A Course on Optimization

Parin Chaipunya — KMUTT, BKK, TH.

Foreword

This lecture note covers important classical topics in both constrained and unconstrained optimization, each will be treated from the theoretical aspect following by algorithmic studies. This document is designed for senior bachelor and master students in mathematics.

Contents

§ 1. Introduction

An optimization problem consists of three main ingredients:

- a decision space X,
- a constraint set $C \subset X$, and
- an objective function $f : U \to \mathbb{R}$, where $U \subset X$ is nonempty,

and aims at finding a point $\overline{x} \in C$ such that $f(\overline{x}) \leq f(x)$ for all $x \in C$. This can be conventionally written as

$$Opt(f, C) \quad \begin{cases} \min & f(x) \\ \text{s.t.} & x \in C. \end{cases}$$

For the problem to make sense, we may require $C \cap U \neq \emptyset$. It does not hurt then to assume that $C \subset U$. We shall develop a shorthand notation $\overline{x} \in Opt(f, C)$ to denote that $\overline{x}$ is a solution to the optimization problem $Opt(f, C)$. Concerning with the problem $Opt(f, C)$, the point $x \in X$ is said to be *feasible* if it belongs to the constraint set C. If $C = X$, then any $x \in X$ is feasible and the optimization problem is said to be *unconstrained*. We shall write $Opt(f) := Opt(f, X)$ for an unconstrained optimization problem.

Notice that the constraint set C above can be just anything, hence the general form of $Opt(f, C)$ is said to be equipped with an *abstract constraint*. Of course, the constraint C can be described in a more explicit form, i.e. by inequalities and equalities:

$$C = \left\{ x \in X \;\middle|\; \begin{array}{ll} g_i(x) \leq 0, & \forall i = 1, 2, \cdots, r \\ h_j(x) = 0, & \forall j = 1, 2, \cdots, l \end{array} \right\}$$

where the (vector) functions $g = (g_1, g_2, \cdots, g_r) : X \to \mathbb{R}^r$ and $h = (h_1, h_2, \cdots, h_l) : X \to \mathbb{R}^l$ are given. In this case, $Opt(f, C)$ will be represented by $Opt(f, g, h)$ with

$$Opt(f, g, h) \quad \begin{cases} \min & f(x) \\ \text{s.t.} & g_i(x) \leq 0 \quad \forall i = 1, 2, \cdots, r \\ & h_j(x) = 0 \quad \forall j = 1, 2, \cdots, l. \end{cases}$$

There are chances that the optimization problem that we are facing has either inequality constraints alone with no inequalities or equality constraints alone without inequalities. In such cases, we may respectively set $h \equiv 0$ or $g \equiv -1$.

In this note, we will only consider the case where X is a Euclidean space $\mathbb{R}^n$. Be cautious that eventhough some results generalize directly into infinite-dimensional (Hilbert or Banach) spaces, many of them may not be so.

1.1 Some Examples

In this section, we take a look at some optimization problems. Of course, we may consider minimizing a function $f(x) = 2.5x^4 - x^3 + 5x^2 - 1$ over an equality constraint $x^2 + 2y^2 = 5$ and an inequality constraint $x + y \leq 1$, but this is too academic and far from real-world applications. So, let us give some motivating examples and also to observe how the problems may be formulated from practical situations.

Example 1.1 (Container design)**.** A soft drink manufacturer would like to produce a cylindrical can that would hold 330 mL of liquid. The manufacturer wants to decide the dimension of this can so that the material used to make this can is minimized.

Example 1.2 (Steel enhancement cost)**.** To enhance a certain type of steel to achieve endurance rate $E > 0$, we may add some additional materials $M_1, \cdots, M_n$. Each unit of the material M_i has an enhancement rate $E_i \geq 0$ and the processing cost c_i. The objective here is to minimize the processing cost while the steel is enhanced to the desired endurance E.

Example 1.3 (Portfolio optimization)**.** A man has \$20,000 worth for investment and he is interested to invest in four different financial instruments as follows.

Choice 1. Buy *stock X* which is selling at \$20 per share.

Choice 2. Purchase a *European call* options to buy a share of stock X at \$15 in exactly 6 months time. The options are selling for \$10.

Choice 3. Raise more funds for investment by selling the European call options above.

Choice 4. Purchase a 6-month riskless zero-coupon bonds having a face value of \$100 at the price of \$90.

This man has determined that there are three *equally likely* scenarios that may occur to the stock X, namely

Scenario 1. Stock X sells for \$20 per share in 6 months.

Scenario 2. Stock X sells for \$40 per share in 6 months.

Scenario 3. Stock X sells for \$12 per share in 6 months.

Due to the risks involved, there is a margin on the total number of European call options that you can sell, in this case set to 500. Also, a person is limited to the maximum of 5000 calls.

The aim of this man is to make an investment plan for the available choices.

Example 1.4 (A linear system as an optimization problem)**.** Solving a linear system $Ax = b$ for $x \in \mathbb{R}^n$, where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, may exploits the unconstrained optimization structure of the objective function $f(x) = \frac{1}{2}x^\top Ax - b^\top x$.

Example 1.5 (Linear regression)**.** Suppose that we have collected some experimental data $y_1, \cdots, y_m$ at each initial variable $x_1, \cdots, x_m \in \mathbb{R}^n$, where it is known that y_i depends linearly on x_i. Then the linear regression aims at finding the best affine approximation $r(x) = \langle a, x\rangle + b$ to all y_i's, where $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ represents the gradient and offset of the affine approximation.

1.2 Existence of a Minimizer

The existence of a minimizer is the first question that we would investigate, and it is really simple and general. Suppose throughout the section that $f : \mathbb{R}^n \to \mathbb{R}$.

Theorem 1.6 (Weierstraß theorem)**.** *Let f be a lsc function and C is a nonempty compact subset of $\mathbb{R}^n$. Then $Opt(f, C)$ has a solution.*

Proof. Let (x_n) be a sequence in C for which $f(x_n) \to \inf_C f$. Since C is compact, we may assume that (x_n) is convergent to some $x^* \in C$. Using the lower semicontinuity of f, we get $\inf_C f = \liminf_n f(x_n) \geq f(x^*)$. This shows that $\inf_C f$ is finite and that x^* is a minimizer of f over C. ∎

The existence of a constrained solution above may be adopted to the unconstrained problem too.

Corollary 1.7. *Let f be a lsc function in which some sublevel set that is nonempty and compact. Then $Opt(f)$ has a solution.*

Corollary 1.8. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a lsc function in which there is a nonempty bounded sublevel set. Then $Opt(f)$ has a solution.*

The uniqueness of a minimizer is usually not achievable and most of the time not of interests. However, in the case where the objective function is strictly convex, we may obtain the uniqueness of a minimizer.

1.3 Global vs. Local Solutions

All the solutions of $Opt(f, C)$ described earlier are known more specifically as the *global* solution. In some context, we may not be previleged enough to be able to obtain the global solution. Hence we may also talk about the *local* solution concept. A feasible point $\overline{x} \in C$ is called a *local solution* to $Opt(f, C)$ if there is a neighborhood N of $\overline{x}$ such that $\overline{x} \in Opt(f, C \cap N)$. We write $\overline{x} \in LOpt(f, C)$ to denote that $\overline{x}$ is a local solution of $Opt(f, C)$. Obviously, $Opt(f, C) \subset LOpt(f, C)$.

We can also talk about the solution in a strict sense: A point $x^* \in C$ is said to be a *strict* solution of $Opt(f, C)$ if $f(x^*) < f(x)$ for all $x \in C$. The strict local solution is defined likewise.

1.4 Some Calculus

Always let $f : U \subset \mathbb{R}^n \to \mathbb{R}$, where U is an open set. Let us recall some differential calculus regarding such a function f.

Definition 1.9. Let $x \in U$ and $d \in \mathbb{R}^n$. The *directional derivative of f at x in the direction d* is defined by

$$f'(x; d) := \lim_{t \to 0+} \frac{f(x + td) - f(x)}{t},$$

if the above limit exists.

Clearly, we have $f'(x; \alpha d) = \alpha f'(x; d)$ for $\alpha \geq 0$. If $f'(x; -d) = -f'(x; d)$, then we have $f'(x; d) = \lim_{t \to 0} \frac{f(x+td)-f(x)}{t}$.

Definition 1.10. A function f is said to be *Gâteaux differentiable at $x \in U$* if the directional derivatives $f'(x; d)$ exist for all directions $d \in \mathbb{R}^n$ and is linear in d. That is, there exists a linear function $df_x : \mathbb{R}^n \to \mathbb{R}$, called the *Gâteaux derivative* of f at x, in which

$$df_x(d) = f'(x; d)$$

for all $d \in \mathbb{R}^n$. We say that f is *Gâteaux differentiable* if it is so for all $x \in U$.

The Gâteaux differentiability is closely related to the existence of the gradient ∇f. Recall that, for each $i = 1, \cdots, n$, the partial derivative $\frac{\partial f}{\partial x_i}(x)$ is defined by $f'(x; e_i)$ where e_i is the vector with 1 in the ith entry and the remainig entries are 0. The *gradient of* f *at* x is $\nabla f(x) := \left[\frac{\partial f}{\partial x_1}(x) \cdots \frac{\partial f}{\partial x_n}(x)\right]^\top$.

Proposition 1.11. *If f is Gâteaux differentiable at x, then $\nabla f(x)$ exists and $df_x(d) = \langle \nabla f(x), d \rangle = \nabla f(x)^\top d$.*

Proof. Left as a student's exercise. ∎

Another useful result for Gâteaux differentiable function is the following version of mean value theorem.

Theorem 1.12. *Let $f : U \to \mathbb{R}$ be a Gâteaux differentiable function. If $x, y \in U$ are distinct points such that $[x, y] \subset U$, then there is a point $z \in]x, y[$ such that*

$$f(y) = f(x) + \langle \nabla f(z), y - x \rangle.$$

Nexy, let us consider a stronger notion of differentiability.

Definition 1.13. The function f is said to be *Fréchet differentiable at* $x \in U$ if there exists a linear function $\ell : \mathbb{R}^n \to \mathbb{R}$, called the *Fréchet derivative at* x (written $\langle \ell, x \rangle = \ell(x)$), such that

$$\lim_{\|h\| \to 0} \frac{f(x+h) - f(x) - \langle \ell, h \rangle}{\|h\|} = 0.$$

We say that f is *Fréchet differentiable* if it is so at all $x \in U$.

In Landau's little "oh" notation, we may write the Fréchet differentiability as

$$f(x+h) = f(x) + \langle l, h \rangle + o(\|h\|).$$

Theorem 1.14. *If f is Fréchet differentiable at $x \in U$, then f is continuous at x and is Gâteaux differentiable. Moreover, the Fréchet derivative ℓ is identical to df_x (and hence to ∇f).*

The students are invited to find an example of a function that is (a) Gâteaux differentiable but not continuous, and (b) Gâteaux differentiable but not Fréchet differentiable. In fact, the converse can be achievable from the conitnuous differentiability. Hence, in finite-dimensional setting, the Fréchet differentiability is vert relevant and will always be adopted when higher differentiability of f is assumed.

Theorem 1.15. *Let f be Gâteaux differentiable at x and all the partial derivatives are continuous at x. Then f is Fréchet differentiable at x.*

In all cases, if f is differentiable (either Gâteaux or Fréchet), then each $x \in U$ is assigned to a linear function df_x. This can be viewed as a map $df : U \to L(\mathbb{R}^n, \mathbb{R})$ that assigns x with df_x, where $L(\mathbb{R}^n, \mathbb{R})$ is the Banach space of all linear functions from $\mathbb{R}^n$ into $\mathbb{R}$. Note, in this setting, that $L(\mathbb{R}^n, \mathbb{R})$ can be identified by $\mathbb{R}^n$ (df_x by $\nabla f(x)$). Hence we use the identification $df : U \to \mathbb{R}^n$ through $df(x) = df_x \equiv \nabla f(x)$.

Definition 1.16. We say that f is continuously differentiable if it is *Fréchet differentiable* and its derivative df is continuous.

Now, we may adopt a similar differentiation approach to a vector function $g = (g_1, \cdots, g_m) : U \to \mathbb{R}^m$ by representing the derivative dg_x with the Jacobian matrix $Jg(x) = [\nabla g_1(x) \cdots \nabla g_m(x)]^\top$. We may state formally as follows.

Definition 1.17. A vector function g defined as above is called *Fréchet differentiable at $x \in U$* if there exists a linear map $dg_x : \mathbb{R}^n \to \mathbb{R}^m$, called the *Fréchet derivative at* x, satisfying

$$\lim_{\|h\| \to 0} \frac{\|g(x+h) - g(x) - dg_x h\|}{\|h\|} = 0.$$

If such differentiability holds for all $x \in U$, then we say that g is *Fréchet differentiable.*

Similar calculation as in the real-valued case may be carried out to show that $dg_x d = Jg(x)d = [\nabla g_1(x) \cdots \nabla g_m(x)]^\top d$ for all $d \in \mathbb{R}^n$. Moreover, breaking down the norm in the numerator, we see that g is Fréchet differentiable at x if and only if each coordinate function g_i is.

Now, let us rewind back and consider a scalar function $f : U \to \mathbb{R}$. Since its derivative is a vector function $df : U \to \mathbb{R}^n$, we may consider taking another Fréchet derivative of df, that is $d^2 f := d(df)$. Since we have $df = \left(\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_n}\right)$, the Jacobian argument suggests the Hessian matrix

$$\nabla^2 f := \left[\frac{\partial^2 f}{\partial x_j \partial x_i}\right]_{i,j=1,\cdots,n} = \left[\nabla\left(\frac{\partial f}{\partial x_1}\right) \cdots \nabla\left(\frac{\partial f}{\partial x_n}\right)\right]^\top$$

to be the representation of d^2f. When taking d^2f at a point $x \in U$, the result d^2f_x is of course is a map from U into $L(L(\mathbb{R}^n, \mathbb{R}), \mathbb{R}) \equiv L(\mathbb{R}^n \times \mathbb{R}^n, \mathbb{R}) \equiv \mathbb{R}^{n \times n}$. Hence d^2f can be viewed as a map from U into $\mathbb{R}^{n\times n}$. If this map is continuous, then we say that f is *twice continuously differentiable.* This is equivalent to saying that all component functions, in this case the second-order partial derivatives, are continuous. It is further implied that $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$ for all $i, j = 1, \cdots, n$, so that $\nabla^2 f$ is symmetric. Note that $d^2 f_x$ can further be identified as a bilinear function accepting two variables u and v from $\mathbb{R}^n$. We may have figured that

$$d^2 f_x(u, v) = u^\top \nabla^2 f(x) v$$

for $u, v \in \mathbb{R}^n$.

Finally, we recall the Taylor's approximation for a scalar function in terms of ∇f and $\nabla^2 f$.

Theorem 1.18 (Taylor's approximation)**.** *Let f be twice continuously differentiable. Then for any $x \in U$, $t > 0$ and $h \in \mathbb{R}^n$ with $x + th \in U$, it holds*

$$f(x + th) = f(x) + t\langle \nabla f(x), h\rangle + \frac{t^2}{2}\langle h, \nabla^2 f(x)h\rangle + o(t^2).$$

The following version Mean Value Theorem also plays an important role in our analysis.

Theorem 1.19. *Let f be twice continuously differentiable, $x, y \in U$ be two distinct points with $[x, y] \subset U$. Then there exists a point $z \in]x, y[$ such that*

$$f(y) = f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\langle y - x, \nabla^2 f(z)(y - x)\rangle.$$

1.5 Convexity

For any $x, y \in \mathbb{R}^n$, the line segment joining them is defined by

$$[x, y] := \{(1 - t)x + ty \mid t \in [0, 1]\}.$$

We write $[x, y[$, $]x, y]$ and $]x, y[$ to denote the line segments that exclude the point y, x, and both ends, respectively.

A subset $C \subset \mathbb{R}^n$ is said to be *convex* if $[x, y] \subset C$ for all $x, y \in C$. A function $f : U \to \mathbb{R}$ is said to be *convex* on a convex set $C \subset U$ if the inequality

$$f((1 - \lambda)x + \lambda y) \le (1 - \lambda)f(x) + \lambda f(y) \tag{1.1}$$

holds for any $x, y \in C$ and any $\lambda \in [0,1]$. If f is convex on its domain U (implicitly implied that U is convex), we simply say that f is convex, without specifying the set C. If the inequality (1.1) holds strictly for all $x, y \in C$ with $x \neq y$ and all $\lambda \in]0,1[$, we say that f is *strictly convex* on C.

We may develop a geometric description for a convex function as follows.

Theorem 1.20. *f is convex if and only if its epigraph*

$$\operatorname{epi} f := \{(x,t) \in U \times \mathbb{R} \,|\, f(x) \leq t\}$$

is a convex set in $\mathbb{R}^{n+1}$.

Proof. The proof will be done in class. ∎

Let us now discuss what convexity provides in terms of optimization.

Theorem 1.21. *If $f : U \to \mathbb{R}$ is a convex function and C is a convex set, then $LOpt(f,C) = Opt(f,C)$.*

Proof. The proof will be done in class. ∎

Theorem 1.22. *If $f : U \to \mathbb{R}$ is a strictly convex function and C is a convex set, then $Opt(f,C)$ contains at most one point.*

Proof. The proof will be done in class. ∎

In the following, we characterize the convexity of a function f with its gradient ∇f and Hessian $\nabla^2 f$.

Theorem 1.23. *If $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, then f is convex if and only if the subgradient inequality*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \tag{1.2}$$

holds for all $x, y \in \mathbb{R}^n$.

Proof. (Only if) Take $x, y \in U$ and $t \in (0,1)$. Then

$$f(x + t(y-x)) = f((1-t)x + ty) \leq (1-t)f(x) + tf(y).$$

Rearrangement yields

$$\begin{aligned} f(y) &\geq \frac{f(x+t(y-x))-(1-t)f(x)}{t} \\ &= \frac{f(x+t(y-x))-f(x)+tf(x)}{t} \\ &= \frac{f(x+t(y-x))-f(x)}{t}+f(x). \end{aligned}$$

Letting $t \to 0+$, we get

$$f(y) \geq f(x) + \lim_{t\to 0+} \frac{f(x+t(y-x))-f(x)}{t} = f(x) + \langle \nabla f(x), y-x\rangle.$$

The (If) part is left as an exercise. ■

If f is twice continuously differentiable, we may characterize its convexity with the positivity of its Hessian. Recall first that a matrix $A \in \mathbb{R}^{n\times n}$ is said to be positive semidifinite (resp. definite) if A is symmetric and $d^\top Ad \geq 0$ (resp. $d^\top Ad > 0$) for all $d \in \mathbb{R}^n$. We adopt the notation $A \preceq B$ and $A \prec B$, for any $A, B \in \mathbb{R}^{n\times n}$, whenever $B - A$ is positive semidefinite and positive definite, respectively. In such cases, we sometimes write $B \succeq A$ and $B \succ A$ to emphasize on B. If eigenvalues are accessible for a given matrix A, the positivity of A can be easily checked.

Theorem 1.24. *If A is a symmetric matrix, then*

(i) A is positive semidefinite if and only if all its eigenvalues are non-negative.

(ii) A is positive definite if and only if all its eigenvalus are strictly positive.

Theorem 1.25. *Let f be a twice continuously differentiable function. Then*

(i) f is convex if and only if $\nabla^2 f(x) \succeq 0$ for all $x \in \mathbb{R}^n$, and

(ii) f is strictly convex if and only if $\nabla^2 f(x) \succ 0$ for all $x \in \mathbb{R}^n$

Proof. Let us prove (i) only, as (ii) is proved similarly.

(Only if) Let f be convex and take any $d \in \mathbb{R}^n$. It follows from the previous theorem and Taylor's approximation that

$$f(x) + t\langle \nabla f(x), d\rangle \leq f(x+td) = f(x) + t\langle \nabla f(x), d\rangle + \frac{t^2}{2}\langle d, \nabla^2 f(x) d\rangle + o(t^2).$$

This gives $\langle d, \nabla^2 f(x)d\rangle + o(t^2)/t^2 \geq 0$. The result follows by taking $t \to 0+$.

(If) Conversely, suppose that $\nabla^2 f(x) \succeq 0$ at all $x \in U$. Take any $x, y \in U$ with $x \neq y$. Trom the Mean Value Theorem 1.19, there exists a point $z \in]x, y[$ such that

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle y - x, \nabla^2 f(z)(y - x) \rangle \geq f(x) + \langle \nabla f(x), y - x \rangle.$$

The Theorem 1.23 implies that f is convex. ∎

The second-order characterization above suggests an approach to further strengthen the convexity.

Definition 1.26. A twice continuously differentiable function $f : U \to \mathbb{R}$ is said to *strongly convex* on C with a constant $m > 0$ if $\nabla^2 f(x) \succeq mI$ for all $x \in U$, where I denotes the identity matrix.

Theorem 1.27. *A twice continuously differentiable function is strongly convex on a convex set C with a constant $m > 0$ if and only if $\lambda_{\min}(x) \geq m$ for all $x \in C$, where $\lambda_{\min}(x)$ denotes the minimum eigenvalue of $\nabla^2 f(x)$. In particular, a strongly convex function is strictly convex.*

We also get a sharper subgradient inequality. At the same time, the inequality appeared in the theorem states that a strongly convex function can be supported from below by a quadratic function.

Theorem 1.28. *If $f : U \to \mathbb{R}$ is strongly convex with a constant $m > 0$, then for any $x, y \in U$, the following inequality holds:*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2} \|y - x\|^2.$$

Proof. Let $x \in U$ and $d \in \mathbb{R}^n$. We have

$$u^\top \nabla^2 f(x) u \geq m \|u\|^2.$$

Let $y \in U$ be different from x. Using the Mean Value Theorem 1.19 together with the above inequality, there exists $z \in]x, y[$ such that

$$\begin{aligned} f(y) &= f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} (y - x)^\top \nabla^2 f(z)(y - x) \\ &\geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2} \|y - x\|^2. \end{aligned}$$ ∎

Another way to relate a strongly convex function with its quadratic support is as follows.

Theorem 1.29. *A function f is strongly convex with a constant $m > 0$ if and only if $f - \frac{m}{2}\|\cdot\|^2$ is convex. In particular, if f is strongly convex with a constant $m > 0$, we have*

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y) - \frac{m}{2}\lambda(1-\lambda)\|x-y\|^2$$

for all $x, y \in U$.

Proof. Let g be a function on U defined by $g := f - \frac{m}{2}\|\cdot\|^2$. Notice that if $\nabla^2 f$ exist for all $x \in U$, then the same holds for $\nabla^2 g$ with $\nabla^2 g = \nabla^2 f - mI$. The result immediately follows from the definition of strong convexity.

Next, suppose that f is strongly convex with a constant $m > 0$. Using the Apollonius identity and the convexity of $f - \frac{m}{2}\|\cdot\|^2$ yield

$$\begin{aligned} f((1-\lambda)x + \lambda y) &- \frac{m}{2}[(1-\lambda)\|x\|^2 + \lambda\|y\|^2 - \lambda(1-\lambda)\|x-y\|^2] \\ &= f((1-\lambda)x + \lambda y) - \frac{m}{2}\|(1-\lambda)x + \lambda y\|^2 \\ &\leq (1-\lambda)f(x) + \lambda f(y) - (1-\lambda)\frac{m}{2}\|x\|^2 - \lambda\frac{m}{2}\|y\|^2, \end{aligned}$$

and we have arrived at the desired conclusion. ∎

This leads to the following elegant result.

Theorem 1.30. *Suppose that $f : U \to \mathbb{R}$ is a strongly convex function (with a constant $m > 0$) and there exists $x \in U$ in which $S_{f(x)}$ is closed. Then f has a unique minimizer.*

Proof. Let x be as in the hypothesis and take $y \in S_{f(x)}$. Then

$$\begin{aligned} 0 &\geq f(y) - f(x) \\ &\geq \langle \nabla f(x), y - x \rangle + \frac{m}{2}\|y-x\|^2 \\ &\geq -\|\nabla f(x)\|\|y-x\| + \frac{m}{2}\|y-x\|^2. \end{aligned}$$

This implies $\|\nabla f(x)\| \geq \frac{m}{2}\|y-x\|$, so that

$$\begin{aligned} \operatorname{diam} S_x &= \sup_{y,z \in S_{f(x)}} \|y-z\| \\ &\leq \sup_{y \in S_{f(x)}} \|y-x\| + \sup_{z \in S_{f(x)}} \|z-x\| \\ &\leq \frac{4}{m}\|\nabla f(x)\| < \infty. \end{aligned}$$

Hence $S_{f(x)}$ is bounded and nonempty. Since $S_{f(x)}$ is also closed, it is compact. The existence of a minimizer follows from Corollary 1.7, while the uniqueness follows from the strict convexity (see Theorem 1.22). ∎

Note that if $U = \mathbb{R}^n$, then $S_{f(x)}$ is closed for all $x \in U$ if f is strongly convex. This leads to our final corollary.

Corollary 1.31. *A strongly convex function $f : \mathbb{R}^n \to \mathbb{R}$ has a unique minimizer.*

§ 2. Unconstrained Optimization Problems – Theory

Unless otherwise specified, always assume that $X = \mathbb{R}^n$ equipped with the Euclidean norm $\|\cdot\|$. Suppose that $U \subset \mathbb{R}^n$ is nonempty and open, and $f : U \to \mathbb{R}$ is a given objective function.

2.1 Optimality Conditions

Ironically, we have no knowledge of how to solve $Opt(f)$ directly from its definition. All the existed methods rely on transforming $Opt(f)$ into some other forms that we feel more confortable with. In this case, we may *cautiously* turn $Opt(f)$ into a critical point equation $\nabla f(x) = 0$. Note that the two problems are not generally equivalent. With the following theorem, we may see that a critical point equation for ∇f is a relaxation of $Opt(f)$.

Theorem 2.1 (First-Order Necessary Optimality Condition). *Suppose that f is continuously differentiable, then*

$$\overline{x} \in LOpt(f) \implies \nabla f(\overline{x}) = 0.$$

Proof. To be proved in class. ∎

Theorem 2.2 (Second-Order Necessary Optimality Condition). *Suppose that f is twice continuously differentiable, then*

$$\overline{x} \in LOpt(f) \implies \nabla^2 f(\overline{x}) \succeq 0.$$

Proof. Let $\overline{x} \in LOpt(f)$. We know that $\nabla f(\overline{x}) = 0$ and for any $h \in \mathbb{R}^n$, $f(\overline{x} + th) \geq f(\overline{x})$ for $t \in \mathbb{R} \setminus \{0\}$ sufficiently close to 0. Using the Taylor's approximation (Theorem 1.18), we have

$$\frac{t^2}{2}\langle h, \nabla^2 f(\overline{x})h\rangle + o(t^2) \geq 0.$$

Dividing both sides by t^2 and letting $t \to 0$ yield

$$h^\top \nabla^2 f(\overline{x}) h \geq 0,$$

and the proof is finished. ∎

To have the sufficiency, we need the (local) convexity of f through the positivity of the Hessian.

Theorem 2.3 (Second-Order Sufficient Optimality Condition)**.** *Suppose that f is twice continuously differentiable and $\nabla f(\overline{x}) = 0$. If $\nabla^2 f(\overline{x}) \succ 0$, then $\overline{x} \in LOpt(f)$.*

Proof. Fix $h \in \mathbb{R}^n$. Since $\nabla^2 f$ is continuous and $h^\top \nabla^2 f(\overline{x})h > 0$, then there is a neighborhood $N \subset U$ of $\overline{x}$ in which $h^\top \nabla^2 f(x)h > 0$ for all $x \in N$. The Mean Value Theorem 1.19 implies that for any $x \in N \setminus \{\overline{x}\}$, there exists $z_x \in]\overline{x}, x[$ in which

$$f(x) = f(\overline{x}) + \frac{1}{2} h^\top \nabla^2 f(z_x) h \geq f(\overline{x}).$$

This shows that $\overline{x} \in LOpt(f)$. ∎

Discussion 2.4. Is it possible to replace $\nabla^2 f(\overline{x} \succ 0)$ with $\nabla^2 f(\overline{x}) \succeq 0$ in the above theorem ? Discuss in general terms and from the theoretical perspective from the above proof.

Theorem 2.5 (Sufficient Optimality Condition for Convex Functions)**.** *Suppose that f is continuously differentiable and convex, then*

$$\nabla f(\overline{x}) = 0 \implies \overline{x} \in Opt(f).$$

Proof. Follows from the subgradient inequality 1.2. ∎

If one wish to find a (local) maximum instead of minimum, the criterion can easily be reverted and the sufficient condition for such case would be $\nabla^2 f(\overline{x}) \prec 0$ at a critical point $\overline{x} \in U$. A more complicate question is how to classify a critial point $\overline{x}$ that is neither a local mimimzer nor a local maximizer. Such a critical point is said to be *saddle.*

Theorem 2.6. *Let f be twice continuously differentiable and $\nabla f(\overline{x}) = 0$. If $\nabla^2 f(\overline{x})$ is indefinite (i.e. having both positive and negative eigenvalues), then $\overline{x}$ is a saddle point.*

Proof. Take a positive eigenvalue λ of $\nabla^2 f(\overline{x})$ and let $v \in \mathbb{R}^n$ be an associated normalized eigenvector. Let $r > 0$ be sufficiently small so that $\overline{x} + rv \in U$. Using Taylor's approximation (Theorem 1.18) and the fact that $\|v\| = 1$, we get

$$f(\overline{x} + tv) = f(\overline{x}) + \frac{t^2}{2} v^\top \nabla^2 f(\overline{x}) v + o(t^2) = f(\overline{x}) + \frac{\lambda t^2}{2} + o(t^2)$$

for $t \in (0, r)$. Note that we may squeeze $t > 0$ sufficiently small so that $o(t^2) > -\frac{\lambda t^2}{2}$. Combining this with the above approximation, we get

$$f(\overline{x} + tv) > f(\overline{x})$$

for small $t > 0$. This means $\overline{x}$ is not a local maximizer.

The proof can be repeated for a negative eigenvalue to ensure that $\overline{x}$ is not a local minimizer. ∎

Discussion 2.7. What happens a linear function is minimized without a constraint?

2.2 Quadratic functions

Quadratic functions deserves a special attention in optimization theory. This is because the computation of its gradient and Hessian is a light work, and it carries good numerical behaviors. Moreover, in view of Taylor's approximation, any twice continuously differentiable function can be very accurately approximated with a quadratic function. Due to these joyful attributes, many applications were formulated as minimizing an appropriate quadratic functions.

Definition 2.8. A *quadratic function* over $\mathbb{R}^n$ is a function $f : \mathbb{R}^n \to \mathbb{R}$ of the form

$$f(x) = x^\top Ax + b^\top x + c,$$

where $A \in \mathbb{R}^{n \times n}$ is symmetric, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$.

Proposition 2.9. *Let f be a quadratic function given as in the above definition. Then for any $x \in \mathbb{R}^n$, $\nabla f(x) = 2Ax + b$ and $\nabla^2 f(x) = 2A$. In particular, it is twice continuously differentiable and its convexity can be determined from A.*

Proof. The proof is done by direction calculations. ∎

The following proposition is also a direct computation.

Proposition 2.10. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a quadratic function given by*

$$f(x) := \frac{1}{2} x^\top Ax - b^\top x \qquad \text{for all } x \in \mathbb{R}^n.$$

Then

(i) $\nabla f(x) = 0$ *if and only if* $Ax = b$*,*

(ii) if $A \succeq 0$*, then* $\overline{x} \in Opt(f) \iff A\overline{x} = b$*,*

(iii) if $A \succ 0$*, then* A *is invertible and* $\overline{x} := A^{-1}b \in Opt(f)$*.*

In the above problem, the function f has a minimizer if and only if the system $Ax = b$ is consistent (at least one solution exists). Let us complement the study with the inconsistent case.

Example 2.11 (Overdetermined linear systems)**.** When the system is inconsistent (no solution), one may look for a point x in which Ax best approximates the target b, i.e. minimizing $\|Ax - b\|$. Usually we consider minimizing $\frac{1}{2}\|Ax - b\|^2$ instead, as it is quadratic, smooth and has the same minimizer as $\|Ax - b\|$.

Let us now state the formulation precisely. Consider a linear system

$$Ax = b,$$

where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ with $m > n$ and that $\operatorname{rank}(A) = n$ (i.e. overdetermined). Define a function $f : \mathbb{R}^n \to \mathbb{R}$ by

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

for $x \in \mathbb{R}^n$. To realize that this is quadratic, observe that

$$f(x) = \frac{1}{2}\|Ax - b\|^2 = \frac{1}{2}(Ax - b)^\top(Ax - b) = \frac{1}{2}x^\top(A^\top A)x - b^\top Ax + \frac{1}{2}\|b\|^2,$$

for all $x \in \mathbb{R}^n$. Since $A^\top A$ is symmetric, the function f is quadratic. We may compute ∇f and $\nabla^2 f$ following the earlier discussion. One may then observe the convexity of f.

Example 2.12 (Linear regression and data fitting)**.** We have described before the formulation of linear regression. If the set $\{(x_i, y_i)\}_{i=1,\cdots,m}$ represents the collection of known samples, where y_i's depend linearly on x_i's. Suppose that x_i's belong to $\mathbb{R}^n$ and y_i's belong to $\mathbb{R}$. We are looking for $a \in \mathbb{R}^n$ in which $a^\top x_i \approx y_i$ for all $i = 1, \cdots, m$. We do so my minimizing the mean square of the difference between $a^\top x_i$ and y_i, i.e. $\sum_i (a^\top x_i - y_i)^2 = \sum_i (x_i^\top a - y_i)^2 = \|Xa - y\|^2$ where $X = [x_1 \cdots x_m]^\top$ and $y = [y_1 \cdots y_m]^\top$. The optimization problem is then

$$\min_{a\in\mathbb{R}^n} \|Xa - y\|^2.$$

In fact, a linear regression problem allows the measurements y_i's to depend linearly on x_i's. This case is in fact included readily in such a linear setting above (how?).

Example 2.13 (Denoising: quadratic regularization)**.** Suppose that a measurement $b \in \mathbb{R}^n$ was captured from an original data $x \in \mathbb{R}^n$. Usually, the measurement b is noisy and $b = x + n$, for some $n \in \mathbb{R}^n$ representing an unknown noise vector. To remove the noise we try to minimize n, but then we are minimizing the data fitting term $\|b - x\|$ where we will naturally get $x = b$, which is still noisy.

To over come this, we need to exploit some *a priori* information that the original data x is *smooth* in the sense that x_i and x_{i+1} are relatively close. Adding this smoothing term into the objective, we get the objective function $f : \mathbb{R}^n \to \mathbb{R}$ defined for each $x \in \mathbb{R}^n$ by

$$f(x) := \underbrace{\|x - b\|^2}_{\text{data fitting term}} + \underbrace{\lambda \sum_{i=1}^{n-1} |x_i - x_{i+1}|^2}_{\text{smoothing term}} = \|x - b\|^2 + \lambda \|Sx\|^2,$$

where $\lambda > 0$ is regularization weight and $S \in \mathbb{R}^{(n-1)\times n}$ is a matrix given by

$$S := \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix}.$$

§ 3. Unconstrained Optimization Problems – Algorithms

Always suppose that $f : \mathbb{R}^n \to \mathbb{R}$. A general scheme used in searching for a minimizer takes the following form.

Algorithm 3.1. General search scheme.

```
Initialization:
    Pick a start point $x^0 \in \mathbb{R}^n$.
    Set $k \leftarrow 0$.
While: $\nabla f(x^k) \neq 0$;
    Choose $t_k > 0$ and $d^k \in \mathbb{R}^n$.
    Compute $x^{k+1} \leftarrow x^k + t_k d^k$.
    Update $k \leftarrow k + 1$.
```

In the above setting, the vector d^k is called the *search direction* and the positive scalar t_k is known as the step length (or step size). The main question is: how to come up with how the directions d^k and step lengths t_k should be determined so that $x^k \to x^* \in Opt(f)$, or at least $f(x^k) \to \inf f$?

Usually, we require d^k to be a *descent* direction, i.e. decrease the value of the objective function f. However, to define such decrease by comparing $f(x^k)$ and $f(x^{k+1})$ may be impractical since x^{k+1} also depends on the step length t_k. To avoid such drawback and state such a decrease purely with the current position x and the direction d, we may use the following definition.

Definition 3.2. A vector $d \in \mathbb{R}^n$ is called a *descent* direction at $x \in \mathbb{R}^n$ if $f'(x; d) < 0$.

The terminology "descent" is used here since if d is a descent direction, then one can see from the definition of $f'(x; d)$ that there is $\varepsilon > 0$ such that $f(x + td) < f(x)$ for all $t \in (0, \varepsilon)$. This has to be used with caution since here is generally no guarantee that $f(x + td) < f(x)$ would hold for any $t > 0$. In fact, it is most likely that $f(x + td) > f(x)$ when $t > 0$ is chosen too large, i.e. overshooting may occur.

There are then several approaches to determine the step sizes t_k's at each iteration. Suppose that the point x^k has been computed and the gradient descent direction $d^k \neq 0$ is used. We shall list the following few of the most famous techniques.

a) **Constant step-size rule.** Take some $\bar{t} > 0$ and let $t_k := \bar{t}$ for all $k \in \mathbb{N}$.

b) **Vanishing step-size rule.** Choose *a priori* a decreasing positive real sequence (t_k) in which $t_k \to 0$.

c) **Exact linesearch rule.** For each $k \in \mathbb{N}$, take $t_k = \arg\min_{t>0} f(x^k + td^k)$.

d) **Armijo's backtracking linesearch rule.** Choose *a priori* an acceptable rate of descent $\alpha \in (0, 1)$, an initial step length $s > 0$, and a decremental ratio $\beta \in (0, 1)$. For each $k \in \mathbb{N}$, define $t_k := \beta^i s$ (here β is raised to the power i), where $i \in \mathbb{N} \cup \{0\}$ is the least integer satisfying

$$f(x^k + s\beta^i d^k) \leq f(x^k) + \alpha s\beta^i \langle \nabla f(x^k), d^k \rangle. \tag{3.1}$$

The existence of such $i \in \mathbb{N} \cup \{0\}$ in this rule needs a closer look. Let us suppose that no such $i \in \mathbb{N} \cup \{0\}$ exists, so that at $t := \beta^i s$ (for any $i \in \mathbb{N} \cup \{0\}$), it holds

$$f(x^k + td^k) > f(x^k) + \alpha t \langle \nabla f(x^k), d^k \rangle. \tag{3.2}$$

Recall also that

$$f(x^k + td^k) = f(x^k) + t\langle \nabla f(x^k), d^k \rangle + o(t).$$

Combining the two, we obtain that

$$(\alpha - 1)\langle \nabla f(x^k), d^k \rangle + \frac{o(t)}{t} < 0.$$

Letting $t \to 0+$, we obtain $(\alpha - 1)\langle \nabla f(x^k), d^k \rangle \leq 0$. This is a contradiction, because $(\alpha - 1) < 0$ and $\langle \nabla f(x^k), d^k \rangle < 0$.

Discussion 3.3. Give a geometric interpretation of the Armijo's backtracking linesearch rule.

Always assume that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. In this section, we give a preliminary convergence analysis for general descent methods in which the descent directions d^k is bounded towards the direction of steepest descent $-\nabla f(x^k)$ (the discussion will be given in the next section). This can be made precise as follows.

Let $\varepsilon > 0$ be a given deviation threshold. Suppose that x^k has been computed and $\nabla f(x^k) \neq 0$. We assume that the unit direction d^k is chosen so that the angle θ_k between the two vectors $-\nabla f(x^k)$ and d^k satisfies

$$\cos\theta_k = \frac{\langle -\nabla f(x^k), d^k\rangle}{\|\nabla f(x^k)\|} \in (\varepsilon, 1], \quad \|d^k\| = 1. \tag{3.3}$$

Note from this condition that d^k is a descent direction, since we have

$$\langle \nabla f(x^k), d^k\rangle < -\varepsilon\|\nabla f(x^k)\| \le 0.$$

In the following theorem, an accumulation properties of any descent method satisfing the above bound is discussed. Note, however, that there is no guarantee either for the full convergence nor the existence of a limit point.

Theorem 3.4. *Let (x^k) be a sequence generated from Algorithm 3.1 where the search directions satisfy (3.3) and the Armijo's backtracking linesearch rule is applied with parameters α, s, β. If $\overline{x} \in \mathbb{R}^n$ is a limit point of (x^k), then $\nabla f(\overline{x}) = 0$.*

Proof. Suppose that Algorithm 3.1 generates an infinite sequence and let $\overline{x}$ is a limit point of x^k so that there is a subsequence (x^{k_q}) of (x^k) in which $\lim_q x^{k_q} = \overline{x}$. We may also assume that d^{k_q} converges to some $\overline{d} \in \mathbb{R}^n$ with $\|\overline{d}\| = 1$. From the linesearch rule, we have

$$f(x^{k_q}) - f(x^{k_q} + t_{k_q} d^{k_q}) \ge -\alpha t_{k_q}\langle \nabla f(x^{k_q}), d^{k_q}\rangle \ge \varepsilon\alpha t_{k_q}\|\nabla f(x^{k_q})\|. \tag{3.4}$$

Suppose to the contrary that $\nabla f(\overline{x}) \neq 0$. Then the condition (3.3) implies that $\langle \nabla f(\overline{x}), \overline{d}\rangle < 0$. Moreover, since $(f(x^k))$ is decreasing, it converges to some f^*. Together with (3.4), this makes $t_{k_q} \to 0$ as $q \to \infty$. It is reflected in the latter conclusion that the backtracking process at the step k_q is not succeeded at $i = 0$. This means

$$f(x^{k_q}) - f\left(x^{k_q} + \frac{t_{k_q}}{\beta} d^{k_q}\right) < -\alpha\frac{t_{k_q}}{\beta}\langle \nabla f(x^{k_q}), d^{k_q}\rangle.$$

Apply the Mean Value Theorem 1.12 to $f\left(x^{k_q}+\frac{t_{k_q}}{\beta}d^{k_q}\right)$ in the above inequality for each $q\in\mathbb{N}$, there exists $z^{k_q}\in]x^{k_q},x^{k_q}+\beta^{-1}t_{k_q}d^{k_q}[$ such that

$$-\beta^{-1}t_{k_q}\langle\nabla f(z^{k_q}),d^{k_q}\rangle < -\alpha\beta^{-1}t_{k_q}\langle\nabla f(x^{k_q}),d^{k_q}\rangle. \tag{3.5}$$

Since $x^{k_q}\to\overline{x}$, we know that $z^{k_q}\to\overline{x}$ as well. Rearranging (3.5) and letting $q\to\infty$, we get

$$(\alpha-1)\langle\nabla f(\overline{x}),\overline{d}\rangle\leq 0$$

Since $\alpha-1<0$, this contradicts with the above calculation that $\langle\nabla f(\overline{x}),\overline{d}\rangle<0$. We have therefore proved that $\nabla f(\overline{x})=0$. ∎

3.3 Gradient Descent Methods

The gradient descent methods refer to the Algorithm 3.1 where the search direction d_k is chosen to be the negative gradient $-\nabla f(x^k)$ for all $k\in\mathbb{N}\cup\{0\}$, since this is the direction of steepest descent for f from the point x^k.

Discussion 3.5. Let us formally observe that $-\nabla f(x)$ is the steepest descent direction for f at x in the sense that $f'(x;\cdot)$ is minimized (i.e. the most negative) among the directions d of equal magnitude.

Since the steepest direction $d^k:=-\nabla f(x^k)$ satisfies (3.3), we can immediately get the following corollary.

Corollary 3.6. *Let (x^k) be a sequence generated from Algorithm 3.1 using the gradient descent direction and the Armijo's backtracking linesearch rule with parameters α,s,β. If $\overline{x}\in\mathbb{R}^n$ is a limit point of (x^k), then $\nabla f(\overline{x})=0$.*

The full convergence using the steepest descent direction can be guaranteed under a rather strong assumption of strong convexity. This can in fact be relaxed by asking only the Lipschitz continuity of the gradient. We shall postpone to show that strong convexity of f implies the Lipschitz continuity of the gradient until before showing the final convergence analysis for strongly convex functions.

The following convergence criteria is quite useful in situations where the Lipschitz constant L of the gradient ∇f is accessible. Due to the limited applicability regarding the constant L, we shall just state the result without proof and save the space to prove another more applicable ones.

Theorem 3.7. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and continuously differentiable with Lipschitz continuous gradient with constant $L > 0$. Suppose that $Opt(f) \neq \emptyset$. Then the sequence (x^k) generated by the gradient descent method under constant step-size rule with $\bar{t} = \frac{1}{L}$ converges to a minimizer of f. Moreover we have the estimate*

$$f(x_k) - \inf_{\mathbb{R}^n} f = o(1/k).$$

In the case when the Lipschitz constant is inaccessible or expensive, we have to adopt the backtracking linesearch. The convergence analysis for this case relies on the upper bound of the maximum eigenvalues of the Hessians, derivable from the strong convexity itself.

For the remaining of this section, suppose that $f : U \to \mathbb{R}$ is a strongly convex function (with a constant $m > 0$) and there exists $p \in U$ in which $S_{f(p)}$ is closed. The following result implies an existence of $M > 0$ so that $mI \preceq \nabla^2 f(x) \preceq MI$ on $S_{f(p)}$.

Theorem 3.8. *There exists $M > 0$ in which $\nabla^2 f(x) \preceq MI$ for all $x \in S_{f(p)}$.*

Proof. Assume to the contrary that for each $M > 0$, there is $y_M \in S_{f(p)}$ such that $\nabla^2 f(y_M) \npreceq MI$. Take a positive real sequence (M_k) with $M_k \to +\infty$ and a sequence (y_k) in $S_{f(p)}$ with $\nabla^2 f(y_k) \npreceq M_k I$ for each $k \in \mathbb{N}$. Then, for each $k \in \mathbb{N}$, there exists $u_k \in \mathbb{R}^n \neq \{0\}$ for which

$$M_k u_k^\top u_k - u_k^\top \nabla^2 f(y_k) u) k = u_k^\top (M_k I - \nabla^2 f(y_k)) u_k < 0.$$

Rearrange the above inequality, use the Cauchy-Schwarz inequality, and then use the operator norm, we obtain

$$M_k < \frac{u_k^\top \nabla^2 f(y_k) u_k}{\|u_k\|^2} \leq \frac{\|u_k\| \|\nabla^2 f(y_k) u_k\|}{\|u_k\|^2} \leq \sup_{\|v\|=1} \|\nabla^2 f(y_k) v\| = \|\nabla^2 f(y_k)\|. \quad (3.6)$$

Note that $S_{f(p)}$ is a compact set, $\| \cdot \|$ and $\nabla^2 f$ are continuous. Hence $\|\nabla^2 f(\cdot)\|$ is bounded above on $S_{f(p)}$. This prevents the right hand side of (3.6) to diverge to $+\infty$, a contradiction. This guarantees an existence of a constant $M > 0$ such that $\nabla^2 f(x) \preceq MI$ holds for all $x \in S_{f(p)}$. ∎

We suppose throughout the section that $M > 0$ is the constant as in the above theorem. We may derive the following upper quadratic bound result from such $\preceq$-bound.

Theorem 3.9. *The inequality*

$$f(y) \leq f(x) + \nabla f(x)^\top (y-x) + \frac{M}{2}\|y-x\|^2$$

holds for all $x, y \in U$.

Proof. Similar to Theorem 1.28. ∎

We need one additional estimation for the infimum value for strongly convex functions.

Proposition 3.10. *The estimate*

$$\frac{1}{2M}\|\nabla f(x)\|^2 \leq f(x) - \inf_{\mathbb{R}^n} f \leq \frac{1}{2m}\|\nabla f(x)\|^2$$

for all $x, y \in U$.

Proof. Let $x \in U$ and put $h(y) := f(x) + \nabla f(x)^\top (y-x) + \frac{m}{2}\|y-x\|^2$ for $y \in U$. Then h is quadratic and therefore is minimized at $\tilde{y} := x - \frac{1}{m}\nabla f(x)$. We obtain from Theorem 1.28 that

$$\begin{aligned} f(y) &\geq f(x) + \nabla f(x)^\top (y-x) + \frac{m}{2}\|y-x\|^2 \\ &\geq f(x) - \frac{1}{m}\|\nabla f(x)\|^2 + \frac{1}{2m}\|\nabla f(x)\|^2 \\ &= f(x) - \frac{1}{2m}\|\nabla f(x)\|^2. \end{aligned}$$

Taking infimum over all $y \in U$ on the left hand side of the above inequality implies

$$f(x) - \inf_{\mathbb{R}^n} f \leq \frac{1}{2m}\|\nabla f(x)\|^2$$

The lower bound can be derived similarly using Theorem 3.9. ∎

Proposition 3.11. *If $f : \mathbb{R}^n \to \mathbb{R}$ is strongly convex and $M > 0$ satisfies $\nabla^2 f(x) \preceq MI$ for all $x \in \mathbb{R}^n$, then*

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{1}{M}\|\nabla f(y) - \nabla f(x)\|^2$$

for all $x, y \in \mathbb{R}^n$.

Proof. Let $x, y \in \mathbb{R}^n$, then we define the functions f_x and f_y by

$$f_x(z) := f(z) - \nabla f(x)^\top z \qquad \text{and} \qquad f_y(z) := f(z) - \nabla f(y)^\top z \tag{3.7}$$

for all $z \in \mathbb{R}^n$. Then $\nabla f_x(z) = \nabla f(z) - \nabla f(x)$ and $\nabla f_y(z) = \nabla f(z) - \nabla f(y)$. Then x and y minimizes f_x and f_y respectively. We further have from Proposition 3.10 that

$$\begin{aligned} f(y) - f(x) - \nabla f(x)^\top (y - x) &= f_x(y) - f_x(x) \\ &\geq \frac{1}{2M} \|\nabla f_x(y)\|^2 \\ &= \frac{1}{2M} \|\nabla f(y) - \nabla f(x)\|^2. \end{aligned}$$

We may similarly obtain that

$$f(x) - f(y) - \nabla f(y)^\top (x - y) \geq \frac{1}{2M} \|\nabla f(y) - \nabla f(x)\|^2.$$

Adding the two inequalities yields (3.7). ∎

Proposition 3.12. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a strongly convex function and $M > 0$ satisfies $\nabla^2 f(x) \preceq MI$ for all $x \in \mathbb{R}^n$. Then f has a Lipschitz continuous gradient with constant $M > 0$, i.e. the inequality*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

holds for all $x, y \in \mathbb{R}^n$.

Proof. Apply the Cauchy-Schwarz inequality to (3.7). ∎

The above result has granted us to use the gradient descent method with constant step-size as in Theorm 3.7.

New we are in the position to consider the convergence analysis of the gradient descent methods for strongly convex functions with Armijo's backtracking linesearch rule, simpliied to the case $s = 1$. Notice that the test statement for the backtraking (3.2) becomes

$$f(x^k + t_k d^k) > f(x^k) - \alpha t \|\nabla f(x^k)\|^2.$$

Theorem 3.13. *Let $f : U \to \mathbb{R}$ be a strongly convex function such that there is $p \in U$ in which $S_{f(p)}$ is closed, and let $m, M > 0$ be constants such that $mI \preceq \nabla^2 f(x) \preceq MI$ for all $x \in S_{f(p)}$. Then the sequence (x^k), with $x^0 \in S_{f(p)}$, generated by the gradient* *descent method under the Armijo's backtracking linesearch rule with parameters $\alpha, s = 1, \beta$ is convergent to the unique minimizer $\overline{x}$ of f. Moreover, we have the estimate*

$$f(x^{k+1}) - \inf_{\mathbb{R}^n} f \leq (1 - \min\{2\alpha m, 2\alpha\beta m/M\})^k \left(f(x^0) - \inf_{\mathbb{R}^n} f \right). \tag{3.8}$$

Proof. We first claim that the backtracking exit condition

$$f(x^k + td^k) \leq f(x^k) - \alpha t \|\nabla f(x^k)\|^2$$

is satisfied for all $0 \leq t \leq 1/M$. Notice that if $t \in [0, 1/M]$ then $\frac{M}{2}t^2 \leq \frac{M}{2} \cdot \frac{1}{M}t = \frac{t}{2} = -\frac{t}{2} + t$, finally leads to $-t + \frac{M}{2}t^2 \leq -\frac{t}{2}$. From Theorem 3.9, we have

$$\begin{aligned} f(x^k + td^k) &\leq f(x^k) - t\|\nabla f(x^k)\|^2 + \frac{Mt^2}{2}\|\nabla f(x^k)\|^2 \\ &\leq f(x^k) - (t/2)\|\nabla f(x^k)\|^2 \\ &\leq f(x^k) - \alpha t \|\nabla f(x^k)\|^2 \end{aligned}$$

for $t \in [0, 1/M]$. The claim is thus proved.

At any iterate k, if the exit condition holds at $i = 0$, then $t_k = 1$ and we have

$$f(x^{k+1}) = f(x^k + td^k) \leq f(x^k) - \alpha \|\nabla f(x^k)\|^2.$$

Otherwise, $t_k \geq \beta/M$ as we know that $\beta/M < 1/M$. This gives

$$f(x^{k+1}) = f(x^k + t_k d^k) \leq f(x^k) - \alpha \left(\frac{\beta}{M}\right) \|\nabla f(x^k)\|^2.$$

Together, we put $C := \min\{\alpha, (\alpha\beta)/M\}$ and the estimate concludes into

$$f(x^{k+1}) \leq f(x^k) - C\|\nabla f(x^k)\|^2.$$

Subtracting both sides with $\inf_{\mathbb{R}^n} f$ and applying Proposition 3.10, we get

$$\begin{aligned} f(x^{k+1}) - \inf_{\mathbb{R}^n} f &\leq \left[f(x^k) - \inf_{\mathbb{R}^n} f \right] - C\|\nabla f(x^k)\|^2 \\ &\leq (1 - 2mC) \left[f(x^k) - \inf_{\mathbb{R}^n} f \right]. \end{aligned}$$

Continuing inductively, we get

$$f(x^{k+1}) - \inf_{\mathbb{R}^n} f \leq (1 - 2mC)^k \left[f(x^0) - \inf_{\mathbb{R}^n} f \right] \to 0,$$

since $2mC \in (0,1)$. This shows that $f(x^k) \to \inf_{\mathbb{R}^n} f$.

To see the convergence of (x^k), recall that

$$f(x^{k+1}) \geq f(\overline{x}) + \nabla f(\overline{x})^\top (x^{k+1} - \overline{x}) + \frac{m}{2}\|x^{k+1} - \overline{x}\|^2.$$

We thus have $x^k \to \overline{x}$, as desired. ∎

The estimate (3.8) suggests that we may reach a solution within an acceptable tolorence $\varepsilon > 0$, i.e. $f(x^k) - \inf_{\mathbb{R}^n} f < \varepsilon$, within at most

$$\frac{\log([f(x^0) - \inf_{\mathbb{R}^n} f]/\varepsilon)}{\log(1/c)},$$

where $c := (1 - \min\{2\alpha m, 2\alpha\beta m/M\}) \in (0,1)$. From here notice that the situation $\min\{2\alpha m, 2\alpha\beta m/M\} = 2\alpha m$ happens if and only if $\beta/M \geq 1$. This is not likely as β is restricted in the interval $(0,1)$. Hence, in most cases, one would expect $c = 1 - 2\alpha\beta(m/M)$. The quantity M/m provides a bound for the condition number (ratio between the widest and thinnest parts) of the sublevel sets of f. If the ratio M/n is large, then

$$\log(1/c) \approx 2\alpha\beta(m/M).$$

This further suggests that the required number of iterations changes almost linearly with the bound of the condition number. For a quadratic function $f(x) = x^\top Ax + b^\top x + c$, the condition number is the ratio $\lambda_{max}/\lambda_{min}$ of the matrix A.

Discussion 3.14. Consider a quadratic function $f(x_1, x_2) := x_1^2 + \gamma x_2^2$, where $\gamma > 0$ is a fixed parameter. Try using programming the gradient descent method and solve for a minimizer of this function with different values of γ. Note that the condition number of this function is $\frac{\max\{1,\gamma\}}{\min\{1,\gamma\}}$. Discuss what happens to the required number of iterations with varying γ's.

3.4 Conjugate Gradient Methods

In this section, we discuss the conjugate gradient method which was originally designed to solve a quadratic problem whose Hessian is positive definite. This algorithm has an exceptional performance since a finite convergence is guaranteed, with its expanding subspace optimization properties.

We may also view that the conjugate gradient is a scheme designed for solving the linear system

$$Ax = b$$

where $A \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix and $b \in \mathbb{R}^n$. As described earlier, this linear system corresponds to the gradient ∇f of the strictly convex quadratic function

$$f(x) := \frac{1}{2}x^\top Ax - b^\top x.$$

This means we may analyze the convergence of the linear conjugate gradient methods in terms of minimizing f.

Always suppose in this section that $A \in \mathbb{R}^{n\times n}$ is a summetric postive definite matrix. A set $\{p_0, \cdots, p_m\}$ of nonzero vectors in $\mathbb{R}^n$ is said to be *conjugate* w.r.t. A (or A-*conjugate*) if

$$p_i^\top Ap_j = 0 \qquad \text{for all } i \neq j.$$

One may immediately observe that if $\{p_0, \cdots, p_m\}$ is conjugate w.r.t. A, then it is linearly independent.

The conjugate direction methods refer to the class of Algorithm 3.1 where d^k is taken from an A-conjugate set $\{d^0, \cdots, d^n\}$. The step-size rule is then chosen according to the exact linesearch rule, explicitly given by

$$t_k := \frac{(r^k)^\top d^k}{(d^k)^\top Ad^k} \tag{3.9}$$

where $r^k := Ax^k - b$ is the residual term.

Theorem 3.15. *Let (x^k) be a seqence generated by the conjugate direction method with an A-conjugate set $\{d^0, \cdots, d^{n-1}\}$. Then (x^k) converges to the minimizer $\overline{x}$ of f within n steps. Moreover, x^k minimizes f over the affine subspace $x^0 + \operatorname{span}\{d^0, \cdots, d^{k-1}\}$.*

Proof. The A-conjugacy of $D := \{d^0, \cdots, d^{n-1}\}$ implies the linear independence of D. So D spans $\mathbb{R}^n$. Hence we may write $\overline{x} - x^0$ as

$$\overline{x} - x^0 = \sigma_0 x^0 + \cdots \sigma_{n-1}d^{n-1} \tag{3.10}$$

for some $\sigma_0, \cdots, \sigma_{n-1} \in \mathbb{R}$. For $k = 0, 1, \cdots, n-1\}$, pre-multiplying $(d^k)^\top A$ the above equation, we have

$$\sigma_k = \frac{(d^k)^\top A(\overline{x} - x^0)}{(d^k)^\top Ad^k}. \tag{3.11}$$

If x^k is generated from the conjugate direction method, we have

$$x^k = x^0 + t^0 d^0 + \cdots + t^{k-1}d^{k-1}.$$

Pre-multiplying with $(d^k)^\top A$ gives

$$(d^k)^\top A(x^k - x^0) = 0,$$

and hence

$$(d^k)^\top A(\overline{x} - x^0) = (d^k)^\top A(\overline{x} - x^k) = (d^k)^\top (b - Ax^k) = -(d^k)^\top r^k = -(r^k)^\top d^k.$$

Substitute this into (3.11), we see that $\sigma_k = t_k$, the exact linesearch rule (3.9). In view of (3.10), we see that x^k converges to $\overline{x}$ within n steps.

The subspace expansion property is left as an exercise. ■

The above theorem says that if one has a miracle way to generate an A-conjugate set, then one may solve $Ax = b$ within n steps. One may show that the eigenvectors of A form an A-conjugate set and can thus be used here. However, the real question is how to generate such a conjugate set in an inexpensive fashion. The very first, and perhaps most famous method, is the *conjugate gradient method*, whose main idea is to replace the steepest descent direction $-\nabla f(x^k) = -r^k$ with $-r^k + \beta_{k+1} d^k$. Here, the extra term d^k will help breaking the orthogonality in the zig-zag pattern between steps of the gradient descent by the amount β_{k+1}. This weight β_{k+1} is chosen precisely so that $d^{k+1} = -r^{k+1} + \beta_{k+1} d^k$ is A-conjugate to d^k, i.e. $(d^{k+1})^\top A d^k = 0$. We may find that

$$\beta_{k+1} = \frac{(r^{k+1})^\top A d^k}{(d^k)^\top A d^k}.$$

Let us now formally state the Conjugate Gradient Algorithm in the following.

Algorithm 3.16. Conjugate Gradient Algorithm.

```
Initialization:
    Pick a start point $x^0 \in \mathbb{R}^n$.
    Set $r^0 \leftarrow Ax^0 - b$, $d^0 \leftarrow -r^0$, $k \leftarrow 0$.
While: $r^k \neq 0$;
    $t_k \leftarrow -\frac{(r^k)^\top d^k}{(d^k)^\top A d^k}$.
    $x^{k+1} \leftarrow x^k + t_k d^k$.
    $r^{k+1} \leftarrow Ax^{k+1} - b$.
    $\beta_{k+1} \leftarrow \frac{(r^{k+1})^\top A d^k}{(d^k)^\top A d^k}$.
    $d^{k+1} \leftarrow -r^{k+1} + \beta_{k+1} d^k$.
    Update $k \leftarrow k + 1$.
```

Discussion 3.17. Try to prove the convergence of this scheme by showing that the generated directions d^k is A-conjugate to all the previous directions $d^0, \cdots, d^{k-1}$.

§ 4. Constrained Optimization Problems – Theory

Unless otherwise specified, suppose throughout the chapter that $f : \mathbb{R}^n \to \mathbb{R}$, $C \subset \mathbb{R}^n$ is nonempty, and $g : \mathbb{R}^n \to \mathbb{R}^r$ and $h : \mathbb{R}^n \to \mathbb{R}^l$. Moreover, always assume that the functions f, g, h are all continuously differentiable.

We shall study optimality conditions for both a general constrained optimization problem $Opt(f, C)$ and $Opt(f, g, h)$.

4.1 Optimality Conditions for Abstract Constraints

The natural aim is to relate the local optimalityity with the condition saying that the objective will increase if one moves around withing the constraint set. Thus we need first a good notion of *allowed directions* to travel within the constraint and then observe the increase through the directional derivatives along such directions.

Definition 4.1. At a fixed position $x \in C$, a direction $d \in \mathbb{R}^n$ is called *feasible* at x w.r.t. C if there is $\overline{t} > 0$ such that $x + td \in C$ for all $t \in (0, \overline{t})$. If the context of x and C is clear, we only say that d is a feasible direction.

Example 4.2. Let us see some important simple examples.

- If C has an interior point x, then any vector $d \in \mathbb{R}^n$ is a feasible direction at x.
- If C is convex and $x \in C$, then $d \in \mathbb{R}^n$ is a feasible direction if and only if $d = t(y-x)$ for some $t > 0$ and $y \in C$.

The concept of a feasible direction seems to work fine in the above examples. This is because the vicinity of such point x is *locally* convex. One may observe, instead, the following example where there is no feasible directions at all feasible points.

Example 4.3. Let C be the unit circle in $\mathbb{R}^s$ and $x \in C$. Then the only feasible direction at x is the zero vector.

Looking at this example, one may employ the idea of differential geometry where traveling along a curve has a direction that is only *tangent* to the curve. Such a direction does not lie necessarily over the curve if one goes straightly along.

Definition 4.4. Let $x \in C$. A direction $d \in \mathbb{R}^n$ is said to be *tangent* to the set C at x if there is a sequence (z_n) in C converging to x and a sequence (t_n) of positive scalars converging to 0 such that

$$\frac{x_n - x}{t_n} \to d.$$

The set of all directions tangent to C at x is called a *tangent cone* to C at x, denoted by $T_C(x)$.

Example 4.5. Let C be the unit circle in $\mathbb{R}^s$ and $x \in C$. Then the $T_C(x)$ is the directions along the tangent line to C at x.

Example 4.6. Let $f : \mathbb{R} \to \mathbb{R}$ be a function defined by $f(x) := \sqrt{|x|}$ for all $x \in \mathbb{R}$, and define $C := \operatorname{epi} f = \{(x_1, x_2) \in \mathbb{R}^2 \,|\, x_2 \geq f(x_1)\}$. Then what is $T_C(0,0)$?

Example 4.7. Let $f : \mathbb{R} \to \mathbb{R}$ be a function defined by $f(x) := \max\{\arctan(x), 0\}$ for all $x \in \mathbb{R}$, and define $C := \operatorname{epi} f = \{(x_1, x_2) \in \mathbb{R}^2 \,|\, x_2 \geq f(x_1)\}$. Then what is $T_C(0,0)$?

We will see that if C is convex, then the tangency concept is redundant and we may rebound ourselves to the feasible directions.

Theorem 4.8. *(FO-NOC) Let $\overline{x}$ be a local solution of Opt(f, C), where f is continuously differentiable, then*

$$\langle \nabla f(\overline{x}), d \rangle \geq 0 \qquad \textit{for all } d \in T_C(\overline{x}).$$

If, in addition, C is convex, then the above inequality is equivalent to

$$\langle \nabla f(\overline{x}), y - \overline{x} \rangle \geq 0 \qquad \textit{for all } y \in C. \tag{4.1}$$

Proof. Local optimality of $\overline{x}$ means $f(x) - f(\overline{x}) \geq 0$ for all x in a neighborhood of $\overline{x}$. Let x be any point in such a neighborhood. Using the Landau's approach to Fréchet derivative, we have $f(x) - f(\overline{x}) \leq \langle \nabla f(\overline{x}), x - \overline{x} \rangle + o(\|x - \overline{x}\|)$. Then we get $-\langle \nabla f(\overline{x}), x - \overline{x} \rangle \leq o(\|x - \overline{x}\|)$. The remaining of the proof is left as an exercise. ∎

Again, to obtain sufficiency, we assume that f is convex.

Theorem 4.9 (SOC)**.** *Assume that f and C are both convex and f is continuously differentiable. If $\overline{x} \in C$ satisfies (4.1), then $\overline{x} \in$ Opt(f, C).*

Proof. The result follows directly from the subgradient inequality (1.2). ∎

The inequality (4.1) is known in the literature as the *variational inequality (VI)* associated to ∇f over C. There is a huge literature on VI associated to different types of mappings for different purposes and applications, e.g. economics, game theory, and PDEs.

4.2 KKT Conditions for Equality and Inequality Constraints

In this section, we derive optimality conditions for the constrained optimization problem that takes the form $\mathit{Opt}(f, g, h)$. With the special structure of C, we may rewrite the optimality condition by exploiting the gradients of all the functions involved, namely ∇f, ∇g_i, and ∇h_j for $i = 1, \cdots, r$ and $j = 1, \cdots, l$.

We first introduce the condition that will be used as our optimality criteria.

Definition 4.10. Consider the problem $\mathit{Opt}(f, g, h)$. A feasible point $x \in \mathbb{R}^n$ is said to satisfy the *Karush-Kuhn-Tucker condition* (briefly, the *KKT condition*) if there exist scalars, called *Lagrange multipliers*, $\lambda_1, \cdots, \lambda_r \geq 0$ and $\mu_1, \cdots, \mu_l \in \mathbb{R}$ such that

$$\begin{cases} \nabla f(x) + \sum_{i=1}^{r} \lambda_i \nabla g_i(x) + \sum_{j=1}^{l} \mu_j \nabla h_j(x) = 0, \\ \lambda_i g_i(x) = 0 \qquad \text{for all } i = 1, \cdots, r. \end{cases}$$

For any $i = 1, \cdots, r$ and $j = 1, \cdots, l$, the multipliers λ_i and μ_j are said to be associated to the constraint g_i and h_j, respectively.

Definition 4.11. Consider the problem $\mathit{Opt}(f, g, h)$. An inequality constraint $g_i(x) \leq 0$ is said to be *active* at a point $x \in \mathbb{R}^n$ if $g_i(x) = 0$. The active inequality index set at x is then defined by

$$\mathcal{A}(x) := \big\{ i \in \{1, \cdots, r\} \,|\, g_i(x) = 0 \big\}.$$

Naturally, it is to be understood that all the equatlity constraints are active, if they are satisfied. The intuition behind an activity of a constraint at a certain point x is as follows: when a constraint is satisfied as an equality, it is at risk of being violated if the point x is moved or inaccurately approximated ever so slightly.

Looking back to the second hypothesis of the KKT condition. It implies that $\lambda_i = 0$ whenever the constraint g_i is inactive. Hence the summation in the first hypothesis of the KKT condition can be reduced only to the active constraints. It simply states that the steepest descent direction of f at x (i.e. $-\nabla f(x)$) is only in the (cofnditioned) linear combination of gradients of the active constraints which is infeasible.

Definition 4.12. A x be a feasible point of the problem $Opt(f, g, h)$ is said to satisfy the *Linear Independence Constraint Qualification* (briefly *LICQ*) if the gradients of all active constraints are linearly independent, i.e. the set $\{\nabla g_i(x) \,|\, i \in \mathcal{A}(x)\} \cup \{\nabla h_1(x), \cdots, \nabla h_l(x)\}$ is linearly independent.

The following theorem gives the general necessary optimality condition for general problem $Opt(f, g, h)$ under LICQ. The proof of this theorem is lengthy and will be left to the interested readers. However, the proofs will be provided for some specific cases or variants of this theorem in subsequent sections.

Theorem 4.13 (General KKT Necessary Optimality Conditions with LICQ)**.** *Suppose that $\overline{x}$ is a local solution of $Opt(f, g, h)$, where f, g, h are all continuously differentiable, and that LICQ holds at this point. Then $\overline{x}$ satisfies the KKT condition.*

The KKT condition can be tricky to use sometimes as it depends also on how well an optimization problem is formulated. Let us illustrate this through the following example.

Example 4.14. Consider the problems

$$\begin{array}{rl} \min & x_1 + x_2 \\ \text{s.t.} & x_1^2 + x_2^2 = 1. \end{array} \tag{4.2}$$

and

$$\begin{array}{rl} \min & x_1 + x_2 \\ \text{s.t.} & (x_1^2 + x_2^2 - 1)^2 = 0. \end{array} \tag{4.3}$$

We observe first that the two problem (4.2) and (4.3) are equivalent in the sense that the objective functions feasible sets are the same. Observe what happens when one apply the KKT condition ?

4.3 KKT Conditions for Inequality Constraints

The KKT necessary condition can be simplified if the equality consitraint is not presented. The proof is also much less complicated in this case.

To fix the idea, we consider in this section an optimization problem of the form

$$\begin{cases} \min & f(x) \\ \text{s.t.} & g_i(x) \leq 0 \quad \forall i = 1, \cdots, r, \end{cases}$$

where all $f, g_1, \cdots, g_r$ are continuously differentiable over $\mathbb{R}^n$. For convenience, the above problem will be referred to as $Opt(f, g)$.

To deliver the KKT condition for the above problem, we need to derive a preliminary form of the KKT condition, known as the Fritz-John condition, from the Gordan's theorem.

Theorem 4.15 (Gordan's Alternative Theorem). *Let $A \in R^{m\times n}$. Then exactly one of the following two systems has a solution:*

A. $Ax < 0$.

B. $p \neq 0$, $A^\top p = 0$, $p \geq 0$.

Theorem 4.16 (Fritz-John conditions). *Let $\overline{x} \in LOpt(f, g)$. Then there exist multipliers $\lambda_0, \lambda_1, \cdots, \lambda_r \geq 0$, which are not all zero, satisfying*

$$\begin{cases} \lambda_0 \nabla f(\overline{x}) + \sum_{i=1}^r \lambda_i \nabla g_i(\overline{x}) = 0, \\ \lambda_i g_i(\overline{x}) = 0 \qquad \textit{for all } i = 1, \cdots, r. \end{cases}$$

Proof. First, we claim that there is no $d \in \mathbb{R}^n$ such that $\langle \nabla f(\overline{x}), d\rangle < 0$ and for all $i \in \mathcal{A}(\overline{x})$, $\langle \nabla g_i(\overline{x}), d\rangle < 0$. Indeed, if there is such $d \in \mathbb{R}^n$, then for $t > 0$ sufficiently small, we get $f(\overline{x}+td) < f(\overline{x})$ and $g_i(\overline{x}+td) < 0$ for all $i = 1, \cdots, r$. This contradicts with the hypothesis that $\overline{x} \in LOpt(f, g)$.

Writing $\mathcal{A}(\overline{x}) = \{i_1, \cdots, i_k\}$ and put $A = [\nabla f(\overline{x})\, \nabla g_{i_1}(\overline{x}) \cdots \nabla g_{i_k}(\overline{x})]^\top$, the conditions of the claim becomes $Ad < 0$. Since this system has no solution, the Gordan's Alternative Theorem implies the existence of $\hat{\lambda} := (\lambda_0, \lambda_{i_1}, \cdots, \lambda_{i_k}) \neq 0$ such that $A^\top \hat{\lambda} = 0$ and $\hat{\lambda} \geq 0$. This is the same as

$$\begin{cases} \lambda_0 \nabla f(\overline{x}) + \sum_{i\in\mathcal{A}(\overline{x})} \lambda_i \nabla g_i(\overline{x}) = 0, \\ \lambda_i g_i(\overline{x}) = 0 \qquad \text{for all } i \in \mathcal{A}(\overline{x}). \end{cases}$$

The conclusion follows by letting $\lambda_i := 0$ for $i \notin \mathcal{A}(\overline{x})$. ∎

A major downside of the Fritz-John condition is the multiplier λ_0 associated to the objective function f which is legal to be 0. When this happens, the Fritz-John condition reduces to the linear dependence of $\nabla g_i(\overline{x})$ without any information of f. Hence, there can be way too many points that satisfy the Fritz-John condition without being optimal. This, again, leads to the LICQ condition that would then force $\lambda_0 \neq 0$ and finally the KKT condition.

Theorem 4.17 (KKT Necessary Optimality Conditions for Inequality Constraints with LICQ)**.** *Suppose that $\overline{x}$ is a local solution of Opt(f, g) and that LICQ holds at this point. Then there exist multipliers $\lambda_1, \cdots, \lambda_r \geq 0$, which are not all zero, satisfying*

$$\begin{cases} \nabla f(\overline{x}) + \sum_{i=1}^{r} \lambda_i \nabla g_i(\overline{x}) = 0, \\ \lambda_i g_i(\overline{x}) = 0 \qquad \textit{for all } i = 1, \cdots, r. \end{cases}$$

Proof. We know the Fritz-John condition holds at $\overline{x}$, hencee there exist multipliers $\hat{\lambda}_0, \hat{\lambda}_1, \cdots, \hat{\lambda}_m \geq 0$, which are not all zero, satisfying

$$\begin{cases} \hat{\lambda}_0 \nabla f(\overline{x}) + \sum_{i=1}^{r} \hat{\lambda}_i \nabla g_i(\overline{x}) = 0, \\ \hat{\lambda}_i g_i(\overline{x}) = 0 \qquad \text{for all } i = 1, \cdots, r. \end{cases} \tag{4.4}$$

If $\hat{\lambda}_0 = 0$, then $\sum_{i=1}^{r} \hat{\lambda}_i \nabla g_i(\overline{x}) = 0$. Since $\hat{\lambda}_i$'s are not all zero, the gradients $\nabla g_i(\overline{x})$'s are linearly dependent. This contradicts our hypothesis, so $\hat{\lambda}_0 \neq 0$. Dividing by $\hat{\lambda}_0$ in all equations of (4.4) gives the desired KKT condition with $\lambda_i := \hat{\lambda}_i / \hat{\lambda}_0$. ∎

We have no doubt in the requirement of the LICQ but this CQ is rather quite technical than being practical. It turns out that we may deduce a more practical CQ when all the functional constraints are convex.

Theorem 4.18 (KKT Necessary Optimality Conditions for Convex Inequality Constraints with Slater's CQ)**.** *Suppose that $\overline{x}$ is a local solution of Opt(f, g), where $g_1, \cdots, g_r$ are all convex, and that the Slater's CQ (i.e. there is a point $\hat{x} \in \mathbb{R}^n$ in which no constraints are active) holds at this point. Then there exist multipliers $\lambda_1, \cdots, \lambda_r \geq 0$, which are not all zero, satisfying*

$$\begin{cases} \nabla f(\overline{x}) + \sum_{i=1}^{r} \lambda_i \nabla g_i(\overline{x}) = 0, \\ \lambda_i g_i(\overline{x}) = 0 \qquad \textit{for all } i = 1, \cdots, r. \end{cases}$$

Proof. Once again, the proof relies on the Fritz-John condition. We know that there exist multipliers $\hat{\lambda}_0, \hat{\lambda}_1, \cdots, \hat{\lambda}_m \geq 0$, which are not all zero, satisfying

$$\begin{cases} \hat{\lambda}_0 \nabla f(\overline{x}) + \sum_{i=1}^{r} \hat{\lambda}_i \nabla g_i(\overline{x}) = 0, \\ \hat{\lambda}_i g_i(\overline{x}) = 0 \qquad \text{for all } i = 1, \cdots, r. \end{cases}$$

Let us assume that $\hat{\lambda}_0 = 0$, which implies $\sum_{i=1}^{r} \hat{\lambda}_i \nabla g_i(\overline{x}) = 0$. The Slater's CQ and the subgradient inequality then gives

$$0 > \sum_{i=1}^{r} \hat{\lambda}_i g_i(\hat{x}) \geq \sum_{i=1}^{r} \hat{\lambda}_i g_i(\overline{x}) + \left\langle \sum_{i=1}^{r} \hat{\lambda}_i \nabla g_i(\overline{x}), \hat{x} - \overline{x} \right\rangle = 0,$$

which is absurd. Therefore $\hat{\lambda}_0 \neq 0$ and the rest of the proof is similar to the previous theorem. ■

4.4 KKT Conditions for Linear Constraints

When all g_i's and h_j's are all affine, then no CQ's are required. The discussion of this section will be taken for an optimization of the form

$$\begin{cases} \min & f(x) \\ \text{s.t.} & a_i^\top x \leq b_i \quad \forall i = 1, \cdots, r, \\ & c_j^\top x = d_j \quad \forall j = 1, \cdots, l, \end{cases} \tag{4.5}$$

where f is continuously differentiable, $a_i, c_j \in \mathbb{R}^n$, and $b_i, d_j \in \mathbb{R}$ for all $i = 1, \cdots, r$ and $j = 1, \cdots, l$.

We shall prove the KKT conditions for the above problem without equality constraints first. For this, we need another alternative theorem.

Theorem 4.19 (Farkas' Alternative Theorem). *Let $A \in \mathbb{R}^{m \times n}$ and $c \in \mathbb{R}^n$. Then exactly one of the following systems has a solution:*

A. $Ax \leq 0$, $c^\top x > 0$.

B. $A^\top y = c$, $y \geq 0$.

Theorem 4.20 (KKT Necessary Optimality Conditions for linear inequality constraints). *Let $\overline{x}$ be a local solution of the problem*

$$\begin{cases} \min & f(x) \\ \text{s.t.} & a_i^\top x \leq b_i \quad \forall i = 1, \cdots, r, \end{cases}$$

where f is continuously differentiable, $a_i \in \mathbb{R}^n$, and $b_i \in \mathbb{R}$ for each $i = 1, \cdots, r$. Then there exist multipliers $\lambda_1, \cdots, \lambda_r \geq 0$ such that

$$\begin{cases} \nabla f(\overline{x}) + \sum_{i=1}^{r} \lambda_i a_i = 0, \\ \lambda_i (a_i^\top \overline{x} - b_i) = 0 \qquad \text{for all } i = 1, \cdots, r. \end{cases}$$

Proof. Observe first that the constraint set for the problem under consideration is convex. Hence the variational inequality

$$\langle \nabla f(\overline{x}), x - \overline{x} \rangle \geq 0$$

holds true for any $x \in \mathbb{R}^n$ satisfying the constraints. Making a change of the variables $y = x - \overline{x}$, we obtain $\nabla f(\overline{x})^\top y \geq 0$ for all $y \in \mathbb{R}^n$ satisfying

$$a_i^\top (y + \overline{x}) \leq b_i \tag{4.6}$$

for $i = 1, \cdots, r$. Note that the above inequality becomes $a_i^\top y \leq 0$ for $i \in \mathcal{A}(\overline{x})$.

We claim that the following implication holds:

$$a_i^\top y \leq 0 \ (\forall i \in \mathcal{A}(\overline{x})) \implies \nabla f(\overline{x})^\top y \geq 0,$$

i.e. the inactive constraints are superfluous. Let $y \in \mathbb{R}^n$ satisfies $a_i^\top y \leq 0$ for all $i \in \mathcal{A}(\overline{x})$. Since $b_i - a_i^\top \overline{x} > 0$ for all $i \notin \mathcal{A}(\overline{x})$, there exists $\alpha > 0$ sufficiently small so that $a_i^\top (\alpha y) \leq b_i - a_i^\top$ for all $i \notin \mathcal{A}(\overline{x})$. We also have $a_i^\top (\alpha y) \leq 0$ for all $i \in \mathcal{A}(\overline{x})$. This means (4.6) holds for αy, and thus $\nabla f(\overline{x})^\top (\alpha y) \geq 0$. Finally we have $\nabla f(\overline{x})^\top y \geq 0$ so that the claim is proved.

If we put $\mathcal{A}(\overline{x}) := \{i_1, \cdots, i_k\}$ and $A := [a_{i_1} \cdots a_{i_k}]^\top$, then the claim says that the system $Ax \leq 0$ with $-\nabla f(\overline{x})^\top x > 0$ has no solution. In view of the Farkas' Alternative Theorem, there exists $\lambda_{i_1}, \cdots, \lambda_{i_k} \geq 0$ such that $\sum_{j=1}^k \lambda_{i_j} a_{i_j} = -\nabla f(\overline{x})$. Letting $\lambda_i := 0$ for $i \notin \mathcal{A}(\overline{x})$, we arrived at the desired conclusion. ∎

The KKT conditions for the full linearly constrained problem (4.5) can be deduced directly from the above results for problems with only linear inequality constraints.

Theorem 4.21 (KKT Necessary Optimality Conditions for linear inequality and equality constraints)**.** *Let $\overline{x}$ be a local solution of the problem described by* (4.5)*. Then there exist multipliers $\lambda_1, \cdots, \lambda_r \geq 0$ and $\mu_1, \cdots, \mu_l \in \mathbb{R}$ such that*

$$\begin{cases} \nabla f(\overline{x}) + \sum_{i=1}^r \lambda_i a_i + \sum_{j=1}^l \mu_j c_j = 0, \\ \lambda_i (a_i^\top \overline{x} - b_i) = 0 \qquad \text{for all } i = 1, \cdots, r. \end{cases}$$

Proof. Derive the required prove from Theorem 4.20. ∎

Discussion 4.22. Notice that the proof of the KKT Necessary Condition for the full linear constraints can be derived by applying the KKT result for the problems with only linear inequality constraints. What makes this technique inapplicable for problems with nonlinear constraints (e.g. Theorem 4.13 from Theorem 4.17)?

4.5 Sufficiency of KKT Conditions under Convexity

When all the ingredients of the problem $Opt(f, g, h)$ are convex (i.e. Convex Program: CP) and continuously differentiable , the KKT conditions are sufficient to guarantee the optimality.

Theorem 4.23. *Let $\overline{x}$ be a feasible point of the problem $Opt(f, g, h)$ where f and g_i's are all convex continuously differentiable, and h_j's are all affine. If $\overline{x}$ is a KKT point, then $\overline{x} \in Opt(f, g, h)$.*

Proof. Suppose that $\lambda_1, \cdots, \lambda_r \geq 0$ and $\mu_1, \cdots, \mu_l \in \mathbb{R}$ are multipliers satisfying the KKT conditions at $\overline{x}$. Note that the function $L(x) := f(x) + \sum_{i=1}^{r} \lambda_i g_i(x) + \sum_{j=1}^{l} \mu_j h_j(x)$, defined for $x \in \mathbb{R}^n$, is convex. Moreover, the hypothesis that $\overline{x}$ is a KKT point implies that $\nabla L(\overline{x}) = 0$. This means $\overline{x}$ minimizes L over $\mathbb{R}^n$. Using complementarity condition of a KKT point, it follows that

$$f(\overline{x}) \leq L(\overline{x}) \leq L(x) = f(x) + \sum_{i=1}^{r} \lambda_i g_i(x) + \sum_{j=1}^{l} \mu_j h_j(x) \leq f(x)$$

for any feasible point x. Hence we get $\overline{x} \in Opt(f, g, h)$. ■

It is noteworthy that eventhough the convexity provides sufficiency of an optimal point through the lens of the KKT conditions, the necessity still relies on good CQs. The only exceptional case where no CQs are needed is when f is convex while g_i's and h_j's are affine.

We may conclude some important points here: (1) Under the Slater's CQ, Convex Program is equivalent to its KKT system, and (2) the Linear Program (LP) is equivalent to its KKT system unconditionally.

4.6 Linear Programs and Duality

Let us briefly discuss about a Linear Program (LP), which refers to the problem $Opt(f, g)$ where f is linear and g_i's are all affine. Due to the equivalence of the LP and its KKT system, we know that a solution of an LP exists if and only if the multipliers exist in the corresponding KKT system.

Suppose that $f = \langle c, \cdot \rangle$ and $g_i = \langle a_i, \cdot \rangle - b_i$ for $c, a_1, \cdots, a_r \in \mathbb{R}^n$ and $b_1, \cdots, b_r \in \mathbb{R}$. Let $\overline{x}$ be a solution of $Opt(f, g)$, then there exist multipliers $\overline{y}_1, \cdots, \overline{y}_r \geq 0$ in which

$$\sum_{i=1}^{r} \overline{y}_i a_i = -c \tag{4.7}$$

and for all $i = 1, \cdots, r$, it holds

$$\overline{y}_i(\langle a_i, \overline{x}\rangle - b_i) = 0. \tag{4.8}$$

Putting $\overline{y} := (\overline{y}_1, \cdots, \overline{y}_r)$ and $b := (b_1, \cdots, b_r)$, the above KKT condition leads to

$$\langle b, \overline{y}\rangle + \langle c, \overline{x}\rangle = \langle b, \overline{y}\rangle - \sum_{i=1}^{r} \overline{y}_i \langle a_i, \overline{x}\rangle = \sum_{i=1}^{r} \overline{y}_i (b_i - \langle a_i, \overline{x}\rangle) = 0,$$

so that we have

$$\langle c, \overline{x}\rangle = \langle -b, \overline{y}\rangle. \tag{4.9}$$

Now, if $x \in \mathbb{R}^n$ is a feasible point of $Opt(f, g)$, then (4.9) gives

$$\langle c, x\rangle \geq \langle -b, \overline{y}\rangle.$$

On the other hand, if $y \in \mathbb{R}^r$ satisfies $y_1, \cdots, y_r \geq 0$ and $\sum_{i=1}^{r} y_i a_i = -c$, then

$$\langle b, y\rangle + \langle c, \overline{x}\rangle = \sum_{i=1}^{r} y_i (b_i - \langle a_i, \overline{x}\rangle) \geq 0,$$

where the last inequality follows from the feasibility of $\overline{x}$. This means

$$\langle c, \overline{x}\rangle \geq \langle -b, y\rangle.$$

Thus, any $y \in \mathbb{R}^r$ with nonnegative entries that satisfies (4.7) provides a lower bound to the infimum for $Opt(f, g)$. Moreover, $\overline{y}$ maximizes $\langle b, \cdot\rangle$ over all such y. That is, $\overline{y}$ is a solution of the so-called *dual problem*:

$$\begin{cases} \max & \langle -b, y\rangle \\ \text{s.t.} & \langle a, y\rangle = -c \\ & y_i \geq 0 \quad \text{for all } i = 1, \cdots, r, \end{cases}$$

where $a = (a_1, \cdots, a_r)$. The vector $y \in \mathbb{R}^r$ is called *dual feasible* if it satisfies the constraint of the dual problem.

Let us conclude the observations of this section.

Theorem 4.24. *For any primal feasible point $x \in \mathbb{R}^n$ and any dual feasible point $y \in \mathbb{R}^r$, we have the weak duality:*

$$\langle c, x\rangle \leq \langle -b, y\rangle.$$

Moreover, if $\overline{x} \in \mathbb{R}^n$ is primal optimal and $\overline{y} \in \mathbb{R}^r$ is dual optimal, then the strong duality holds:

$$\langle c, \overline{x}\rangle = \langle -b, \overline{y}\rangle.$$

4.7 Programming packages

Most of the prominent programming languages (e.g. MatLab, Python, Julia, Scilab, etc.) offers a ready-to-use package for solving Linear Programs (LP) and Lineary constrained Quadratic Program (LQP). This is, most of the time, sufficient in medium scale problems. Let us demonstrate the use of such packages via an application of LQP in binary classification model called Support Vector Machine (SVM).

Discussion 4.25 (Binary classification with Support Vector Machine)**.** Given a data set $\{(x_i, y_i)\}_{i=1,\cdots,m}$, where $x_i \in \mathbb{R}^n$ is the collected data and $y_i \in \{-1, +1\}$ is the binary label of x_i into either the class $+1$ or the other class -1. For convenience, suppose that $x_1, \cdots, x_p$ belong to the class $+1$ and $x_{p+1}, \cdots, x_m$ belong to the class -1. For the simplicity, suppose that the two classes are separable by a hyperplane, i.e. there is a linear separator L given by a function

$$g(\cdot) := \langle w, \cdot\rangle + b,$$

where $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$, satisfying $g(x_i) > 0$ and $g(x_j) < 0$ for all $i = 1, \cdots, p$ and $j = p + 1, \cdots, m$.

The aim of the Support Vector Machine (SVM) is to find the best linear separator in the sense that it produces highest accuracty when generalized to the unseen data. This means the linear separator must have the margin to the data as large as possible. Noting that the distance between a given point $x \in \mathbb{R}^n$ and the plane L defined by the function g above can be computed by

$$d(x, L) := \frac{|\langle w, x\rangle + b|}{\|w\|},$$

the margin can be computed by

$$\min_{i=1,\cdots,m} d(x_i, L) = \min_{i=1,\cdots,m} \frac{|\langle w, x_i\rangle + b|}{\|w\|}.$$

An SVM can then be modeled as an optimization problem

$$\begin{cases} \max_{w,b} & \left\{ \min_{i=1,\cdots,m} \frac{|\langle w, x_i\rangle + b|}{\|w\|} \right\} \\ \text{s.t.} & \langle w, x_i\rangle + b > 0 \quad \text{for all } i = 1, \cdots, p, \\ & \langle w, x_j\rangle + b < 0 \quad \text{for all } j = p+1, \cdots, m. \end{cases}$$

The objective function of the above model is hard to solve. We shall reduce it finally to a LQP.

Note first that if (w, b) is a solution of the above optimization problem, then so is $(\lambda w, \lambda b)$ for any $\lambda > 0$. This means there is a solution $(\overline{w}, \overline{b})$ in which

$$\min_{i=1,\cdots,m} |\langle \overline{w}, x_i\rangle + \overline{b}| = 1.$$

This does not change anything to the aim of an SVM model. Moreover, this condition implies that the constraints can be improved. The new formulation reads

$$\begin{cases} \max_{w,b} & \|w\|^{-1} \\ \text{s.t.} & \min_{i=1,\cdots,m} |\langle w, x_i\rangle + b| = 1 \\ & \langle w, x_i\rangle + b \geq 1 \quad \text{for all } i = 1, \cdots, p, \\ & \langle w, x_j\rangle + b \leq -1 \quad \text{for all } j = p+1, \cdots, m. \end{cases}$$

We may then equivalently write $\max \|w\|^{-1}$ as $\min \frac{1}{2}\|w\|^2$ and also notice that the condition $\min_{i=1,\cdots,m} |\langle \overline{w}, x_i\rangle + \overline{b}| = 1$ can be dropped from the above formulation. This is because if $\min_{i=1,\cdots,m} |\langle w, x_i\rangle + b| > 1$ holds at a feasible point (w, b), then the constraints are not active and we may improve the objective value by choosing $(\lambda w, \lambda b)$ with $\lambda := (\min_{i=1,\cdots,m} |\langle w, x_i\rangle + b|)^{-1}$. Hence, the model for an SVM becomes

$$\begin{cases} \min_{w,b} & \frac{1}{2}\|w\|^2 \\ \text{s.t.} & \langle w, x_i\rangle + b \geq 1 \quad \text{for all } i = 1, \cdots, p, \\ & \langle w, x_j\rangle + b \leq -1 \quad \text{for all } j = p+1, \cdots, m, \end{cases}$$

which is now an LQP.

§ 5. Constrained Optimization Problems – Algorithms

In this chapter, we shall study some approaches to solve a constrained optimization problem numerically. The first and most naive approach is to modify the gradient descent algorithm using the projection operator to pull the infeasible update onto the constraint set.

5.1 Projected Gradient Algorithms

The projected gradient algorithm (or gradient projection algorithm) is a modification of the gradient descent algorithm that extends to any constrained optimziation problem whose constraint set is closed and convex.

Before we can officially state the algorithm, we first need to define a metric projection and study some of its properties.

Proposition 5.1. *Let $C \subset \mathbb{R}^n$ be a nonempty set, $x \in \mathbb{R}^n$, and define $d_x(\cdot) := \frac{1}{2}\|\cdot - x\|^2$. The following statements hold.*

(i) If C is closed, then d_x attains a minimizer over C.

(ii) If C is convex, then d_x has at most one minimizer over C.

Proof. (i) Note that d_x is strongly convex so that its level sets are bounded. Since d_x is bounded below, a minimizer over C of d_x belongs to $S_\alpha \cap C$, for some $\alpha \in \mathbb{R}$, which is compact by the closedness of C. The Weierstraß theorem 1.6 ensures the existence of a minimizer over C.

(ii) Suppose that d_x has two distinct minimizers u and v over C. The strict convexity of d_x implies that $\frac{1}{2}u + \frac{1}{2}v \in C$ but $d_x(\frac{1}{2}u + \frac{1}{2}v) < \frac{1}{2}d_x(u) + \frac{1}{2}d_x(v) = \inf_C d_x$, which is a contradiction. Therefore d_x has at most one minimizer over C. ∎

The above proposition guarantees that if $C \subset \mathbb{R}^n$ is nonempty, closed, and convex, then d_x has a unique minimizer over C for any $x \in \mathbb{R}^n$. This leads to the definition of a metric projection onto the set C.

Definition 5.2. Let $C \subset \mathbb{R}^n$ be nonempty, closed, and convex. Then the *metric projection over the set* C is the map $P_C : \mathbb{R}^n \to C$ given by

$$P_C(x) := \arg\min_{u \in C} \frac{1}{2}\|u - x\|^2$$

for all $x \in \mathbb{R}^n$.

The following theorem gives a geometric characterization for a metric projection operator.

Theorem 5.3. *Let $C \subset \mathbb{R}^n$ be nonempty, closed, and convex. Let $x \in \mathbb{R}^n$ and $z \in C$, then $z = P_C(x)$ if and only if the inequaltiy*

$$\langle x - z, y - z \rangle \leq 0 \tag{5.1}$$

for all $y \in C$.

Proof. Since finding the projection is a convex optimization problem, so its solution $z = P_C$ can be characterized by a variational inequality $\langle \nabla f(z), y - z \rangle \geq 0$ for all $y \in C$, where $f(u) := \frac{1}{2}\|u - x\|^2$. Since $\nabla f(u) = u - x$ for all $u \in \mathbb{R}^n$, the aforementioned variational inequality becomes (5.1). ∎

The above characterization also implies the Lipschitz continuity of a metric projection.

Proposition 5.4. *Let $C \subset \mathbb{R}^n$ be nonempty, closed, convex. Then P_C is nonexpansive, i.e. Lipchitz continuous with constant $L = 1$.*

Proof. Take any $x, y \in \mathbb{R}^n$. Then (5.1) gives

$$\langle x - P_C(x), P_C(y) - P_C(x) \rangle \leq 0$$

and

$$\langle y - P_C(y), P_C(x) - P_C(y) \rangle \leq 0.$$

Adding both inequalities, we obtain

$$\langle (x - y) + (P_C(y) - P_C(x)), P_C(y) - P_C(x) \rangle \leq 0,$$

which can be rearranged into

$$\begin{aligned} \|P_C(y) - P_C(x)\|^2 &\leq \langle y - x, P_C(y) - P_C(x) \rangle \\ &\leq \|y - x\|\|P_C(y) - P_C(x)\|. \end{aligned}$$

The result is thus proved. ∎

With the help of Theorem 5.3, we may further reformulate a variational inequality as a fixed point equation associated to the operator $u \mapsto P_C(u - \sigma \nabla f(u))$.

Lemma 5.5. *Let $C \subset \mathbb{R}^n$ be closed and convex, and $f : \mathbb{R}^n \to \mathbb{R}$ a differentiable function. Let $\sigma > 0$. Then $\overline{x}$ solves the variational inequality $\langle \nabla f(\overline{x}), y - \overline{x} \rangle \geq 0$ for all $y \in C$ if and only if $\overline{x} = P_C(\overline{x} - \sigma \nabla f(\overline{x}))$.*

Proof. Using Theorem 5.3, $\overline{x} = P_C(\overline{x} - s\nabla f(\overline{x}))$ if and only if

$$\langle (\overline{x} - s\nabla f(\overline{x})) - \overline{x}, y - \overline{x} \rangle \leq 0$$

for all $y \in C$, which is equivalent to the required variational inequality. ∎

Now we are ready to state and give a convergence analysis of the Projected Gradient Algorithm.

Algorithm 5.6. Projected Gradient Algorithm.

```
Initialization:
    Pick a start point $x^0 \in \mathbb{R}^n$ and an unconstrained descent step length $\sigma > 0$.
    Set $k \leftarrow 0$.
While: $k = 0$ or $\|d^k\| \neq 0$;
    $y^k \leftarrow P_C(x^k - \sigma \nabla f(x^k))$.
    $d^k \leftarrow y^k - x^k$.
    Determine the step-size according to the Armijo linesearch $t_k = \beta^i$ (i.e. $s = 0$), where $i \in \mathbb{N} \cup \{0\}$ is the smallest integer satisfying (3.1).
    $x^{k+1} \leftarrow x^k + t_k d^k$.
    Update $k \leftarrow k + 1$.
```

Let us emphasize again on the Armijo linesearch rule.

Discussion 5.7. Notice that the Armijo linesearch rule needs a new consideration here, since the earlier proof of its well-definedness requires the direction d^k to be a descent direction, i.e. $\langle \nabla f(x^k), d^k \rangle < 0$. In fact, it can be shown that such condition holds, thanks to Theorem 5.3.

Theorem 5.8. *Let $C \subset \mathbb{R}^n$ be nonempty, closed, and convex, and $f : \mathbb{R}^n \to \mathbb{R}$ a differentiable function. Let (x^k) be a sequence generated by the Projected Gradient Algorithm described above. If $\overline{x}$ is a limit point of the sequence (x^k), then it satisfies the variational inequality (4.1).*

Proof. Suppose that (x^k) is an infinite sequence. This means d^k and $\nabla f(x^k)$ are nonzero for all $k \in \mathbb{N}$. The Armijo's linesearch rule implies that

$$f(x^k) - f(x^{k+1}) = f(x^k) - f(x^k + t_k d^k) \geq -\alpha t_k \langle \nabla f(x^k), d^k \rangle. \tag{5.2}$$

Let $\overline{x} \in C$ be a limit point of (x^k) and choose a subsequence (x^{k_q}) of (x^k) so that $\lim_q x^{k_q} = \overline{x}$. In the above inequality, we see that the right hand side is strictly positive since $\nabla f(x^k)$ and d^k are nonzero and d^k is a descent direction by Discussion 5.7. This means $(f(x^k))$ is decreasing and converges to some $f^* \in \mathbb{R}$. Moreover, since f is continuous, we know that $f^* = \lim_q f(x^{k_q}) = f(\overline{x})$. This implies that the left hand side of (5.2) tends to 0. This further gives

$$\lim_q t_{k_q} \langle \nabla f(x^{k_q}), d^{k_q} \rangle = 0. \tag{5.3}$$

Recall from the definition of y^k's, we may see from the continuity of P_C (Proposition 5.4) that (y^{k_q}) is convergent to some $\overline{y} \in C$, and so is (d^{k_q}). Let $\overline{d} := \lim_q d^{k_q}$, we may see that $\overline{d} = \overline{y} - \overline{x}$.

We claim that $\langle f(\overline{x}), \overline{d} \rangle = 0$. In case $t_{k_q} \not\to 0$, we obtain from (5.3) that $\langle f(\overline{x}), \overline{d} \rangle = 0$. Otherwise if $g_{k_q} \to 0$, Theorem 1.12 implies that there is $s_{k_q} \in]0, t_{k_q}[$ in which

$$f(x^{k_q} + t_{k_q} d^{k_q}) - f(x^k) = \langle \nabla f(x^{k_q} + t_{k_q} d^{k_q}), d^{k_q} \rangle$$

for each $q \in \mathbb{N}$. Letting $q \to \infty$, we get $\langle f(\overline{x}), \overline{d} \rangle = 0$ as needed. From this and Theorem 5.3, we know that

$$0 \geq \langle (\overline{x} - \sigma \nabla f(\overline{x})) - \overline{y}, \overline{x} - \overline{y} \rangle = \sigma \langle \nabla f(\overline{x}), \overline{d} \rangle + \|\overline{d}\|^2 = \|\overline{d}\|^2,$$

showing that $\overline{d} = 0$ and hence $\overline{x} = \overline{y} = P_C(\overline{x} - \sigma \nabla f(\overline{x}))$. Lemma 5.5 then brings us to the conclusion of the theorem. ∎

5.2 Duality and Uzawa's Algorithm

In this section, we consider the constrained optimization problem where the constraint set is described by equalities and inequalities. We shall focus on the case where the optimization problem is described by its KKT system. We shall see in the following that a KKT system provides a duality view of maximizing with respect to the multipliers against minimizing over the decision variables.

Let V and M be any two sets. Recall that a pair $(u, \lambda) \in V \times M$ is said to be a *saddle point* of a function $L : V \times M \to \mathbb{R}$ if

$$\inf_{v \in V} L(v, \lambda) = L(u, \lambda) = \sup_{\mu \in M} L(v, \mu).$$

Notice that the inequality

$$\sup_{\mu \in M} \inf_{v \in V} L(v, \mu) \leq \inf_{v \in V} \sup_{\mu \in M} L(v, \mu)$$

always hold, but the reverse inequality is not true in general. However, it holds at the saddle point. Indeed, if (u, λ) is a saddle point of L, then

$$\inf_{v \in V} \sup_{\mu \in M} L(v, \mu) \leq \sup_{\mu \in M} L(u, \mu) = L(u, \lambda) = \inf_{v \in V} L(v, \lambda) \leq \sup_{\mu \in M} \inf_{v \in V} L(v, \mu).$$

From this simple observation, we have the following propostion.

Proposition 5.9. *If (u, λ) is a saddle point of a function $L : V \times M \to \mathbb{R}$, then*

$$\sup_{\mu \in M} \inf_{v \in V} L(v, \mu) = L(u, \lambda) = \inf_{v \in V} \sup_{\mu \in M} L(v, \mu). \tag{5.4}$$

Now, we associate with the problem $Opt(f, g)$, where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^r$, a *Lagrangian function* $L : \mathbb{R}^n \times \mathbb{R}^r_+ \to \mathbb{R}$ given by

$$L(v, \mu) := f(v) + \sum_{i=1}^{r} \mu_i g_i(v) = f(v) + \mu^\top g(v).$$

for $(v, \mu) \in \mathbb{R}^n \times \mathbb{R}^r_+$. We next give a relationship between $Opt(f, g)$ and the saddle points of its Lagragian function.

Theorem 5.10. *If $(u, \lambda) \in \mathbb{R}^n \times \mathbb{R}^r_+$ is a saddle point of the Lagrangian L, then the point u is feasible and optimal to $Opt(f, g)$.*

Proof. Notice that $L(u, \mu) \leq L(u, \lambda)$ can be expressed as

$$\sum_{i=1}^{r} (\mu_i - \lambda_i) g_i(u) \leq 0$$

for all $\mu \in \mathbb{R}^r_+$. This implies that $g_i(u) \leq 0$ for all $i = 1, \cdots, r$, showing that u is feasible. Taking $\mu = 0$ in the above inequality, we get $\sum_{i=1}^{r} g_i(u) \geq 0$. Therefore, we have $\sum_{i=1}^{r} \lambda_i g_i(u) = 0$. Finally, for any feasible point $v \in \mathbb{R}^n$, we have

$$f(u) = f(u) + \sum_{i=1}^{r} \lambda_i g_i(u) = L(u, \lambda) \leq L(v, \lambda) = f(v) + \sum_{i=1}^{r} \lambda_i g_i(v) \leq f(v),$$

where the last inequality follows from the feasibility of v. ∎

Theorem 5.11. *Suppose that f and g_i's are all convex continuously differentiable functions, and that LICQ is satisfied for $Opt(f,g)$. If $u \in Opt(f,g)$, then there exists $\lambda \in \mathbb{R}^r_+$ such that (u,λ) is a saddle point of the Lagrangian L.*

Proof. Since $u \in Opt(f,g)$ and the constraint is qualified with LICQ, there exist multipliers $\lambda = (\lambda_1, \cdots, \lambda_r) \in \mathbb{R}^r_+$ satisfying $\nabla_1 L(u,\lambda) = 0$ and $\lambda_i g_i(u) = 0$ for all $i = 1, \cdots, r$. Since $L(\cdot, \lambda)$ is convex, it follows that $L(u,\lambda) = \inf_{v\in\mathbb{R}^n} L(v,\lambda)$. Moreover, the feasibility of u and the complementarity condition imply that

$$L(u,\mu) = f(u) + \sum_{i=1}^{r} \mu_i g_i(u) \leq f(u) = f(u) + \sum_{i=1}^{r} \lambda_i g_i(u) = L(u,\lambda)$$

for all $\mu \in \mathbb{R}^r_+$. Hence, (u,λ) is a saddle point of L. ∎

Looking at the above Thoerems 5.10 and 5.11, if a saddle point (u,λ) for the Lagrangian function can be calculated, then we disgard λ and keep u as the solution of $Opt(f,g)$ that we are after. In view of Proposition 5.9, the vector λ can be obtained solving the following maximization problem

$$\max_{\mu\in\mathbb{R}^r_+} \ q(\mu), \qquad q(\mu) := \inf_{v\in\mathbb{R}^n} L(v,\mu).$$

This problem is known as the *dual problem* of $Opt(f,g)$, where we now refer to $Opt(f,g)$ as the *primal problem*. Likewise, the decision variable for the primal problem is called *primal variable*, while the one for the dual problem is called *dual variable*. Note that the dual problem may be viewed as a constrained problem over $\mathbb{R}^r$ with the constraints $\mu_i \geq 0$ for all $i = 1, \cdots, r$, which can be handled pretty easily. We have the following result, which follows from the existence of multipliers in the KKT.

Theorem 5.12. *Suppose that $Opt(f,g)$ has a solution, where f and g_i's are all convex continuous differentiable functions. Then L has a saddle point.*

The Uzawa's algorithm is just the Projected Gradient Algorithm applied to the dual problem in the fashion that: to each dual variable $\mu \in \mathbb{R}^r_+$, the parametrized unconstrained problem $Opt(L(\cdot,\mu))$ has a solution $u_\mu \in \mathbb{R}^n$. Then at each iteration $k \in \mathbb{N} \cup \{0\}$, the update on the dual variable λ^k (obtained by the projected steepest ascent) gives rise to the primal variable $u^k \in Opt(L(\cdot,\lambda^k))$. As $k \to \infty$, we wish that (u^k, λ^k) would converge to a saddle point of L, or at least u^k would converge to a solution of $Opt(f,g)$.

The Uzawa's algorithm is available due to the nice explicit formulae of the projection $P_+ := P_{\mathbb{R}^r_+} : \mathbb{R}^r \to \mathbb{R}^r_+$ and the gradient ∇q of the dual objective function q when u_μ is available for each μ. In particular, we have the following coordinatewise expressions

$$P_+(\nu)_i = \max\{\nu_i, 0\} \quad \text{and} \quad \nabla q(\mu)_i = g_i(u_\mu)$$

for each $i = 1, \cdots, r$.

Algorithm 5.13. Uzawa's Algorithm.

```
Initialization:
    Pick a start point $\lambda^0 \in \mathbb{R}^r_+$ and an unconstrained descent step length $\sigma > 0$.
    Set $k \leftarrow 0$.
While: $\|d^k\| \neq 0$;
    $u^k$ : Chosen from $\arg\min_{v\in\mathbb{R}^n} L(\cdot, \lambda^k)$.
    $\lambda^{k+1} \leftarrow P_+(\lambda^k + \sigma\nabla q(\lambda^k))$,
            i.e. $\lambda_i^{k+1} = \max\{0, \lambda_i^k + \sigma g_i(u^k)\}$ for $i = 1, \cdots, r$.
    Update $k \leftarrow k + 1$.
```

We then consider the convergence of the Uzawa's algorithm for *Opt*(f, g) with linear inequality constraints.

Theorem 5.14. *Consider the problem Opt(f, g) such that f is α-strongly convex and $g(v) = Av - b$ for each $v \in \mathbb{R}^n$, where $A \in \mathbb{R}^{r\times n}$ and $b \in \mathbb{R}^r$. Suppose that the constraint set in nonempty and $\sigma \in (0, 2\alpha/\|A\|^2)$. Then the sequence (u^k) generated from the Uzawa's algorithm is convergent to the unique solution of Opt(f, g). Moreover, if* $\operatorname{rank}(A) = r$*, then (λ^k) also converges to the unique solution of the dual problem.*

Proof. Note first that *Opt*(f, g) is the problem of minimizing a strongly convex function over a closed convex set, hence it has a unique solution, denoted with u. Moreover, for each $\mu \in \mathbb{R}^r_+$, $L(\cdot, \mu)$ is also strongly convex so that u_μ exists and is unique. According to Theorem 5.11, there is $\lambda \in \mathbb{R}^r_+$ in which (u, λ) is a saddle point of L. Hence we have

$$\nabla_1 L(u, \lambda) = \nabla f(u) + A^\top \lambda = 0$$

since $L(u, \lambda) = \inf_{v\in\mathbb{R}^n} L(v, \lambda)$, and also

$$\langle g(u), \mu - \lambda\rangle \leq 0 \qquad (\forall \mu \in \mathbb{R}^r_+)$$

since $L(u, \lambda) = \sup_{\mu\in\mathbb{R}^r_+} L(u, \mu)$. The latter inequality is equivalent to

$$\langle (\lambda + \sigma g(u)) - \lambda, \mu - \lambda\rangle \leq 0$$

for all $\mu \in \mathbb{R}^r_+$ and any fixed $\sigma > 0$. Hence, the saddle condition is equivalent to the system

$$\begin{cases} \nabla f(u) + A^\top \lambda = 0, \\ \lambda = P_+(\lambda + \sigma g(u)). \end{cases}$$

Let $k \in \mathbb{N} \cup \{0\}$. With similar calculation applied to u^k and λ^{k+1}, we analogously obtain

$$\begin{cases} \nabla f(u^k) + A^\top \lambda^k = 0, \\ \lambda^{k+1} = P_+(\lambda^k + \sigma g(u^k)). \end{cases} \tag{5.5}$$

Combining the two systems and using the nonexpansivity of P_+, we get

$$\begin{cases} \nabla f(u^k) - \nabla f(u) + A^\top(\lambda^k - \lambda) = 0, \\ \|\lambda^{k+1} - \lambda\| \le \|\lambda^k - \lambda + \sigma A(u^k - u)\|. \end{cases}$$

From the above system, the Apollonius identity, the Cauchy-Schwarz inequality, and α-strong monotonicity of ∇f, we get

$$\begin{aligned} \|\lambda^{k+1} - \lambda\|^2 &\le \|\lambda^k - \lambda\| + 2\sigma\langle A^\top(\lambda^k - \lambda), u^k - u\rangle + \sigma^2\|A(u^k - u)\|^2 \\ &\le \|\lambda^k - \lambda\| - 2\sigma\langle \nabla f(u^k) - \nabla f(u), u^k - u\rangle + \sigma^2\|A\|^2\|A(u^k - u)\|^2 \\ &\le \|\lambda^k - \lambda\| - 2\sigma\alpha\|u^k - u\|^2 + \sigma^2\|A\|^2\|A(u^k - u)\|^2 \\ &= \|\lambda^k - \lambda\| - \sigma(2\alpha - \sigma\|A\|^2)\|A(u^k - u)\|^2. \end{aligned} \tag{5.6}$$

Since $\sigma \in \left(0, \frac{2\alpha}{\|A\|^2}\right)$, the above inequality reduces to $\|\lambda^{k+1} - \lambda\| \le \|\lambda^k - \lambda\|$. This means $(\|\lambda^k - \lambda\|)$ is decreasing and bounded below, so it is convergent. Rearranging (5.6), we obtain

$$\sigma(2\alpha - \sigma\|A\|^2)\|A(u^k - u)\|^2 \le \|\lambda^k - \lambda\| - \|\lambda^{k+1} - \lambda\|^2.$$

Letting $k \to \infty$, the right hand side goes to 0 and we obtain the desired convergence $u^k \to u$.

Next, we show that (λ^k) is convergent when $\operatorname{rank}(A) = r$. Since $\operatorname{rank}(A) = \operatorname{rank}(A) = \operatorname{rank}(A^\top) = r$ if and only if the linear system

$$A^\top \nu = b \tag{5.7}$$

has a unique solution (in ν) for any $b \in \mathbb{R}^n$. For $b = -\nabla f(u)$, the unique solution is the dual optimum point λ. Now, since (λ^k) is bounded, it possesses a convergent subsequence (λ^{k_q}) with a limit $\lambda' \in \mathbb{R}^r_+$. From (5.5), we have $\nabla f(u^{k_q}) + A^\top \lambda^{k_q} = 0$.

Letting $q \to \infty$, we obtain $\nabla f(u) + A^\top \lambda' = 0$. Since (5.7) has a unique solution, it must be the case that $\lambda' = \lambda$. Since this process can be repeated with any subsequence of (λ^k), the entire sequence (λ^k) is convergent to λ. ■

It should be observed that we need to solve a minimization problem at each iterate k to obtain u^k. This can be a little uncomfortable, but at least this subproblem is unconstrained. Moreover, since the problem at hand is a convex program, this reduces to finding a critical point. In particular, when $f(v) = \frac{1}{2} v^\top C v - d^\top v$ is strictly convex, the point u^k can be obtained by solving the linear system $Cv - d + A^\top \lambda^k = 0$.

Discussion 5.15. Derive the Uzawa's algorithm for a strongly convex program with linear inequality and equality constraints. Can we guarantee the convergence?

Discussion 5.16. Apply the Uzawa's algorithm to obtain the optimal separating hyperplane in the SVM model, as discussed in Discussion 4.25.