentiable.) The problems of minimizing f and minimizing g are clearly equivalent. Compare the gradient method and Newton's method, applied to f and g. How are the search directions related? How are the methods related if an exact line search is used?

Exercise 5.7. Minimizing a quadratic function. Consider the problem of minimizing a quadratic function:

$$\text{minimize} \quad f(x) = (1/2)x^T P x + q^T x + r,$$

where $P \in \mathbf{S}^n$ (but we do not assume $P \succeq 0$).

(a) Show that if $P \not\succeq 0$, i.e., the objective function f is not convex, then the problem is unbounded below.
(b) Now suppose that $P \succeq 0$ (so the objective function is convex), but the optimality condition $Px^\star = -q$ does not have a solution. Show that the problem is unbounded below.

Exercise 5.8. Minimizing a quadratic-over-linear fractional function. Consider the problem of minimizing the function $f : \mathbf{R}^n \to \mathbf{R}$, defined as

$$f(x)=\frac{\|Ax-b\|_2^2}{c^Tx+d}, \quad \mathbf{dom} f=\left\{x \mid c^Tx+d>0\right\}.$$

We assume $\mathbf{rank} A=n$ and $b \notin \mathscr{R}(A)$.

(a) Show that f is closed.
(b) Show that the minimizer x^* of f is given by

$$x^\star=x_1+tx_2$$

where $x_1=\left(A^TA\right)^{-1}A^Tb, x_2=\left(A^TA\right)^{-1}c$, and $t\in\mathbf{R}$ can be calculated by solving a quadratic equation.

Exercise 5.9. Initial point and sublevel set condition. Consider the function $f(x)=x_1^2+x_2^2$ with domain $\mathbf{dom} f=\{(x_1,x_2) \mid x_1>1\}$.

(a) What is $p^\star$?
(b) Draw the sublevel set $S=\left\{x \mid f(x)\le f\left(x^{(0)}\right)\right\}$ for $x^{(0)}=(2,2)$. Is the sublevel set S closed? Is f strongly convex on S ?
(c) What happens if we apply the gradient method with backtracking line search, starting at $x^{(0)}$? Does $f\left(x^{(k)}\right)$ converge to $p^\star$?

Exercise 5.10. Do you agree with the following argument? The ℓ_1-norm of a vector $x\in\mathbf{R}^m$ can be expressed as

$$\|x\|_1=(1/2)\inf_{y\succ 0}\left(\sum_{i=1}^m x_i^2/y_i+\mathbf{1}^Ty\right).$$

Therefore the ℓ_1-norm approximation problem

$$\text{minimize } \|Ax-b\|_1$$

is equivalent to the minimization problem

$$\text{minimize} \quad f(x,y)=\sum_{i=1}^m\left(a_i^Tx-b_i\right)^2/y_i+\mathbf{1}^Ty, \tag{5.14.3}$$

with $\mathbf{dom} f=\{(x,y)\in\mathbf{R}^n\times\mathbf{R}^m \mid y\succ 0\}$, where a_i^T is the i th row of A. Since f is twice differentiable and convex, we can solve the ℓ_1-norm approximation problem by applying Newton's method.

Exercise 5.11. Backtracking line search. Suppose f is strongly convex with $mI\preceq\nabla^2 f(x)\preceq MI$. Let Δx be a descent direction at x. Show that the backtracking stopping condition holds for

$$0<t\le-\frac{\nabla f(x)^T\Delta x}{M\|\Delta x\|_2^2}.$$

Use this to give an upper bound on the number of backtracking iterations.

Exercise 5.12. Let $f : \Re^n \mapsto \Re$ be a given function.

(a) Consider a vector x^* such that f is convex over a sphere centered at x^*. Show that x^* is a local minimum of f if and only if it is a local minimum of f along every line passing through x^* [i.e., for all $d \in \Re^n$, the function $g : \Re \mapsto \Re$, defined by $g(\alpha) = f(x^* + \alpha d)$, has $\alpha^* = 0$ as its local minimum].

(b) Assume that f is not convex. Show that a vector x^* need not be a local minimum of f if it is a local minimum of f along every line passing through x^*. Hint: Use the function $f : \Re^2 \mapsto \Re$ given by

$$f(x_1, x_2) = \left(x_2 - px_1^2\right)\left(x_2 - qx_1^2\right),$$

where p and q are scalars with $0 < p < q$, and $x^* = (0,0)$. Show that $f\left(y, my^2\right) < 0$ for $y \neq 0$ and m satisfying $p < m < q$, while $f(0,0) = 0$.

Exercise 5.13. (Exact Penalty Functions)

Let $f : Y \mapsto \Re$ be a function defined on a subset Y of $\Re^n$. Assume that f is Lipschitz continuous with constant L, i.e.,

$$|f(x) - f(y)| \leq L\|x - y\|, \quad \forall x, y \in Y.$$

Let also X be a nonempty closed subset of Y, and c be a scalar with $c > L$.

(a) Show that if x^* minimizes f over X, then x^* minimizes

$$F_c(x) = f(x) + c \inf_{y \in X} \|y - x\|$$

over Y.

(b) Show that if x^* minimizes $F_c(x)$ over Y, then $x^* \in X$, so that x^* minimizes f over X.

Exercise 5.14. (Approximate Minima of Convex Functions)

Let X be a closed convex subset of $\Re^n$, and let $f : \Re^n \mapsto (-\infty, \infty]$ be a closed convex function such that $X \cap \mathrm{dom}(f) \neq \varnothing$. Assume that f and X have no common nonzero direction of recession. Let X^* be the set of global minima of f over X (which is nonempty and compact), and let $f^* = \inf_{x \in X} f(x)$. Show that:

(a) For every $\varepsilon > 0$ there exists a $\delta > 0$ such that every vector $x \in X$ with $f(x) \leq f^* + \delta$ satisfies $\min_{x^* \in X^*} \|x - x^*\| \leq \varepsilon$.

(b) If f is real-valued, for every $\delta > 0$ there exists a $\varepsilon > 0$ such that every vector $x \in X$ with $\min_{x^* \in X^*} \|x - x^*\| \leq \varepsilon$ satisfies $f(x) \leq f^* + \delta$.

(c) Every sequence $\{x_k\} \subset X$ satisfying $f(x_k) \to f^*$ is bounded and all its limit points belong to X^*.

Exercise 5.15. Gradient descent and nondifferentiable functions.

(a) Let $\gamma > 1$. Show that the function

$$f(x_1,x_2) = \begin{cases} \sqrt{x_1^2+\gamma x_2^2} & |x_2| \le x_1 \\ \dfrac{x_1+\gamma|x_2|}{\sqrt{1+\gamma}} & \text{otherwise} \end{cases}$$

is convex. You can do this, for example, by verifying that

$$f(x_1,x_2) = \sup\left\{x_1y_1 + \sqrt{\gamma}x_2y_2 \mid y_1^2+y_2^2 \le 1, y_1 \ge 1/\sqrt{1+\gamma}\right\}.$$

Note that f is unbounded below. (Take $x_2 = 0$ and let x_1 go to $-\infty$.)

(b) Consider the gradient descent algorithm applied to f, with starting point $x^{(0)} = (\gamma, 1)$ and an exact line search. Show that the iterates are

$$x_1^{(k)} = \gamma\left(\frac{\gamma-1}{\gamma+1}\right)^k, \quad x_2^{(k)} = \left(-\frac{\gamma-1}{\gamma+1}\right)^k.$$

Therefore $x^{(k)}$ converges to $(0,0)$. However, this is not the optimum, since f is unbounded below.

Exercise 5.16. Prove the lower complexity bound for strongly convex problems using the following worst case example.

$$f(x) = \tfrac{\mu(Q-1)}{4}\left[\tfrac{1}{2}\langle Ax,x\rangle - \langle e_1,x\rangle\right], \tag{5.14.4}$$

where $e_1 := (1,0,\ldots,0)$ and A is a symmetric matrix in $\mathbb{R}^{n\times n}$ given by

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\ 0 & 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & \kappa \end{pmatrix} \quad \text{with} \quad \kappa = \tfrac{\sqrt{Q}+3}{\sqrt{Q}+1}. \tag{5.14.5}$$

References