Non-Linear Optimization 447/550 Notes

Farid Rajkotia Zaheer

1 Mathematical Preliminaries

Theorem 1.1. *(**Spectral Factorization Theorem:**) Let $A \in \mathbb{R}^{n\times n}$ be a $n \times n$ symmetric matrix. Then there exists an orthogonal matrix U i.e. $UU^T = U^TU = I$ and a diagonal matrix D such that*

$$U^TAU = D.$$

Here the columns of U are eigenvectors of A and form an orthogonal basis. The entries of D are the eignvalues of A.

Theorem 1.2. *(**Mean Value Theorem**) Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable. Then, for every $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$ in $\mathbb{R}^n$ there exists $\boldsymbol{z} \in (\boldsymbol{x}, \bar{\boldsymbol{x}})$ where $(\boldsymbol{x}, \bar{\boldsymbol{x}}) := \{x + s(\bar{\boldsymbol{x}} - \boldsymbol{x}) : 0 < s < 1\}$, is the line segment joining $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$ excluding the end points such that,*

$$f(\boldsymbol{x}) = f(\bar{\boldsymbol{x}}) + \nabla f(\boldsymbol{z})^T(\boldsymbol{x} - \bar{\boldsymbol{x}}).$$

Theorem 1.3. *(**Linear Approximation Theorem**) Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable. Then, for every $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$ in $\mathbb{R}^n$ there exists $\boldsymbol{z} \in (\boldsymbol{x}, \bar{\boldsymbol{x}})$*

$$f(x) = f(\bar{x}) + \nabla f(\bar{x})^T(x - \bar{x}) + \frac{1}{2}(x - \bar{x})^T\nabla^2 f(z)(x - \bar{x}).$$

Theorem 1.4. *(**Quadratic Approximation Theorem**)*

2 Graphical Optimization

A neat method for solving both linear and non-linear programming problems up to 3 dimensions can be done by plotting out the feasible region of the problem at hand. Explicitly, we may apply this technique if the number of variables in the objective function does not exceed 2.

The following are the steps needed to obtain a graphical solution:

1. Draw the boundaries of each of the constraint functions. The region that is obtained is the feasible set.
2. Draw several contours of the objective function and determine in which direction the function is decreasing/increasing (depending if the optimization is minimizing or maximizing).
3. Select the point that belongs to the feasible set and is the lowest/highest value of the objective function.

In order to determine in which direction the objective function is decreasing or increasing can be resolved by recalling the directional derivative from vector calculus.

Definition 2.1. ***(Directional Derivative)*** *Let $f : \mathbb{R}^n \to R$, the directional derivative of f at x in the direction d is,*

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

Given the limit exists and is finite.

Notice that if $f'(x : d) > 0$ then for $t > 0$ small enough we have that $f(x + td) > f(x)$. So we deduce that the function value increases in the direction d as we vary over small enough step size t.

Then we know the gradient of f in the direction of some vector d is given as,

$$\nabla f(x)^T d = \nabla f(x) \cdot d = \|\nabla f(x)\| \|d\| \cos\theta.$$

So, $\nabla f(x)^T d > 0$ iff $\theta \in [0, 2\pi)$.

Definition 2.2. ***(Level Set)*** *Let $f : \mathbb{R}^n \to \mathbb{R}$, the level set of f is the set of all points where the function takes on a constant value, c.*

When $n = 2$, the level set is a level curve, when $n = 3$, the level set is a level surface etc.

Recall the well known fact from vector calculus,

Theorem 2.3. *∇f is a vector perpendicular to each point of the level set of f.*

It is important to understand that it is not neccesarily the case that the solution to the optimization problem is also the unconstrained optimum. Note, that the unconstrained optimum, is the minimizer or maximizer of the objective function regardless of the constraints i.e. independent of the feasible set.

We may now turn our attention to looking at the setup of a non-linear programming problem and how to go about solving it graphically.

Example 2.4. *Consider the non-linear minimization problem*

$$\begin{aligned} \min \quad & f(x, y)x^2 + y^2 - 4y + 4 \\ s.t \quad & g_1(x, y)x^2 - y + 2 \leq 2 \\ & g_2(x, y)x + y - 6 \leq 0. \end{aligned} \tag{1}$$

We may plot the feasible region by plotting the boundaries of the constraint functions i.e. $x^2 - y + 2 = 2$ and $x + y - 6 = 0$.

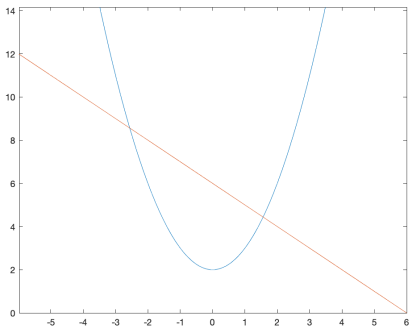

Figure 1: The feasible set is the bounded region.

By now plotting contours of the objective function we may find the optimal solution that minimizes the function.

3 Unconstrained Optimization

Unconstrained optimization refers to the set of linear or non-linear optimization problems where there are no constraints of any kind placed on the objective function. In other words, the feasible set of the problem is the entire space over which the objective function is defined in the problem. In such problems we seek to find global or local optima for the objective function.

3.1 Global and Local Optima

Consider the set $S \subset \mathbb{R}^n$ and the mapping $f : S \to \mathbb{R}$. Then,

Definition 3.1. *$\boldsymbol{x}^*$ is called the **global minimum point** of f over S if $f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*)$ for any $\boldsymbol{x} \in S$.*

Definition 3.2. *$\boldsymbol{x}^*$ is called a **strict global minimum point** of f over S if $f(\boldsymbol{x}) > f(\boldsymbol{x}^*)$ for any $\boldsymbol{x}^* \neq \boldsymbol{x} \in S$.*

Definition 3.3. *$\boldsymbol{x}^*$ is called the **global maximum point** of f over S if $f(\boldsymbol{x}) \leq f(\boldsymbol{x}^*)$ for any $\boldsymbol{x} \in S$.*

Definition 3.4. *$\boldsymbol{x}^*$ is called a **strict global maximum point** of f over S if $f(\boldsymbol{x}) < f(\boldsymbol{x}^*)$ for any $\boldsymbol{x}^* \neq \boldsymbol{x} \in S$.*

The set S on which the optimization of f is performed is called the *feasible set* and any point $\boldsymbol{x} \in S$ is called a *feasible solution*. For unconstrained optimization problems, S will be implicitly taken to be the whole space.

A vector $\boldsymbol{x}^* \in S$ is called a *global optimum* if it is either a global minimum or maximum.

The maximal and minimal values of f over S are defined as the supremum and infimum respectively,

$$\max\{f(\boldsymbol{x}) : \boldsymbol{x} \in S\} := \sup\{f(\boldsymbol{x}) : \boldsymbol{x} \in S\}$$
$$\min\{f(\boldsymbol{x}) : \boldsymbol{x} \in S\} := \inf\{f(\boldsymbol{x}) : \boldsymbol{x} \in S\}$$

Furthermore, the set of all global maximizers and minimizers of f over S is denoted respectively as,

$$\operatorname{argmax}\{f(\boldsymbol{x}) : \boldsymbol{x} \in S\}$$
$$\operatorname{argmin}\{f(\boldsymbol{x}) : \boldsymbol{x} \in S\}.$$

Keep in mind there could be several global minimal and maximal points, however minimal and maximal values are unique. This comes from the fact that global minimal and maximal points belong to the domain of the objective function whereas minimal and maximal values belong to the range.

As a trivial example, consider any constant function, $f : \mathbb{R}^n \to \mathbb{R}$ i.e. $f(\boldsymbol{x}) = c$. It is clear that the set of all minimizers and maximizers are equivalent and include all points in $\mathbb{R}^n$. However the maximal (and in this case minimal) value is unique namely, $c \in \mathbb{R}$.

Definition 3.5. *$\boldsymbol{x}^*$ is called the **local minimum point** of f over S if there exists $r > 0$ such that $f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*)$ for any $\boldsymbol{x} \in S \cap B(\boldsymbol{x}^*, r)$.*

Definition 3.6. *$\boldsymbol{x}^*$ is called the **strict local minimum point** of f over S if there exists $r > 0$ such that $f(\boldsymbol{x}) > f(\boldsymbol{x}^*)$ for any $\boldsymbol{x}^* \neq \boldsymbol{x} \in S \cap B(\boldsymbol{x}^*, r)$.*

Definition 3.7. *$\boldsymbol{x}^*$ is called the **local maximum point** of f over S if there exists $r > 0$ such that $f(\boldsymbol{x}) \leq f(\boldsymbol{x}^*)$ for any $\boldsymbol{x} \in S \cap B(\boldsymbol{x}^*, r)$.*

Definition 3.8. *$\boldsymbol{x}^*$ is called the **strict local maximum point** of f over S if there exists $r > 0$ such that $f(\boldsymbol{x}) < f(\boldsymbol{x}^*)$ for any $\boldsymbol{x}^* \neq \boldsymbol{x} \in S \cap B(\boldsymbol{x}^*, r)$.*

3.2 First Order Optimality Condition

We know that a for a differentiable one dimensional function $f(x)$ defined over some interval, if there exists some local maximum or minimum x^*, then $f'(x^*) = 0$. The multidimensional generalization states that the gradient of the function will be zero at local optima. We refer to this as the *first order optimality condition* since it relies on constraints on the first derivatives of a function.

Theorem 3.9. *(**First Order Neccesary Optimality Condition**). Let $f : U \to \mathbb{R}$ be a function defined on the set $U \subseteq \mathbb{R}^n$. Suppose that $\boldsymbol{x}^* \in \text{int}(U)$ is a local optimum point and that all partial derivatives of f exist at $\boldsymbol{x}^*$. Then $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.*

Proof. Let $i \in \{1, 2, ..., n\}$ and consider the one dimensional function $g(t) = f(\boldsymbol{x}^* + te_i)$. Notice that g is differentiable at $t = 0$ and that,

$$g'(0) = \frac{\partial f}{\partial x_i}(\boldsymbol{x}^*).$$

Since $\boldsymbol{x}^*$ is a local optimum of f it follows that $t = 0$ is a local optimum of g which implies that $g'(0) = 0$. So we gain the equality,

$$\frac{\partial f}{\partial x_i}(\boldsymbol{x}^*) = 0.$$

Furthermore, this will hold for all i in the index set and the claim follows i.e. $\nabla f(\boldsymbol{x}^*) = 0$. □

It should be noted the converse of the statement is not true i.e. there exist points that are not local optima that have gradient zero. For example $f(x) = x^3$.

We refer to points in the interior of the set whose gradient is zero as *stationary points*.

3.3 Matrix Classifications

Definition 3.10. *(**Positive Semidefinitness**).*

1. *A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is called **positive definite**, denoted $A \succ 0$, if $\boldsymbol{x}^T A\boldsymbol{x} > 0$ for all $\boldsymbol{x} \neq \mathbf{0} \in \mathbb{R}^n$.*
2. *A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is called **positive semidefinite**, denoted $A \succcurlyeq 0$, if $\boldsymbol{x}^T A\boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$.*

We may by symmetry derive the definition of negative definite and negative semidefinite matrices from Definition 3.10.

Definition 3.11. *(**Indefiniteness**) A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is called indefinite if there exist $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^n$ such that $\boldsymbol{x}^T A\boldsymbol{x} < 0$ and $\boldsymbol{y}^T A\boldsymbol{y} > 0$.*

It is not easy in general to check whether a matrix is positive definite or positive semidefinite (same for negative). One way to check is to characterize the eigenvalues of these matrices.

Theorem 3.12. *(**Eigenvalue Charachterization Theorem**). Let A be a symmetric $n \times n$ matrix then,*

(a) A is positive definite if and only if all its eigenvalues are positive.

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

(c) A is negative definite if and only if all its eigenvalues are negative.

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

(e) A is indefinite if and only if it has at least one positive eigenvalue and at least one negative eigenvalue.

Moreover, without computer assistance it is still cumbersome to use eigenvalue characterization to determine the definitness of a matrix as all eigenvalues need to be calculated. As one can tell, without computer assitance calculating eigenvalues for any matrix greater than 3×3 is tedious and error prone.

An alternate method is through *Leading Principal Minors*. Recall some terminology from linear algebre, namely,

- The **leading principal submatrix of order** k of an $n \times n$ matrix $\boldsymbol{A}$ is obtained by deleting the last $n-k$ rows and columns of $\boldsymbol{A}$.
- The determinant of the leading principal submatrix is called the **leading principal minor** of $\boldsymbol{A}$.

Theorem 3.13. *(**Leading Principal Minors Criterion**) A matrix is*

- *positive definite if and only if all its leading principal minors and positive.*
- *negative definite if and only if all its odd principal minors are negative and even principal minors are positive.*
- *indefinite if one of its kth order leading principal minors is negative for an even k or if there are two odd leading principal minors that have different signs.*

We may articulate conditions for positive/negative semi-definitness by changing positive/negative to non-negative/non-positive respectively.

Yet another check of matrix definiteness is by checking if the matrix is diagonally dominant. Such techniques can easily be found in any text on Linear Algebra. The moral of the story here is a that in general to check matrix definitenss is a hard computational problem especially in practical problems that arise in applied mathematics. As such there are various techniques one can use to test matrix definiteness.

3.4 Second Order Optimality Conditions

We state the necessary and sufficent second order optimality conditions.

Theorem 3.14. *(**necessary second order optimality conditions**). Let $f : U \to \mathbb{R}$ be a function defined on the open set $U \subseteq \mathbb{R}^n$. Suppose f is twice continuously differentiable over U and that $\boldsymbol{x}^*$ is a sationary point. Then the following hold:*

(a) If $\boldsymbol{x}^*$ is a local minimum point of f over U, then $\nabla^2 f(\boldsymbol{x}^*) \succcurlyeq 0$.

(b) If $\boldsymbol{x}^*$ is a local maximum point of f over U, then $\nabla^2 f(\boldsymbol{x}^*) \preccurlyeq 0$.

Proof. (a) Since $\boldsymbol{x}^*$ is a local minimum point of f over U there exists a ball centred at $\boldsymbol{x}^*$ with radius r i.e. $B(\boldsymbol{x}^*, r) \subseteq U$ for which $f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*)$ for all $\boldsymbol{x} \in B(\boldsymbol{x}^*, r)$. Let $\boldsymbol{y}$ be a non-zero vector in $\mathbb{R}^n$. For any $0 < \alpha < \frac{r}{\|\boldsymbol{y}\|}$, we have that,

$$\boldsymbol{x}^*_\alpha \equiv \boldsymbol{x}^* + \alpha \boldsymbol{y} \in B(\boldsymbol{x}^*, r).$$

Hence, for any such α we also have that,

$$f(\boldsymbol{x}^*_\alpha) \geq f(\boldsymbol{x}^*).$$

By the Linear Approximation Theorem, we also know there exists a non-zero vector $\boldsymbol{z}_\alpha \in [\boldsymbol{x}^*, \boldsymbol{x}^*_\alpha]$ such that,

$$f\left(\boldsymbol{x}^*_\alpha\right) - f\left(\boldsymbol{x}^*\right) = \nabla f\left(\boldsymbol{x}^*\right)^T \left(\boldsymbol{x}^*_\alpha - \boldsymbol{x}^*\right) + \frac{1}{2}\left(\boldsymbol{x}^*_\alpha - \boldsymbol{x}^*\right)^T \nabla^2 f\left(\boldsymbol{z}_\alpha\right)\left(\boldsymbol{x}^*_\alpha - \boldsymbol{x}^*\right).$$

Since $\boldsymbol{x}^*$ is a stationary point, the equation reduces to,

$$f(\boldsymbol{x}^*_\alpha) - f(\boldsymbol{x}^*) = \frac{\alpha}{2} \boldsymbol{y}^T \nabla^2 f(\boldsymbol{z}_\alpha) \boldsymbol{y}.$$

It then follows for any $\alpha \in (0, \frac{r}{\|\boldsymbol{y}\|})$ the inequality $\boldsymbol{y}^T \nabla^2 f(\boldsymbol{z}_\alpha)\boldsymbol{y} \geq 0$ holds. Finally, using the fact that $\boldsymbol{z}_\alpha \to \boldsymbol{x}^*$ as $\alpha \to 0^+$ and the continuity of the Hessian, we obtain that,

$$\boldsymbol{y}^T \nabla^2 f(\boldsymbol{x}^*)\boldsymbol{y} \geq 0.$$

Since this will hold for any $\boldsymbol{y} \in \mathbb{R}^n$, the claim is established.

(b) The proof for (b) follows directly by applying the procedure above to $-f$.

□

It is important to keep in mind the distinction between the necessary and sufficient conditions for second order optimality. Theorem 3.14 is valid only when given a stationary point with the information that it is a local max or min of f over the set on which it is defined.

Note that Theorem 3.14 implies the stronger property of strict local optimality.

Definition 3.15. *(**Saddle Point**) Let $f : U \to \mathbb{R}$ be a function defined on the open set $U \subseteq \mathbb{R}^n$. Suppose that f is continuously differentiable over U. A stationary point $\boldsymbol{x}^*$ is called a saddle point if it is neither a local minimum point or local maximum point of f over U.*

Theorem 3.16. *(**sufficient second order optimality conditions**). Let $f : U \to \mathbb{R}$ be a function defined on the open set $U \subseteq \mathbb{R}^n$. Suppose f is twice continuously differentiable over U and that $\boldsymbol{x}^*$ is a sationary point. Then the following hold:*

(a) If $\nabla^2 f(\boldsymbol{x}^*) \succ 0$, then $\boldsymbol{x}^*$ is a strict global minimum point of f over U.

(b) If $\nabla^2 f(\boldsymbol{x}^*) \prec 0$, then $\boldsymbol{x}^*$ is a strict global maximum point of f over U.

Proof. We will prove part (a), part (b) follows dually by using the same argument for $-f$. We know that $\boldsymbol{x}^*$ is a stationary point satisfying $\nabla^2 f(\boldsymbol{x}^*) \succ 0$. Since the Hessian is continuous it follows that there exists a ball $B(\boldsymbol{x}^*, r) \subseteq U$ such that $\nabla^2 f(\boldsymbol{x}^*) \succ 0$ for any $\boldsymbol{x} \in B(\boldsymbol{x}^*, r)$. By the Linear Approximation Theorem, we have that for any $\boldsymbol{x} \in B(\boldsymbol{x}^*, r)$ there exists a vector $\boldsymbol{z}_{\boldsymbol{x}} \in [\boldsymbol{x}^*, \boldsymbol{x}]$ for which,

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) = \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^T \nabla^2 f(\boldsymbol{z}_{\boldsymbol{x}})(\boldsymbol{x} - \boldsymbol{x}^*).$$

Since $\nabla^2 f(\boldsymbol{z}_{\boldsymbol{x}}) \succ 0$, it follows that for any $\boldsymbol{x} \in B(\boldsymbol{x}^*, r)$ such that $\boldsymbol{x} \neq \boldsymbol{x}^*$ the inequality $f(\boldsymbol{x}) > f(\boldsymbol{x}^*)$ holds. The claim follows as this means that $\boldsymbol{x}^*$ is a strict local minimum of f over U. □

Theorem 3.17. *(**Sufficient Condition for a Saddle Point**) Let $f : U \to \mathbb{R}$ be a function defined on the open set $U \subseteq \mathbb{R}^n$. Suppose that f is twice continuously differentiable over U and that $\boldsymbol{x}^*$ is a stationary point. If $\nabla^2 f(\boldsymbol{x}^*)$ is an indefinite matrix, then $\boldsymbol{x}^*$ is a saddle point of f over U.*

An important issue to consider is wether a function has a global minimizer or maximizer. This is an issure regarding attainment or existence. In order to develop tools to investigate this, we recall some important properties of closed and bounded sets on $\mathbb{R}^n$.

Theorem 3.18. *(**Weiestrass' Theorem**) Let f be a continuous function defined over a non-empty, compact set $D \subset \mathbb{R}^n$. Then there exists a global minimum of f overe D.*

Recall from elementary analysis, $D \subset \mathbb{R}^n$ being closed and bounded implies D is compact. So, given f is continuous, we will always be guaranteed that there exists a global optimum in D.

Note, that Weiestrass' Theorem shows the existence of such a point, it does not however say anything about the attainment of the optima. Furthermore, when the set D is not compact, Weiestrass' theorem does not guarantee any attainment of extrema. However we may still recover information from the following definition.

Definition 3.19. *(**Coerciveness**) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function. The function is called coercive if,*

$$\lim_{\|\boldsymbol{x}\| \to \infty} f(\boldsymbol{x}) = \infty.$$

Coercive functions exemplify the fact that the value of $f(x)$ cannot remain bounded when defined over an unbounded set $A \subseteq \mathbb{R}^n$ and f coercive.

The important and most applied property (for our purposes) of coercive functions is that they always attain their global minima when defined over closed sets. This leads to the following fact.

Theorem 3.20. *(**Coercive Functions and Existence of Global Minima**) Let $f : D \subset \mathbb{R}^n \to \mathbb{R}$ such that f is coercive. Then, $f(x)$ has atleast one global minimum point.*

Proof. Pick $\boldsymbol{x}_0 \in \mathbb{R}^n$ and set $f(\boldsymbol{x}_0) = M$. By the definition of coercivness it follows that there exists $R_M \geq 0$ such that,

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0), \ \forall \|\boldsymbol{x}\| \geq R_M.$$

Observe that the global minimum will not be in the set $\{\boldsymbol{x} : \|\boldsymbol{x}\| > R_M\}$. However, since the set $\{\boldsymbol{x} : \|\boldsymbol{x}\| \leq R_M\}$ is closed and bounded, by Weiestrass' Theorem we are guaranteed that there is a global minimum. The claim follows as we have established that the global minimum exists. □

3.5 Global Optimality Conditions

Theorem 3.21. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function such that $\nabla^2 f(\boldsymbol{x}) \succcurlyeq 0$ $\forall \boldsymbol{x} \in \mathbb{R}^n$. Let $\boldsymbol{x}^*$ be a stationary point of $f(\boldsymbol{x})$, then $\boldsymbol{x}^*$ is a global minimum of f over $\mathbb{R}^n$.*

We may derive the definition of global maximum by a symmetric argument. Given that the Hessian is negative semi-definite $\forall \boldsymbol{x} \in \mathbb{R}^n$, then a stationary point $\boldsymbol{x}^*$ is a global maximum.

Note here the crucial quantifier in Theorem 3.21, "for all $\boldsymbol{x}$ in $\mathbb{R}^n$". In other words, given that we know the Hessian of the objective function is positive semi-definite for all $\boldsymbol{x}$ in $\mathbb{R}^n$ and we have a stationary point $\boldsymbol{x}^*$, only then can we classify $\boldsymbol{x}^*$ as a global minimum point of f.

3.6 Quadratic Functions

Definition 3.22. *(**Quadratic Function**) A quadratic function over $\mathbb{R}^n$ is a function of the form,*

$$f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} + 2\boldsymbol{b}\boldsymbol{x}^T + c,$$

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

We also gain the helpful identities that are straightforward to prove,

$$\nabla f(\boldsymbol{x}) = 2\boldsymbol{A}\boldsymbol{x} + 2\boldsymbol{b}$$
$$\nabla^2 f(\boldsymbol{x}) = 2\boldsymbol{A}.$$

We may also deduce the following important properties of quadratic functions.

Lemma 3.23. *Let $f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} + 2\boldsymbol{b}\boldsymbol{x}^T + c$ such that $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is symmetric, $\boldsymbol{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Then,*

(a) $\boldsymbol{x}$ is a stationary point of f if and only if $\boldsymbol{A}\boldsymbol{x} = -\boldsymbol{b} \implies \boldsymbol{x} = -\boldsymbol{A}^{-1}\boldsymbol{b}$.

(b) If $\boldsymbol{A} \succcurlyeq 0$, then $\boldsymbol{x}$ is a global minimum point of f if and only if $\boldsymbol{A}\boldsymbol{x} = -\boldsymbol{b}$.

(c) If $\boldsymbol{A} \succ 0$, then $\boldsymbol{x} = -\boldsymbol{A}^{-1}\boldsymbol{b}$ is a strict global minimum point of f.

Following from (c), we may also gain the minimal value of f by $f(\boldsymbol{x}) = c - \boldsymbol{b}^T \boldsymbol{A}^{-1} \boldsymbol{b}$.

Lemma 3.23 follows directly as a consequence of the Global Optimality Conditions.

Lemma 3.24. *Let $f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} + 2\boldsymbol{b}\boldsymbol{x}^T + c$ such that $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is symmetric, $\boldsymbol{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Then f is coercive if and only if $\boldsymbol{A} \succ 0$.*

4 Least Squares

4.1 Solution to Over Determined Systems

Consider the linear system $\boldsymbol{A}\boldsymbol{x} \approx \boldsymbol{b}$, such that $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. We assume that $\boldsymbol{A}$ has a full column rank.

When $m > n$ i.e. there are more equations than unknowns, the system is inconsistent. The classic approach to solving such kinds of problems is through **Least Squares**. Solutions of the form,

$$\boldsymbol{x}_{LS} = \min \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2.$$

Alternatively, the least squares problem is the same as,

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \{f(\boldsymbol{x}) \equiv \boldsymbol{x}^T \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{b}^T \boldsymbol{A}\boldsymbol{x} + \|\boldsymbol{b}\|^2\}.$$

Recall that $\nabla^2 = 2\boldsymbol{A}^T\boldsymbol{A} \prec 0$.

So the unique, optimal solution $\boldsymbol{x}_{LS}$ is the solution $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$,

$$\left(\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{x}_{LS} = \boldsymbol{A}^T\boldsymbol{b} \implies \boldsymbol{x}_{LS} = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{b}.$$

4.2 Data Fitting

An important application of least squares is in linear data fitting. Consider data points $(\boldsymbol{s}_i, t_i)$ where $i = 1, 2, 3, ..., m$ where $s_i \in \mathbb{R}^n$ and $t_i \in \mathbb{R}$. Assume a linear relation of the form,

$$t_i = \boldsymbol{s}_i^T \boldsymbol{x},$$

holds is an approximate sense. In least squares we want to find the parameters vector $\boldsymbol{x} \in \mathbb{R}^n$ that solves,

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \sum_{i=1}^{m} \left(\boldsymbol{s}_i^T\boldsymbol{x} - t_i\right)^2 = \min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{S}\boldsymbol{x} - \boldsymbol{t}\|^2.$$

4.2.1 Least Squares in Non-linear Data Fitting

The least squares appraoch can also be used in non-linear fitting. Suppose we are given a set of points in $\mathbb{R}^2$ of the form (x_i, y_i), where $i = 1, 2, ..., m$. Furthermore, we know apriori that these points are approximately related via a poilynomial of degree at most d,

$$\sum_{i=0}^{d} a_j u_i^j \approx y_i.$$

The least squares approach here seeks $a_0, a_1, ..., a_d$ that are the least squares solution to the linear system,

$$\underbrace{\begin{pmatrix} 1 & u_1 & u_1^2 & \cdots & u_1^d \\ 1 & u_2 & u_2^2 & \cdots & u_2^d \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & u_m & u_m^2 & \cdots & u_m^d \end{pmatrix}}_{\mathbf{U}} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_m \end{pmatrix}$$

Notice then, the solution is well-defined if the $m \times (d+1)$ matrix if of full column rank. The matrix $\boldsymbol{U}$ consists of the first $d+1$ columns of the vandermonde matrix.

4.3 Regularized Least Squares

There are many instances where the least squares solution does not give rise to a good enough approximate solution. For example in systems where there are more variables than equations.

In such cases we must modify the least squares technique to incorporate a *regularization function* $R(\cdot)$. The problem takes the form.

$$\min_{\boldsymbol{x}} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 + \lambda R(\boldsymbol{x}).$$

$\lambda > 0$ and is called the *regularization parameter*. As λ gets larger, more weight is given to the regularization function.

In many cases $R(\boldsymbol{x})$ is taken to be quadratic i.e. $R(\boldsymbol{x}) = \|\boldsymbol{D}\boldsymbol{x}\|^2$ such that $\boldsymbol{D} \in \mathbb{R}^{p\times n}$ is a given matrix. The aim of the regularization function is to then control the norm of $\boldsymbol{D}\boldsymbol{x}$.

$$\min_{\boldsymbol{x}} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 + \lambda\|\boldsymbol{D}\boldsymbol{x}\|^2.$$

Expressing the problem equivalently as,

$$\min_{\boldsymbol{x}}\{f_{RLS}(\boldsymbol{x}) \equiv \boldsymbol{x}^T\left(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{D}^T\boldsymbol{D}\right)\boldsymbol{x} - 2\boldsymbol{b}^T\boldsymbol{A}\boldsymbol{x} + \|\boldsymbol{b}\|^2\}.$$

Since f_{RLS} is quadratic the Hessian is,

$$\nabla^2 f_{RLS}(\boldsymbol{x}) = 2\left(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{D}^T\boldsymbol{D}\right) \succcurlyeq 0.$$

So, an stationary point is a global minimum. Furthermore, the stationary points satisfy $\nabla f_{RLS} = \boldsymbol{0}$ i.e.

$$\left(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{D}^T\boldsymbol{D}\right)\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{b}.$$

Hence, if $\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{D}^T\boldsymbol{D} \succ 0$, then the RLS solution is given by,

$$\boldsymbol{x}_{RLS} = \left(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{D}^T\boldsymbol{D}\right)^{-1}\boldsymbol{A}^T\boldsymbol{b}.$$

4.4 Denoising

A common area where RLS is implemented is in denoising data. Suppose that a a noisy singnal of a measurment is given as,

$$\boldsymbol{b} = \boldsymbol{x} + \boldsymbol{w}.$$

Here $\boldsymbol{x} \in \mathbb{R}^n$ is an unknown signal and $\boldsymbol{w} \in \mathbb{R}^n$ is noise. Given $\boldsymbol{b}$, we want to find a good estimate for $\boldsymbol{x}$. However, the problem,

$$\min \|\boldsymbol{b} - \boldsymbol{x}\|^2,$$

is meaningless as the optimal solution is clearly $\boldsymbol{b} = \boldsymbol{x}$.

In order to articulate a more relevant problem, we must add a regularization term. Supposing we have some a-priori information regarding the signal. For example, we know that is it smooth in some sense. Then, our regularization function can be quadratic. Specifically, the sum of squares of consecutive differences i.e.

$$R(\boldsymbol{x}) = \sum_{i=1}^{n-1}(x_i - x_{i+1})^2.$$

We may also write $R(\boldsymbol{x})$ in matrix form as $\|\boldsymbol{L}\boldsymbol{x}\|^2$, where $L \in \mathbb{R}^{(n-1)\times n}$ is given as,

$$\mathbf{L} = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}$$

Therefore the resulting RLS problem is,

$$\min \|\boldsymbol{x} - \boldsymbol{b}\|^2 + \lambda \|\boldsymbol{L}\boldsymbol{x}\|^2.$$

Furthermore, the optimal solution is given as,

$$\boldsymbol{x}_{RLS}(\lambda) = (\boldsymbol{I} + \lambda \boldsymbol{L}^T \boldsymbol{L})^{-1} \boldsymbol{b}.$$

4.5 Non-Linear Least Squares

So far the least squares methods we have outlined are for solutio to approximate linear equalities. As we know however, non-linear systems arise ubiquitously in practice.

$$f_i(\boldsymbol{x}) = c_i.$$

In such a case, we want to formulate a non-linear least squares problem of the form,

$$\min \sum_{i=1}^{n} (f_i(\boldsymbol{x}) - c_i)^2.$$

As opposed to linear least squares, there is no general method for solving non-linear least squares problems. The Gauss-Newton Method (which will be disscussed later) can be used to solve problems of such form, however this does not guarantee the optimal solution. Rather, only a solution that is stationary.

4.6 Circle Fitting

Suppose we are given m points $\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_m \in \mathbb{R}^n$. The circel fitting problem seeks to find a circle,

$$C(\boldsymbol{x}, r) = \{\boldsymbol{y} \in \mathbb{R}^n : \|\boldsymbol{y} - \boldsymbol{x}\| = r\},$$

that best fits the m points. Note, we are using the term circle, however our procedure is being constructed in general on $\mathbb{R}^n$. Furthermore, observe that $C(\boldsymbol{x}, r)$ is the boundary of the corresponding ball $B(\boldsymbol{x}, r)$.

The non-linear approximate equations assciated with the problem are given as,

$$\|\boldsymbol{x} - \boldsymbol{a}_i\| \approx r, \quad i = 1, 2, ..., m.$$

We wish to deal with differentiable functions and the norm function is not differentiable. Hence,

$$\|\boldsymbol{x} - \boldsymbol{a}\|^2 \approx r^2.$$

The NLS problem here can be formulated as follows,

$$\min_{\boldsymbol{x} \in \mathbb{R}^n,\, r \in \mathbb{R}} \sum_{i=1}^{m} \left(\|x_i - a_I\|^2 - r^2 \right)^2.$$

The trick here is to realize that this NLS problem can be shown to be equivalent to a linear LS problem. It would follow that the optimal solution can be easily obtained.

Note, the problem is the same as,

$$\min_{\mathbf{x}, r} \left\{ \sum_{i=1}^{m} \left(-2\mathbf{a}_i^T \mathbf{x} + \|\mathbf{x}\|^2 - r^2 + \|\mathbf{a}_i\|^2 \right)^2 : \mathbf{x} \in \mathbb{R}^n, r \in \mathbb{R} \right\}.$$

Where we set the regularization function, $R(\boldsymbol{x}) = \|\boldsymbol{x}\|^2 - r^2$. Then re-write the above problem as,

$$\min_{\mathbf{x}\in\mathbb{R}^n, R\in\mathbb{R}} \left\{ f(\mathbf{x}, R) \equiv \sum_{i=1}^{m} \left(-2\mathbf{a}_i^T\mathbf{x} + R + \|\mathbf{a}_i\|^2 \right)^2 : \|\mathbf{x}\|^2 \geq R \right\}.$$

Notice, that the change of variable introduced a constraint namely, $\|\boldsymbol{x}\|^2 \geq R$. We want to show that this constraint can be dropped and the problem is equivalent to the LS,

$$\min_{\mathbf{x},R} \left\{ \sum_{i=1}^{m} \left(-2\mathbf{a}_i^T\mathbf{x} + R + \|\mathbf{a}_i\|^2 \right)^2 : \mathbf{x} \in \mathbb{R}^n, R \in \mathbb{R} \right\}.$$

5 Gradient Methods

5.1 Descent Direction Methods

We now turn to looking at the general unconstrained minimization problem,

$$\min\{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^n\}.$$

We have already seen in the previous section that one procedure to solve such problems is by computing the stationary points i.e. solutions $\nabla f(\boldsymbol{x}) = \mathbf{0}$ and then finding the stationary point with lowest function value.

In general however this will not work. This is due to several reasons, here are a couple:

(i) In a majority of cases it is very difficult to solve the system of mainly non-linear $\nabla f(\boldsymbol{x}) = \mathbf{0}$.

(ii) We may have infinitely many stationary points and finding the one that corresponds to the minimal value may be a difficult optimization problem in and of itself. This could be even more difficult than the original problem!

It is easy to see that finding exact analytical solutions will be (in general) a very difficult task.

Hence using some iterative numerical technique may be an easy way to find a numerical/computational solution. The iterative algorithm we will consider takes the from,

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + t_k\boldsymbol{d}_k, \quad k = 0, 1, 2, ...$$

Here $\boldsymbol{d}_k \in \mathbb{R}^n$ is the is the direction and t_k is the stepsize. In this method, we restrict ourselves, by convention, to descent directions.

Definition 5.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function over $\mathbb{R}^n$. Let a non-zero vector $\boldsymbol{d} \in \mathbb{R}^n$ be called a descent direction of f at $\boldsymbol{x}$ if the directional derivative is negative. Mathematically,*

$$f'(\boldsymbol{x}; \boldsymbol{d}) = \nabla f(\boldsymbol{x})^T \cdot \boldsymbol{d} < 0.$$

It is straightfoward to see that taking small enough step size, moving in the direction of descent leads to a decrease in objective function value. We may state this more rigourously as a Lemma.

Lemma 5.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function and let $\boldsymbol{x} \in \mathbb{R}^n$. Suppose that $\boldsymbol{d}$ is a descent direction of f at $\boldsymbol{x}$. Then there exists $\epsilon > 0$ such that*

$$f(\boldsymbol{x} + t\boldsymbol{d}) < f(\boldsymbol{x}), \ \forall t \in (0, \epsilon].$$

The proof of Lemma 5.2 follows directly from the definiton of the directional derivative.

Here is a schematic on how the Gradient method works within a computation.

Initialization: Pick $\boldsymbol{x}_0 \in \mathbb{R}^n$ arbitrarily.
General Step: For any $k = 0, 1, 2, ...$
(a) Pick a descent direction $\boldsymbol{d}_k$.
(b) Find a step size t_k such that $f(\boldsymbol{x}_k + t_k \boldsymbol{d}_k) < f(\boldsymbol{x}_k)$.
(c) Set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + t_k \boldsymbol{d}_k$.
(d) It stopping criterion is satisfied, then STOP and the output is $\boldsymbol{x}_{k+1}$.

There are still many details missing, hence the schematic and by extension the method remains conceptual.

We must still have rigourous answers to the questions what is the stopping criteria? What stepsize should be taken? What is the starting point? How does one choose a descent direction?

We begin by analyzing stopping criteria. A popular stoping criterion is given by,

$$\|\nabla f(\boldsymbol{x}_{k+1})\| \leq \epsilon.$$

We assume the stepsize is chosen in such a way that $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$. Notice then we have assumed a descent direction as the function value decreases iteration to iteration.

5.1.1 Step-Size Selection

The process of finding the step size is called **line search**. This can be regarded as a minimization problem in and of itself as we try and minimize the one dimensional function,

$$g(t) = f(\boldsymbol{x}_k + t_k \boldsymbol{d}_k).$$

There are many choices for step size selection. We will describe three major ones now.

- **Constant Step-Size:** $t_k = t$ for all k.
- **Exact Line Search:** t_k is a minimizer of f along the ray $\boldsymbol{x}_k + t_k \boldsymbol{d}_k$.

  $$t_k \in \text{argmin}_{t \geq 0} f(\boldsymbol{x}_k + t_k \boldsymbol{d}_k).$$

- **Backtracking:** This method requires three parameters, namely $s > 0$, $\alpha \in (0, 1)$ and $\beta \in (0, 1)$. The choice of t_k is taken by the following procedure. First, t_k is set to equal the intial s. Then, while,

  $$f(\mathbf{x}_k) - f(\mathbf{x}_k + t_k \mathbf{d}_k) < -\alpha t_k \nabla f(\mathbf{x}_k)^T \mathbf{d}_k$$

  we set $\beta t_k \to t_k$. In words, we want to find the smallest step-size such that the condition above is satisfied.

We can begin to see that choosing a step-size is a non-trivial task and in practice will be more of an art than and exact method while solving gradient descent problems.

For constant step size, we know that its advantage is its simplicity, however the question of how to choose a constant remains ambiguous. It then seems as though line search becomes the perfect solution. However, in practice choosing the minimizer of f along the ray may cause long convergence times and computing inefficiencies. Also it may not always be possible to find an exact minimizer.

In a sense then, backtracking becomes the mathematical compromise between the two other methods in step size selection. It does not perform an exact line search but does generate a "good enough" step size. Rigorously, good-enough here means the condition as outlined above is satisfied.

Lemma 5.3. *Let f be a continuously differentiable function on $\mathbb{R}^n$ and let $\boldsymbol{x} \in \mathbb{R}^n$. Suppose that $\boldsymbol{d} \neq \mathbf{0} \in \mathbb{R}^n$ is a descent direction of f at $\boldsymbol{x}$ amd let $\alpha \in (0,1)$. Then there exists $\epsilon > 0$ such the inequality,*

$$f(\mathbf{x}) - f(\mathbf{x} + t\mathbf{d}) \geq -\alpha t \nabla f(\mathbf{x})^T \mathbf{d},$$

Holds for all $t \in [0, \epsilon]$.

We now show an example to find the analytical formula for the exact line search when the objective function is quadratic.

Example 5.4. *Let $f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} + 2\boldsymbol{b}^T\boldsymbol{x} + c$ where $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and positive definite, $\boldsymbol{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Let $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{d} \in \mathbb{R}^n$ be a descent direction of f at $\boldsymbol{x}$.*

We will find the exact analytical formula for finding the step size as generated by line search. Explicitly, we will find a solution for the problem,

$$\min_{t \geq 0} \{f(\boldsymbol{x} + t\boldsymbol{d})\}.$$

So, we have that,

$$\begin{aligned} g(t) = f(\boldsymbol{x} + t\boldsymbol{d}) &= (\boldsymbol{x} + t\boldsymbol{d})^T \boldsymbol{A} (\boldsymbol{x} + t\boldsymbol{d}) + 2\boldsymbol{b}^T (\boldsymbol{x} + t\boldsymbol{d}) + c \\ &= \left(\boldsymbol{d}^T \boldsymbol{A}\boldsymbol{d}\right) t^2 + 2\left(\boldsymbol{d}^T \boldsymbol{A}\boldsymbol{x} + \boldsymbol{d}^T\boldsymbol{b}\right) t + \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + 2\boldsymbol{b}^T\boldsymbol{x} + c \\ &= \left(\boldsymbol{d}^T \boldsymbol{A}\boldsymbol{d}\right) t^2 + 2\left(\boldsymbol{d}^T \boldsymbol{A}\boldsymbol{x} + \boldsymbol{d}^T\boldsymbol{b}\right) t + f(\boldsymbol{x}). \end{aligned}$$

Note, in the second step we use the fact that $\boldsymbol{A} \succ 0$ and hence symmetric. Now, we know that,

$$\begin{aligned} g'(t) &= 2\left(\boldsymbol{d}^T \boldsymbol{A}\boldsymbol{d}\right) t + 2\left(\boldsymbol{d}^T \boldsymbol{A}\boldsymbol{x} + \boldsymbol{d}^T\boldsymbol{b}\right) \\ \nabla f(\boldsymbol{x}) &= 2\left(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}\right). \end{aligned}$$

It then must be the case that $g'(t) = 0$ if and only if,

$$t = \bar{t} \equiv -\frac{\boldsymbol{d}^T \nabla f(\boldsymbol{x})}{2\boldsymbol{d}^T \boldsymbol{A}\boldsymbol{d}}.$$

Since $\boldsymbol{d}$ is a descent direction of f at $\boldsymbol{x}$, it follows that $\boldsymbol{d}^T \nabla f(\boldsymbol{x}) < 0$ and hence $\bar{t} > 0$. This implies that the step size dictated by the exact line search is,

$$\bar{t} = -\frac{\boldsymbol{d}^T \nabla f(\boldsymbol{x})}{2\boldsymbol{d}^T \boldsymbol{A}\boldsymbol{d}}.$$

Note, we have implicitly used the fact that the second dervative with respect to $g(t)$ is always positive.

5.2 Gradient Descent Method

We now turn towards outlining the gradient descent method or just gradient descent. Building from the theory thus far, the gradient descent method takes the descent direction to be the negative gradient of the objective function i.e $\boldsymbol{d}_k = -\nabla f(\boldsymbol{x}_k)$. By definition, this will give us the direction of steepest descent.

As done eariler, we outline the schematic of the gradient descent method program.

Input: $\epsilon > 0$, tolerance parameter.

Initialization: Pick $\boldsymbol{x}_0 \in \mathbb{R}^n$ arbitrarily.

General Step: For $k : 0, 1, 2, ...$
(a) Pick step size via line search on the function $g(t) = f(\boldsymbol{x}_k - t\nabla f(\boldsymbol{x}_k))$.
(b) Set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + t_k \nabla f(\boldsymbol{x}_k)$.
(c) If $\|\nabla f(\boldsymbol{x}_{k+1})\| < \epsilon$, STOP and $\boldsymbol{x}_{k+1}$ is the output.

We now turn to looking at an example of an exact line search within the gradient descent method as implemented on MATLAB.

Example 5.5. *Consider the two dimensional minimization problem,*

$$\min_{x,y} x^2 + 2y^2.$$

It is easy to see that there is only one minimizer and it is global namely $(0,0)$*, where the function value is* 0*. To invoke the gradient method with the step size chosen by exact line search we write a MATLAB script as provided below.*

The program finds the exact step size via exact line search to problems of the form,

$$\min_{\boldsymbol{x}\in R^n} \left\{ \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} + 2\boldsymbol{b}^T \boldsymbol{x} \right\}.$$

Where $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *is* $\boldsymbol{A} \succ 0$ *and* $\boldsymbol{b} \in \mathbb{R}^n$*. By Example 5.5, we know the exact line search formula for quadratic minimization probelm. Hence, in this case we know the step size is given by the* k*th iteration,*

$$t_k = \frac{\|\nabla f(\boldsymbol{x}_k)\|^2}{2\nabla f(\boldsymbol{x}_k)^T \boldsymbol{A} \nabla f(\boldsymbol{x}_k)}.$$

Keep in mind this is a gradient method and hence the direction of descent is taken to be the negative gradient of the objective function.

We now present the example of applying the gradient method via backtracking in order to select the step size.

5.3 The Condition Number

Consider the quadratic minimization problem,

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \left\{ f(\boldsymbol{x}) \equiv \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} \right\}.$$

Such that $\boldsymbol{A} \succ 0$. The optimal solution is obviously $\boldsymbol{x}^* = \boldsymbol{0}$. We know that the gradient method with exact line search takes the form,

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t_k \boldsymbol{d}_k,$$

where $\boldsymbol{d}_k = \nabla f(\boldsymbol{x}_k) = 2\boldsymbol{A}\boldsymbol{x}_k$. Moreover, we know the step size is given by,

$$t_k = \frac{\boldsymbol{d}_k^T \boldsymbol{d}_k}{2\boldsymbol{d}_k^T \boldsymbol{A}\boldsymbol{d}_k}.$$

Then we have that

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) &= \boldsymbol{x}_{k+1}^T \boldsymbol{A}\boldsymbol{x}_{k+1} \\ &= (\boldsymbol{x}_k - t_k\boldsymbol{d}_k)^T \boldsymbol{A}(\boldsymbol{x}_k - t_k\boldsymbol{d}_k) \\ &= \boldsymbol{x}_k^T \boldsymbol{A}\boldsymbol{x}_k - 2t_k \boldsymbol{d}_k^T \boldsymbol{A}\boldsymbol{x}_k + t_k^2 \boldsymbol{d}_k^T \boldsymbol{A}\boldsymbol{d}_k \\ &= \boldsymbol{x}_k^T \boldsymbol{A}\boldsymbol{x}_k - t_k \boldsymbol{d}_k^T \mathbf{d}_k + t_k^2 \boldsymbol{d}_k^T \boldsymbol{A}\boldsymbol{d}_k \end{aligned}.$$

Now, substituting the expression for t_k into the last step,

$$\begin{aligned} f(\mathbf{x}_{k+1}) = \mathbf{x}_k^T \mathbf{A}\mathbf{x}_k - \frac{1}{4}\frac{\left(\mathbf{d}_k^T \mathbf{d}_k\right)^2}{\mathbf{d}_k^T \mathbf{A}\mathbf{d}_k} &= \mathbf{x}_k^T \mathbf{A}\mathbf{x}_k \left(1 - \frac{1}{4}\frac{\left(\mathbf{d}_k^T \mathbf{d}_k\right)^2}{\left(\mathbf{d}_k^T \mathbf{A}\mathbf{d}_k\right)\left(\mathbf{x}_k^T \mathbf{A}\mathbf{A}^{-1}\mathbf{A}\mathbf{x}_k\right)}\right) \\ &= \left(1 - \frac{\left(\mathbf{d}_k^T \mathbf{d}_k\right)^2}{\left(\mathbf{d}_k^T \mathbf{A}\mathbf{d}_k\right)\left(\mathbf{d}_k^T \mathbf{A}^{-1}\mathbf{d}_k\right)}\right) f(\mathbf{x}_k) \end{aligned}.$$

We now turn to a well known result from matrix theory called Kantorovich's inequality.

Lemma 5.6. *Let $\boldsymbol{A}$ be a real, square, positive definite matrix. Then for any $\boldsymbol{x} \neq \boldsymbol{0} \in \mathbb{R}^n$, the inequality,*

$$\frac{\left(\boldsymbol{x}^T\boldsymbol{x}\right)^2}{\left(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}\right)\left(\boldsymbol{x}^T\boldsymbol{A}^{-1}\boldsymbol{x}\right)} \geq \frac{4\lambda_{\max}(\boldsymbol{A})\lambda_{\min}(\boldsymbol{A})}{\left(\lambda_{\max}(\boldsymbol{A}) + \lambda_{\min}(\boldsymbol{A})\right)^2},$$

holds.

We want to analyse the convergence rate of the gradient method on the quadratic minimization problem. By using the Kantorovich inequality we have that,

$$f(\boldsymbol{x}_{k+1}) \leq \left(1 - \frac{4Mm}{(M+m)^2}\right) f(\boldsymbol{x}_k) = \left(\frac{M-m}{M+m}\right)^2 f(\boldsymbol{x}_k).$$

Where $M = \lambda_{max}(\boldsymbol{A})$ and $m = \lambda_{min}(\boldsymbol{A})$. We may summerize the ideas here in the following lemma.

Lemma 5.7. *Let $\{\boldsymbol{x}_k\}_{k\geq 0}$ be the sequence generated by the gradient descent method via exact line search for solving the quadratic minimization probelm. Then for any k,*

$$f(\boldsymbol{x}_{k+1}) \leq \left(\frac{M-m}{M+m}\right)^2 f(\boldsymbol{x}_k).$$

This inequality implies that, $f(\boldsymbol{x}_k) \leq c^k f(\boldsymbol{x}_0)$, where $c = \left(\frac{M-m}{M+m}\right)^2$.

We have that the sequence of function values is bounded above by a deacreasing geometric sequence. In this case we say that the sequence of a function values converges at a linear rate to the optimal value.

There is great dependence on c, as such we may deduce convergence rates by analysing c. **As c becomes larger, convergence speed becomes slower.** We may also write,

$$c = \left(\frac{x-1}{x+1}\right)^2.$$

Where $x = \frac{M}{m}$ i.e. the ratio between the largest eigenvalue of $\boldsymbol{A}$ and the smallest eigenvalue of $\boldsymbol{A}$.

Definition 5.8. *Let $\boldsymbol{A}$ be an $n \times n$ positive definite matrix. Then the **condition number** of $\boldsymbol{A}$ is defined by,*

$$\chi(\boldsymbol{A}) = \frac{\lambda_{\max}(\boldsymbol{A})}{\lambda_{\min}(\boldsymbol{A})}.$$

We may in general define the condition number for any square matrix, however we will restrict ourselves to the postive definite case for our purposes with respect to optimization.

One will notice that when gradient method take a relatively small number of iterations to converge to the optimal solution, the condition number is small.

Matrices with a small condition number are said to be **well-conditioned**, whereas matrices with large condition number are said to be **ill-conditioned**.

We have restriced ourselves to a class of problems where the objective function is quadratic, hence the Hessian is constant and therefore we gain well defined condition numbers. However, the same analysis will work for non-quadratic objective functions as well.

Convergence Analysis of the Gradient Method

We now turn to providing a more rigorous treatment regarding convergence of the gradient method.

Lipschitz Property of the Gradient

Consider the unconstrained optimization problem,

$$\min\{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^n\}.$$

We assume that the objective function is continuously differentiable and that the gradient is Lipschitz continuous i.e.

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \text{ for any } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

Note that if $\nabla f(\boldsymbol{x})$ is Lipschitz consitinuous with constant L, then it is also Lipschitz continuous with constant $\tilde{L} \geq L$.

The class of functions with Lipschitz gradient with constant L is denoted $C_L^{1,1}(\mathbb{R}^n)$ or just $C_L^{1,1}$. Furthermore, linear functions and quadratic functions fall in the class of $C^{1,1}$.

Theorem 5.9. *Let f be a twice continuously differentiable function over $\mathbb{R}^n$. Then the following two claims are equivalent:*

(a) $f \in C_L^{1,1}(\mathbb{R}^n)$.

(b) $\|\nabla^2 f(\boldsymbol{x})\| \leq L$ *for any* $\boldsymbol{x} \in \mathbb{R}^n$.

In words, Theorem 5.9 says that is a twice continuously differentiable function has Lipschitz gradient it is equivalent to saying that the norm on its Hessian is bounded above by the Lipschitz constant.

Also note that the norm of the Hessian is an induced matrix norm and will generally (for our purposes) be the spectral norm.

The Descent Lemma

An important result for all $C^{1,1}$ functions is that they can be bounded above by qudratic functions over the entire space. This result is known as the **Descent Lemma** and is of paramount importance we doing analysis regarding convergence of gradient based methods.

Lemma 5.10. *Let $f \in C_L^{1,1}(\mathbb{R}^n)$. Then for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,*

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

Proof. □

Convex Sets

We now begin the theoretical discussion of convex sets and convex analysis. The theoretical foundations we build will then the applied within the context of a large sub-filed of optimization called convex optimization.

Definition 5.11. ***Convex Sets:*** *Let $S \subseteq \mathbb{R}^n$, S is said to be convex if for any $\boldsymbol{x}, \boldsymbol{y} \in S$,*

$$\lambda\boldsymbol{x} + (1 - \lambda)\boldsymbol{y} \in S,$$

for any $\lambda \in [0, 1]$.

Geometrially, this definition states that a set is convex only when for every pair of points within the set, those two points can be connect by a line segment and the enitre line segment stays within the set.

As a trivial example, a line segment is a convex set. Consider the following lemma on more important sets that turn out to be convex.

Lemma 5.12. *Let $\boldsymbol{a} \in \mathbb{R}^N\{\boldsymbol{0}\}$ and $b \in \mathbb{R}^n$. Then the following sets are convex,*

(a) The hyper-plane $H = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{a}^T\boldsymbol{x} = \boldsymbol{b}\}$.

(b) The half-space $H^- = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{a}^T\boldsymbol{x} \leq \boldsymbol{b}\}$.

(c) The open-half-space $\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{a}^T\boldsymbol{x} < \boldsymbol{b}\}$.

Proof. Let us prove the convexity of the half-space, the proof for the other two can be extended trivially. Following from the definition of convexity, let $\boldsymbol{x}, \boldsymbol{y} \in H^-$ and $\lambda \in [0, 1]$. We want to show that $\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y} \in H^-$. Take $\boldsymbol{z} \in H^-$ such that,

$$\boldsymbol{a}^T\boldsymbol{z} = \boldsymbol{a}^T(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) = \lambda\boldsymbol{a}^T\boldsymbol{x} + (1-\lambda)\boldsymbol{a}^T\boldsymbol{y} \leq \lambda b + (1-\lambda)b.$$

Where the inequality follows from the definition of the half space i.e. $\boldsymbol{a}^T\boldsymbol{x} \leq b$ and $\boldsymbol{a}^T\boldsymbol{y} \leq b$. Hence the claim follows. □

To establish convexity of more topological sets, we can also show that open and closed balls are convex.

Lemma 5.13. *Let $\boldsymbol{c} \in \mathbb{R}^n$ and $r > 0$ and let $\|\cdot\|$ be some norm on $\mathbb{R}^n$. Then, the closed and open balls,*

$$B(c, r) = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{c} - \boldsymbol{x}\| < r\},\ B[c, r] = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{c} - \boldsymbol{x}\| \leq r\},$$

are convex.

Proof. We prove the convexity of the closed ball, the proof can be trivially extened to open balls. Let $\boldsymbol{x}, \boldsymbol{y} \in B[\boldsymbol{c}, r]$ and $\lambda \in [0, 1]$, it follows that,

$$\|\boldsymbol{c} - \boldsymbol{x}\| \leq r,\ \|\boldsymbol{c} - \boldsymbol{y}\| \leq r.$$

Now let $\boldsymbol{z} = \lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}$, we want to show that $\boldsymbol{z} \in B[\boldsymbol{c}, r]$. Take,

$$\begin{aligned}
\|\boldsymbol{z} - \boldsymbol{c}\| &= \|\lambda\boldsymbol{x} - (1-\lambda)\boldsymbol{y} - \boldsymbol{c}\| \\
&= \|\lambda(\boldsymbol{x} - \boldsymbol{c}) + (1-\lambda)(\boldsymbol{y} - \boldsymbol{c})\| \\
&\leq \|\lambda(\boldsymbol{x} - \boldsymbol{c})\| + \|(1-\lambda)(\boldsymbol{y} - \boldsymbol{c})\| \\
&= \lambda\|\boldsymbol{x} - \boldsymbol{c}\| + (1-\lambda)\|\boldsymbol{y} - \boldsymbol{c}\| \\
&\leq \lambda r + (1-\lambda)r = r.
\end{aligned}$$

Note, the first inequality comes from the triangel inequality. The claim follows. □

Note, that these results will work for all norms on $\mathbb{R}^n$.

Algebraic Operations on Convex Sets

An important property of convexity is that it is preserved under intersections of sets.

Lemma 5.14. *Let $C_i \in \mathbb{R}^n$ where i is in some index set I. If each C_i is convex then,*

$$\bigcap_i C_i,$$

is also convex.

Proof. Let $\boldsymbol{x}, \boldsymbol{y} \in \bigcap_i C_i$ and let $\lambda \in [0,1]$ and assume that C_i is convex for any i. It then follows from the definition of convexity that $\lambda\boldsymbol{x} + (1+\lambda)\boldsymbol{y} \in C_i$ for any i. By the definition of intersection of sets, it must be the case that,

$$\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in \bigcap_i C_i.$$

The claim follows. □

In a similar manner we can prove that convexity is preserved under addition, cartersian product, linear mappings and inverse linear mappings. These results also show that convexity is preserved under translations, rotations and other symmetries.

Note, that convexity is preserved under set intersections however this is not the case for set unions. For example, take two distinct closed balls in $\mathbb{R}^2$. The union of these two sets is clearly not convex as there is not line segmet that connect one point from the first ball to the another point in the second ball while remaining within the set union.

Definition 5.15. ***Convex Combinations*** *Given k vectors $\boldsymbol{x}_1, ..., \boldsymbol{x}_k$, the convex combination of these vectors is,*

$$\sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i,$$

such that $\sum_{i=1}^{k} \lambda_i = 1$ and each λ_i is non-negative.

We can see how then how the definition of the convex set originates. Moreover, we may re-define a convex set to be one, where the convex combinations of vectors within some set remain within the set.

In fact for any m vectors in a convex set, we can prove via induction that the convex combination of these m vectors remains within the set.

Theorem 5.16. *(**Countable Convex Combinations**) Let C be a convex set and let $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_m \in C$. Then for any $\lambda \in \Delta_m$,*

$$\sum_{i=1}^{m} \lambda_i \boldsymbol{x}_i \in C.$$

Proof. We will show this by induction on m. For the base case $m = 1$, the result is trivial as $\boldsymbol{x}_1 \in C$.

The induction hypothesis is that $\sum_{i=1}^{m} \lambda_i \boldsymbol{x}_i \in C$.

We now want to show that $\boldsymbol{z} := \sum_{i=1}^{m+1} \lambda_i \boldsymbol{x}_i \in C$. If $\lambda_{m+1} = 1$ we have that $\boldsymbol{z} = \boldsymbol{x}_{m+1} \in C$ and the claim follows. If $\lambda_{m+1} < 1$, then,

$$\begin{aligned}\boldsymbol{z} &= \sum_{i=1}^{m+1} \lambda_i \boldsymbol{x}_i \\ &= \sum_{i=1}^{m} \lambda_i \boldsymbol{x}_i + \lambda_{m+1} \boldsymbol{x}_{m+1}.\end{aligned}$$

If we now multiply the first term of the second step by one in the form $\frac{1-\lambda_{m+1}}{1-\lambda_{m+1}} = 1$, we may rewrite $\boldsymbol{z}$ as,

$$\boldsymbol{z} = (1 - \lambda_{m+1}) \sum_{i=1}^{m} \frac{\lambda_i}{1 - \lambda_{m+1}} \boldsymbol{x}_i + \lambda_{m+1} \boldsymbol{x}_{m+1}.$$

Let,

$$\boldsymbol{v} := \sum_{i=1}^{m} \frac{\lambda_i}{1 - \lambda_{m+1}} \boldsymbol{x}_i.$$

Then, $\boldsymbol{v}$ is a convex combination of the first m terms and by the induction hypothesis $\boldsymbol{v} \in C$. Furthermore, by Definition **??**,

$$\boldsymbol{z} = (1 - \lambda_{m+1})\boldsymbol{v} + \lambda_{m+1} \boldsymbol{x}_{m+1} \in C.$$

The induction is complete and the claim is established. □

Definition 5.17. ***Convex Hulls*** *Let $S \subseteq \mathbb{R}^n$, then the convex hull of S, denoted* $\text{conv}(S)$*, is the set comprising all convex combinations of S,*

$$\text{conv}(S) \equiv \left\{ \sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i : \boldsymbol{x}_1, ..., \boldsymbol{x}_k \in S,\ \lambda \in \Delta_k,\ k \in \mathbb{N} \right\}.$$

We can say that $\text{conv}(S)$ is the smallest convex set containing S. This fact requires a proof.

Lemma 5.18. *Let $S \subseteq \mathbb{R}^n$, if $S \subseteq T$ for some convex set T, then* $\text{conv}(S) \subseteq T$.

Proof. Suppose that $S \subseteq T$ and take $\boldsymbol{z} \in \text{conv}(S)$. By the definition of the convex hull, $\boldsymbol{z}$ can be represented as the convex combination,

$$\boldsymbol{z} = \sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i,$$

of vectors $\boldsymbol{x}_i \in \text{conv}(S)$. But then, by the convexity of T and the fact that $\boldsymbol{x}_i \in T$, it must be the case that $\boldsymbol{z} \in T$. The claim follows. □

We now turn to presenting a central idea from Caratheodory.

Theorem 5.19. ***Caratheodory's Theorem*** *Let $S \subset \mathbb{R}^n$ and let $\boldsymbol{x} \in \text{conv}(S)$. Then there exists $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ such that $\boldsymbol{x} \in \text{conv}(\{\boldsymbol{x}_1, ..., \boldsymbol{x}_{n+1}\})$. That is there exists $\lambda \in \Delta_{n+1}$ such that,*

$$\boldsymbol{x} = \sum_{i=1}^{n+1} \lambda_i \boldsymbol{x}_i.$$

In words, Caratheodory's Theorem says that any element of the convex hull of a subset $S \subseteq \mathbb{R}^n$ can be represented as a convex combination of no more than $n + 1$ vectors of S.

Proof. $\boldsymbol{x} \in \text{conv}(S)$. By Defintion of the convex hull, we know that there exists $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k \in S$ and $\lambda \in \Delta_k$ such that,

$$\boldsymbol{x} = \sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i.$$

We may assume without loss of generality that $\lambda_i > 0$ otherwise the term would simply be omitted from the convex combination. Also, keep in mind $\sum_{i=1}^{k} \lambda_i = 1$. If $k \leq n+1$ the claim is proven.

Assume $k > n+1$, then the vectors, $\boldsymbol{x}_2 - \boldsymbol{x}_1, \boldsymbol{x}_3 - \boldsymbol{x}_1, ..., \boldsymbol{x}_k - \boldsymbol{x}_1$, which are more than n vectors in $\mathbb{R}^n$ are necessarily linearly dependent. So, there exists $\mu_2, \mu_3, ..., \mu_k$ which are not all zeros such that,

$$\sum_{i=2}^{k} \mu_i (\boldsymbol{x}_i - \boldsymbol{x}_1) = 0.$$

If we define,

$$\mu_1 := -\sum_{i=2}^{k} \mu_i,$$

We obtain that,

$$\sum_{i=1}^{k} \mu_i \boldsymbol{x}_i = 0.$$

Where not all coefficents μ_i are zero and they satisfy $\sum_{i=1}^{k} \mu_i = 0$. □

5.4 Convex Cones

A set S is called a cone if it satisfies the following property, for any $\boldsymbol{x} \in S$ and $\lambda \geq 0$ the inclusion $\lambda \boldsymbol{x} \in S$ is satisfied.

We now propose a simple and elegant characterization of convex cones.

Lemma 5.20. *A set is a convex cone if and only if the following properties hold,*

(a) $\boldsymbol{x}, \boldsymbol{y} \in S \implies \boldsymbol{x} + \boldsymbol{y} \in S$.

(b) $\boldsymbol{x} \in S \ \lambda \geq 0 \implies \lambda \boldsymbol{x} \in S$.

Proof. (convex cone $\implies$ (a),(b)). Assume S is a convex cone, property (b) follows from the definition. To prove (a), assumed $\boldsymbol{x}, \boldsymbol{y} \in S$. By the convexity of S we have that $\frac{1}{2}(\boldsymbol{x} + \boldsymbol{y}) \in S$. Then, since S is a cone, we have that $\boldsymbol{x} + \boldsymbol{y} = 2 \cdot \frac{1}{2}(\boldsymbol{x} + \boldsymbol{y}) \in S$.

((a),(b) $\implies$ convex cone). A set S satisfies properties (a) and (b). By property (b), S is a cone. To show convexity, let $\boldsymbol{x}, \boldsymbol{y} \in S$ and let $\lambda \in [0, 1]$. Then, $\lambda \boldsymbol{x}, (1-\lambda)\boldsymbol{y} \in S$ by property (b). Thus, by property (a), $\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in S$ thererby establishing convexity of S. □

Definition 5.21. ***(Conic Combinations)*** *Given k points such that $\boldsymbol{x}_1, ..., \boldsymbol{x}_k \in \mathbb{R}^n$, a conic combination of these k vectors, is another vector of the form,*

$$\sum_{i=1}^{k} \lambda_1 \boldsymbol{x}_i,$$

such that each $\lambda \in \mathbb{R}^k_+$.

Now we have the elementary albeit important result.

Lemma 5.22. *Let C be a convex cone, and let $\boldsymbol{x}_1, ..., \boldsymbol{x}_k \in C$ and $\lambda_1, .., \lambda_k \geq 0$. Then the convex combination,*

$$\sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i \in C.$$

In words, the conic combination remains within the convex cone.

Proof. By the lemma on convex cones, it must be the case that,

$$\lambda_1 \boldsymbol{x}_1 + \lambda_2 \boldsymbol{x}_2 + ... + \lambda_k \boldsymbol{x}_k = \sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i \in C.$$

Where, $\lambda_k \in \mathbb{R}^k_+$. The claim follows as this is the conic combination of the k vectors in C. □

Just as for convex sets and convex hulls, we may naturally propose a definition of the conic hull.

Definition 5.23. ***(Conic Hull)*** *Let $S \subset \mathbb{R}^n$. The conic hull of S, denoted* $\text{cone}(S)$*, is the set comprising all conic combinations of vectors within S,*

$$\text{cone}(S) := \left\{ \sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i : \boldsymbol{x}_i \in S,\ \lambda \in \mathbb{R}^k_+,\ k \in \mathbb{N} \right\}.$$

As we stated for convex hulls, the conic hull of S is the smallest convex cone containing S.

Now the natural question is, whether we can establish an analog of Caratheodory's theorem for convex cones. Interestingly enough, we may establish an even stronger result.

Theorem 5.24. ***(Conic Representation Theorem)*** *Let $S \subset \mathbb{R}^n$ and let $\boldsymbol{x} \in \text{cone}(S)$. Then there exists k linearly independent vectors, $\boldsymbol{x}_1, ..., \boldsymbol{x}_k \in S$ such that $\boldsymbol{x} \in \text{cone}(\{\boldsymbol{x}_1, ..., \boldsymbol{x}_k\})$. That is, there exists $\lambda \in \mathbb{R}^k_+$ such that,*

$$\boldsymbol{x} = \sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i.$$

In addition, $k \leq n$.

Definition 5.25. ***(Basic Feasible Solutions)*** *Let $P = \{\boldsymbol{x} \in \mathbb{R}^n : A\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \geq 0\}$ where $A \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$. Suppose that the rows of A are linearly independent. Then $\tilde{\boldsymbol{x}}$ is a basic feasible solution (bfs) of P if the the columns of A corresponding to the indices of the positive values of $\tilde{\boldsymbol{x}}$ are linearly independent.*

Definition 5.26. ***(Extreme Point)*** *Let $S \subseteq \mathbb{R}^n$. A point $\boldsymbol{x} \in S$ is called an extreme point of S if there do not exist $\boldsymbol{x}_1, \boldsymbol{x}_2 \in S$ where $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$ and $\lambda \in (0, 1)$ such that $\boldsymbol{x} = \lambda \boldsymbol{x}_1 + (1 - \lambda)\boldsymbol{x}_2$.*

That is, an extreme point of a set is a point that cannot be expressed as a trivial convex combination of any other distinct points in the set. Furthermore, the set of all extreme points of S is denoted $\text{ext}(S)$. Geometrically, the set of extreme points of a convex polytope consists of all its verticies.

Convex Functions

We now turn towards the ideas and concepts related to convex functions. These play an essential part in the theory of convex optimization and more generally in non-linear optimization and obviously convex analysis.

Definition 5.27. *(**Convex Function**) A function $f : C \subseteq \mathbb{R}^n \to R$ where C is convex, is called convex if,*

$$f(\lambda \boldsymbol{x} + (1 - \lambda \boldsymbol{y})) \leq \lambda f(\boldsymbol{x}) + (1 - \lambda) f(\boldsymbol{y}),$$

$\forall\, \boldsymbol{x}, \boldsymbol{y} \in C$ *and* $\lambda \in [0, 1]$.

The definition is also known as the *fundemental inequality*. Furthermore, if we do not allow equality in the fundemental inequality, then f is known as a *strictly convex* function.

Another important and related concept is that of *concavity*. We say that a function is concave if $-f$ is convex. More formally,

Definition 5.28. *(**Concave Function**) A function $f : C \subseteq \mathbb{R}^n \to R$ where C is convex, is called concave if,*

$$f(\lambda \boldsymbol{x} + (1 - \lambda \boldsymbol{y})) \geq \lambda f(\boldsymbol{x}) + (1 - \lambda) f(\boldsymbol{y}),$$

$\forall\, \boldsymbol{x}, \boldsymbol{y} \in C$ *and* $\lambda \in [0, 1]$.

Theorem 5.29. *(**Jensen's Inequality**) Let $f : C \to R$ be a convex function, where $C \subseteq \mathbb{R}^n$ is a convex set. Then for any $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k \in C$ and $\lambda \in \Delta_k$ the following inequality holds,*

$$f\left(\sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i\right) \leq \sum_{i=1}^{k} \lambda_i f(\boldsymbol{x}_i).$$

Proof. We will prove the claim by induction on k. For the base case, $k = 1$, the statement is trivial and follows from the fefinition of the convex function.

The induction hypothesis is that,

$$f\left(\sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i\right) \leq \sum_{i=1}^{k} \lambda_i f(\boldsymbol{x}_i). \tag{2}$$

We want to show that the result will hold for $k + 1$. Suppose that $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_{k+1} \in C$ and $\lambda \in \Delta_{k+1}$. Let,

$$\boldsymbol{z} := \sum_{i=1}^{k+1} \lambda_i \boldsymbol{x}_i.$$

We want to establish that,

$$f(\boldsymbol{z}) \leq \sum_{i=1}^{k} \lambda_i f(\boldsymbol{x}_i).$$

If $\lambda_{k+1} = 1$, we have that $\boldsymbol{z} = \boldsymbol{x}_{k+1}$ and the claim follows straightforwardly. If $\lambda_{k+1} < 1$, then

$$\begin{aligned}
f(\boldsymbol{z}) &= f\left(\sum_{i=1}^{k+1} \lambda_i \boldsymbol{x}_i\right) \\
&= f\left(\sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i + \lambda_{k+1} \boldsymbol{x}_{k+1}\right) \\
&= f\left((1 - \lambda_{k+1}) \sum_{i=1}^{k} \frac{\lambda_i}{1 - \lambda_{k+1}} \boldsymbol{x}_i + \lambda_{k+1} \boldsymbol{x}_{k+1}\right) \\
&\leq (1 - \lambda_{k+1}) f\left(\sum_{i=1}^{k} \frac{\lambda_i}{1 - \lambda_{k+1}} \boldsymbol{x}_i\right) + \lambda_{k+1} f(\boldsymbol{x}_{k+1}).
\end{aligned}$$

Let,

$$\boldsymbol{v} := \sum_{i=1}^{k} \frac{\lambda_i}{1-\lambda_{k+1}} \boldsymbol{x}_i.$$

It follows that $\boldsymbol{v}$ is a convex combination of k points in the convex set C. By the induction hypothesis, $f(\boldsymbol{v}) \leq \sum_{i=1}^{k} \frac{\lambda_i}{1-\lambda_{k+1}} f(\boldsymbol{x}_i)$. The induction is thereby complete and the claim follows. □

First Order Characterization of Convex Functions

Convex functions are not necessarily differentiable. When they are however, we can replace terms in Jensen's inequality with the gradient of the convex function. We will see through the *Gradient Inequality*, tangent hyperplanes of convex functions i.e. gradients of said functions are underestimates.

As one can see, this will play a vital role in optimization problems and finding minima of functions that are convex.

Theorem 5.30. ***(The Gradient Inequality)*** *Let $f : C \subseteq \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function defined over a convex set C. Then f is convex if and only if,*

$$f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}) \leq f(\boldsymbol{y}),$$

for any $\boldsymbol{x}, \boldsymbol{y} \in C$.

Proof. Suppose f is a convex function. If $\boldsymbol{x} = \boldsymbol{y}$, the claim follows trivially. Assume $\boldsymbol{x} \neq \boldsymbol{y}$. By the definition of f being convex, we have that,

$$f(\lambda \boldsymbol{x} + (1 - \lambda \boldsymbol{y})) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$$
$$\frac{f(\boldsymbol{x} + \lambda(\boldsymbol{y} - \boldsymbol{x}) - f(\boldsymbol{x})}{\lambda} \leq f(\boldsymbol{y}) - f(\boldsymbol{x}).$$

Notice, letting $\lambda \to 0^+$, the quotient on the left hand side will converge to the directional derivative of f at $\boldsymbol{x}$ in the $\boldsymbol{y} - \boldsymbol{x}$ direction.

$$f'(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) \leq f(\boldsymbol{y}) - f(\boldsymbol{x}).$$

By assumption, f is continuously differentiable, so it follows that,

$$f'(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) \leq \nabla f(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}).$$

To prove the reverse, assume the gradient inequality holds. Let $\boldsymbol{z}, \boldsymbol{v} \in C$ and let $\lambda \in (0, 1)$. We want to show that,

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{v}) \leq \lambda f(\boldsymbol{z}) + (1-\lambda) f(\boldsymbol{v}).$$

Let $\boldsymbol{u} = \lambda \boldsymbol{z} + (1-\lambda)\boldsymbol{v} \in C$. Then,

$$\boldsymbol{z} - \boldsymbol{u} = \frac{\boldsymbol{u} - (1-\lambda)\boldsymbol{v}}{\lambda} - \boldsymbol{u} = -\frac{1-\lambda}{\lambda}(\boldsymbol{v} - \boldsymbol{u}).$$

We now imploy the Gradient Inequality on the pairs $\{\boldsymbol{z}, \boldsymbol{u}\}$ and $\{\boldsymbol{v}, \boldsymbol{u}\}$.

$$f(\boldsymbol{u}) + \nabla f(\boldsymbol{u})^T (\boldsymbol{z} - \boldsymbol{u}) \leq f(\boldsymbol{z})$$
$$f(\boldsymbol{u}) - \frac{\lambda}{1-\lambda} \nabla f(\boldsymbol{u})^T (\boldsymbol{z} - \boldsymbol{u}) \leq f(\boldsymbol{v}),$$

Multiplying the first inequality by $\frac{\lambda}{1-\lambda}$ and adding to the second, we get,

$$\frac{1}{1-\lambda} f(\boldsymbol{u}) \leq \frac{\lambda}{1-\lambda} f(\boldsymbol{z}) + f(\boldsymbol{v}).$$

Which can the be re-arranged,

$$f(\boldsymbol{u}) \leq \lambda f(\boldsymbol{z}) + (1-\lambda) f(\boldsymbol{v}).$$

Hence, the claim follows, f is convex function. □

Proposition 5.31. *(**Sufficency of Stationarity under Convexity**) Let f be a continuously differentiable convex function over the convex set $C \subseteq \mathbb{R}^n$. Suppose the $\nabla f(\boldsymbol{x}^*) = 0$ then $\boldsymbol{x}^*$ is a global minimizer of f over C.*

Proof. Let $\boldsymbol{z} \in C$, applying the gradient inequality and using the fact that $\nabla f(\boldsymbol{x}^*) = 0$, we get that,

$$f(\boldsymbol{x}^*) \leq f(\boldsymbol{z}),\ \forall\, \boldsymbol{z} \in C.$$

So, $\boldsymbol{x}^*$ is a global minimizer of f over C. □

Note that proposition becomes a neccesary condition when $C = \mathbb{R}^n$.

Theorem 5.32. *(**Necessity and Sufficency of Stationarity under Convexity**) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable convex function. Then $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ if and only if $\boldsymbol{x}^*$ is a global minimum point of f over $\mathbb{R}^n$.*

Now we turn to establishing strict convexity of quadratic convex functions.

Theorem 5.33. *(**Convexity and strict convexity of quadratic functions with positive semidefinite matrices**) Let $f : \mathbb{R}^n \to \mathbb{R}$ be the quadratic function of the form $\boldsymbol{x}^T A\boldsymbol{x} + 2\boldsymbol{b}^T\boldsymbol{x} + c$ where $A \in \mathbb{R}^{n\times n}$ is a symmetric matrix, $\boldsymbol{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Then, f is convex if $A \succcurlyeq 0$ and strictly convex if $A \succ 0$.*

Proof. We start by proving convexity. We know from the theorem on the gradient inequality, the convexity of f is equivalent to the validity of the gradient inequality. So, we want to show that,

$$\boldsymbol{x}^T A\boldsymbol{x} + 2\boldsymbol{b}^T\boldsymbol{x} + c + 2(A\boldsymbol{x} + \boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{x}) \leq \boldsymbol{y}^T A\boldsymbol{y} + 2\boldsymbol{b}^T\boldsymbol{y} + c.$$

Where the inequality holds for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. With some rearangement we may write the inequality as,

$$(\boldsymbol{y} - \boldsymbol{x})^T A(\boldsymbol{y} - \boldsymbol{x}) \geq 0.$$

Let $\boldsymbol{d} = \boldsymbol{y} - \boldsymbol{x}$, the we have that $\boldsymbol{d}^T A\boldsymbol{d} \geq 0$ for any $\boldsymbol{d} \in R^n$, therefore we have that $A \succcurlyeq 0$.

We now want to show strict convexity. This will amount to using the strict gradient inequality,

$$\boldsymbol{x}^T A\boldsymbol{x} + 2\boldsymbol{b}^T\boldsymbol{x} + c + 2(A\boldsymbol{x} + \boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{x}) < \boldsymbol{y}^T A\boldsymbol{y} + 2\boldsymbol{b}^T\boldsymbol{y} + c.$$

By the analogous argument we get that $\boldsymbol{d}^T A\boldsymbol{d} > 0$, which is equivalent to saying that $A \succ 0$ and the claim follows. □

Another type of first order characterization is the monotonicity of the gradient. In the one dimensional case, this characterization states that the derivative is non-decreasing. However, we will need to generalize for n-dimensions.

Theorem 5.34. *(**Monotonicity of the Gradient**) Suppose that $f : C \subset \mathbb{R}^n \to R$ is a continuously differentiable function and C is a convex set. Then, f is convex over C if and only if,*

$$\left(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\right)^T (\boldsymbol{x} - \boldsymbol{y}) \geq 0,$$

for any $\boldsymbol{x}, \boldsymbol{y} \in C$.

Second Order Characterizations of Convex Functions

When the function is twice continuously differentiable, the convexity of the function is characterized by the positive semidefiniteness of the Hessian matrix.

Theorem 5.35. *(**Second order characterization of convexity**) Let $f : C \subseteq \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function defined over the convex set C. Then f is convex if and only if the Hessian is positive semidefinite i.e. $\nabla^2 f(x) \succcurlyeq 0$ for any $\boldsymbol{x} \in C$.*

To prove the theorem it suffices to show that the gradient inequality holds true. For the opposite direction, we take f to be convex and then show that the Hessian is positive semi-definite.

We can also state conditions for strict convexity.

Theorem 5.36. *(**Sufficient second order condition for strict convexity**) Let f be a twice continuously differentiable function over a convex set $C \subset \mathbb{R}^n$. Suppose $\nabla^2 f \succ 0$, then f is strictly convex over C.*

The proof of the theorem follows the same form in one direction for the general second order characterization for convexity.

6 Convex Optimization

We now begin a central concept in non-linear optimization, namely convex optimization. Our efforts here will be to minimize a convex function $f : C \subset \mathbb{R}^n \to \mathbb{R}$, over the closed, convex set C.

We will consider explicit convex optimization problems of the form,

$$\min\ f(\boldsymbol{x}),\ \text{s.t.}\ g_i(\boldsymbol{x}) \leq 0,\ h_j(\boldsymbol{x}) \leq 0.$$

Where i, j belong to some finite index set. For the problem to be convex, inequality constraints must be themselves convex functions and equality constraints must be affine. We now note a hall mark of convex problems. The implications of the following theorem are obvious.

Theorem 6.1. *(**Local** min**=Global** min) Let $f : C \to \mathbb{R}$ be a convex function over the convex set C. Let $\boldsymbol{x}^* \in C$ be a local minimum of f over C. Then $\boldsymbol{x}^*$ is a global minimum of f over C.*

Proof. □

It follows by only a slight modification, that for a strictly convex function over a convex set, a local minimum is indeed a strict global minimum.

Definition 6.2. *(**Optimal Set**) The optimal set of a convex problem, is the set of all minimizers. Denoted,*

$$\text{argmin}\{f(\boldsymbol{x}) : \boldsymbol{x} \in C\} \tag{3}$$

We may extend this notion to any general optimization problem. Furthermore, **we may show that the optimal set of a convex problem is itself convex.**

We may also define probelms such as minimizing concave functions over concave sets as convex problems by a symmetry argument. Just multiply the objective function by -1.

All linear programming problems are convex problems.

6.1 Orthogonal Projection Operator

Given a non-empty closed, convex set C. We defin the orthogonal projection operator as $P_C : \mathbb{R}^n \to C$ as,

$$P_C(\boldsymbol{x}) = \text{argmin}\{\|\boldsymbol{y} - \boldsymbol{x}\|^2 : \boldsymbol{y} \in C\}.$$

Here $\boldsymbol{x}$ is fixed and not necessarily in C.

Geometrically, the orthogonal projection operator returns a point in C that is closest to the point $\boldsymbol{x} \in \mathbb{R}^n$. Obviously, if $\boldsymbol{x} \in C$, then $P_C(\boldsymbol{x}) = \boldsymbol{x}$.

We also notice that the OPO is defind via a convex optimization problem. Namely, minimizing the quadratic, square of the norm of the difference over the convex set C.

We now show that P_C is in fact well defined.

Theorem 6.3. *(**First Projection Theorem**) Let C be a non-empty closed convex set. Then, the problem,*

$$P_C(\boldsymbol{x}) = \text{argmin}\{\|\boldsymbol{y} - \boldsymbol{x}\|^2 : \boldsymbol{y} \in C\},$$

has a unique optimal solution.

Proof. Since the objective function is quadratic with positive definite matrix we know that the function must be coercive. It follows that it must attain at least one optimal solution. Furthermore, we may deduce that the function is strictly convex, we can conclude there exists only one optimal solution and it is attained by coerciveness. □

We know that the norm induces a metric,

$$d(\boldsymbol{x}, C) = \min_{y \in C} \|\boldsymbol{x} - \boldsymbol{y}\|.$$

We may also express this metric interms of the OPO,

$$d(\boldsymbol{x}, C) = \|\boldsymbol{x} - P_C(\boldsymbol{x})\|.$$

7 Optimization over Convex Sets

We now consider a class of optimization problems, where f is a continuously differentiable function over a closed convex set C.

Instead now of considering stationary points of a function within an unconstrained problem. We must now consider stationary points of a problem.

Definition 7.1. *(**Stationary points of a constrained problem**) Let f be a continuously differentiable function over a closed convex set C. Then $\boldsymbol{x}^* \in C$ is called a stationary point of the problem if,*

$$\nabla f(\boldsymbol{x}^*)^T(\boldsymbol{x} - \boldsymbol{x}^*) \geq 0,$$

for any $\boldsymbol{x} \in C$.

We can see, that geometrically, stationarity here means that there are no feasible descent directions at the optimal $\boldsymbol{x}^*$. Furthermore, we can deduce that stationarity is a necessary condition for a local minimum of the problem.

Theorem 7.2. *(**Stationarity as a necessary optimality condition**) Let f be a continuously differentiable function over a closed convex set C and let $\boldsymbol{x}^*$ be a local minimum of the problem. Then $\boldsymbol{x}^*$ is a stationary point of the problem.*

7.1 Stationarity in Convex Problems

Stationarity is a necessary optimality condition for local optimiality. However, when the objective function is also convex, stationarity is a necessary and sufficient condition for opitmality.

Theorem 7.3. *Let f be a continuously differentiable convex function over the closed and convex set C. Then $\boldsymbol{x}^*$ is a sationary point of the constrained optimization problem if and only if $\boldsymbol{x}^*$ is an optimal solution of the problem.*

7.2 Orthogonal Projection Revisited

We now revist the idea of the orthogonal projection operator to put forward the second projection theorem.

To give a general idea, given a closed convex set C, $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in C$. The second projection theorem asserts that the angle between $\boldsymbol{x} - P_C(\boldsymbol{x})$ and $\boldsymbol{y} - P_C(\boldsymbol{x})$ is obtuse i.e. greater than or equal to 90 degrees. More formally we have the following.

Theorem 7.4. ***(Second projection theorem)*** *Let C be a closed convex set and let $\boldsymbol{x} \in \mathbb{R}^n$. Then $\boldsymbol{z} = P_C(\boldsymbol{x})$ if and only if,*

$$(\boldsymbol{x} - P_C(\boldsymbol{x}))^T(\boldsymbol{y} - P_C(\boldsymbol{x})) \leq 0,$$

for any $\boldsymbol{y} \in C$.

8 Optimality Conditions for Linearly Constraiend Problems

In the last section we discussed the concept of stationarity within optimization problems where the objective function is differentiable and the feasible set is closed and convex. We saw that the condition is quite tricky to use apart from some simpler feasible sets.

In this section we will look at constructing a stronger optimality condition (at least from linearly constrained problems) that will be easier to use while performing optimizations. These will be the so called Karush-Kuhn-Tucker conditions for linearly contrained problems. Later on in these notes we will also extend the KKT conditions to other types of optimization problems.

8.1 Seperation and Alternative Theorems

We begin by stating a somewhat straight forward fact about convex sets that will provide a strong foundation for what is to follow.

Given a set $S \subset \mathbb{R}^n$, a hyperplane $H = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{a}^T\boldsymbol{x} = b\}$ is said to strictly seperate a point $\boldsymbol{y} \notin S$ if $\boldsymbol{a}^T\boldsymbol{y} > b$ and $\boldsymbol{a}^T\boldsymbol{x} \leq b$ for all $\boldsymbol{x} \in S$ and $\boldsymbol{y} \notin S$. Given this simple idea, consider the following theorem.

Theorem 8.1. ***(Strict Seperation Theorem)*** *Let $C \subset \mathbb{R}^n$ be a closed and convex set and let $\boldsymbol{y} \notin C$. Then there exists non-zero p and $\alpha \in \mathbb{R}$ such that,*

$$\boldsymbol{p^T y > \alpha}$$

and,

$$\boldsymbol{p^T x \leq \alpha,}$$

for all $\boldsymbol{x} \in C$.

We now use the theorem to prove Farka's lemma, which is regarded as an alternative theorem. Reason being, it states that exactly one of the two systems is feasible.

Theorem 8.2. *(**Farka's Lemma**) Let $C \in \mathbb{R}^n$ and $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Then, exactly one of the following systems has a solution.*

(i) $\boldsymbol{Ax} \leq \boldsymbol{0}$, $\boldsymbol{c}^T\boldsymbol{x} > 0$.

(ii) $\boldsymbol{A}^T\boldsymbol{y} = \boldsymbol{c}$, $\boldsymbol{y} \geq \boldsymbol{0}$.

Before carrying on, consider the following illustration. Let $\boldsymbol{c}^T = (-1, 9)$ and

$$\boldsymbol{A} = \begin{pmatrix} 1 & 5 \\ -1 & 2 \end{pmatrix}.$$

Stating the system (i) is infeasible means that the system $\boldsymbol{Ax} \leq \boldsymbol{0}$ implies that the inequality $\boldsymbol{c}^T\boldsymbol{x} \leq 0$. This is true as adding twice the second inequality to the first estabilshes the result. It follows that $\boldsymbol{c}^T$ can be written as a conic combination of rows of $\boldsymbol{A}$.

Now, in generality, the question becomes whether a set of linear inequalities in a base system can be expressed as linear inequalities in some new system if and only if the new inequalities can be written as conic combinations of inequalities in the base system. Amazingly, Farka's lemma says yes!

With this as a motivation, we may re-formulate Farka's Lemma to make it easier to prove.

Theorem 8.3. *(**Farka's Lemma, second formulation**) Let $\boldsymbol{c} \in \mathbb{R}^n$ and $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Then the following are equivalent,*

(A) The implication $\boldsymbol{Ax} \leq 0 \implies \boldsymbol{c}^T\boldsymbol{x} \leq 0$ holds true.

(B) There exists $\boldsymbol{y} \in \mathbb{R}^m_+$ such that $\boldsymbol{A}^T\boldsymbol{y} = \boldsymbol{c}$.

Another alternative theorem to consider is Gordon's Theorem.

Theorem 8.4. *(**Gordon's Alternative Theorem**) Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ then exactly one of the following systems has a solution.*

(A) $\boldsymbol{Ax} < \boldsymbol{0}$.

(B) $\boldsymbol{p} \neq \boldsymbol{0}$, $\boldsymbol{A^T p = 0}$, $\boldsymbol{p} \geq \boldsymbol{0}$.

9 KKT Conditions

We now use Gordon's theorem to state a special case of the KKT conditions. These conditions in turn extend the idea of stationarity as discussed in the previous section.

Theorem 9.1. *(**KKT Conditions for Linearly Constrained Problems**) Consider the minimization problem,*

$$\begin{array}{ll} \min & f(\boldsymbol{x}) \\ s.t. & \boldsymbol{a}_i^T\boldsymbol{x} \leq b_i, \quad i = 1, 2, \ldots, m \end{array}$$

Where f is continuously differentiable over $\mathbb{R}^n$, $\boldsymbol{a}_1, ..., \boldsymbol{a}_m \in \mathbb{R}^m$ and $b_1, ..., b_m \in \mathbb{R}$ and let $\boldsymbol{x}^*$ denote the local minimum point of the problem.

Then there exist Lagrange multipliers $\lambda_1, ..., \lambda_m \geq 0$ such that,

$$\nabla f(\boldsymbol{x}^*) + \sum_1^m \lambda_i \boldsymbol{a}_i = 0,$$

and

$$\lambda_i(\boldsymbol{a}_i^T\boldsymbol{x}^* - b_i) = 0, \quad i = 1, 2, ..., m.$$

Proof. Since $\boldsymbol{x}^*$ is a local minimum point of the problem, it follows from the stationary condition of differentiable functions over convex sets that,

$$\nabla f(\boldsymbol{x}^*)^T(\boldsymbol{x}-\boldsymbol{x}^*) \geq 0,$$

for all $\boldsymbol{x} \in \mathbb{R}^n$ such that $\boldsymbol{x}$ satisfies the constraints i.e. $\boldsymbol{a}_i^T\boldsymbol{x} \leq b_i$.

Let the set of active constraints be denoted,

$$I(\boldsymbol{x}^*) := \{i : \boldsymbol{a}_i^T\boldsymbol{x}^* = b_i\}.$$

Furthermore, we introduce the change of variable $\boldsymbol{y} = (\boldsymbol{x}-\boldsymbol{x}^*)$, making the stationarity condition $\nabla f(\boldsymbol{x}6*)^T\boldsymbol{y} \geq 0$ for any $\boldsymbol{y}$ satisfying $\boldsymbol{a}_i^T(\boldsymbol{x}^*+\boldsymbol{y}) \leq b_i$. That is for any $\boldsymbol{y} \in \mathbb{R}^n$ such that,

$$\begin{aligned} \boldsymbol{a}_i^T\boldsymbol{y} &\leq 0, \;\; i \in I(\boldsymbol{x}^*) \\ \boldsymbol{a}_i^T\boldsymbol{y} &\leq b_i - \boldsymbol{a}_i^T\boldsymbol{x}^*, \;\; i \notin I(\boldsymbol{x}^*). \end{aligned}$$

We now want to show that the latter inequalities can be removed. That is $\boldsymbol{a}_i^T\boldsymbol{y} \leq 0 \implies \nabla f(\boldsymbol{x}^*)^T\boldsymbol{y} \geq 0$.

Suppose then that $\boldsymbol{y}$ satisfies, $\boldsymbol{a}_i^T\boldsymbol{y} \leq 0, \quad i \in I(\boldsymbol{x}^*)$. Since $b_i - \boldsymbol{a}_i^T\boldsymbol{x}^* > 0$ for all $i \notin I(\boldsymbol{x}^*)$, it follows there exists a small enough $\alpha > 0$ for which $\boldsymbol{a}_i^T\boldsymbol{y} \leq b_i - \boldsymbol{a}_i^T\boldsymbol{x}^*$. Since in addition $\boldsymbol{a}_i^T\alpha\boldsymbol{y} \leq 0$, it follows by stationarity that $\nabla f(\boldsymbol{x}^*)^T(\alpha\boldsymbol{y}) \geq 0$ and hence $\nabla f(\boldsymbol{x}^*)^T\boldsymbol{y} \geq 0$. We have thus shown that,

$$\boldsymbol{a}_i^T\boldsymbol{y} \leq 0 \text{ for all } i \in I\left(\boldsymbol{x}^*\right) \Rightarrow \nabla f\left(\boldsymbol{x}^*\right)^T\boldsymbol{y} \geq 0.$$

Then, by employing Farka's lemma, there exists a $\lambda_i \geq 0$ such that

$$-\nabla f(\boldsymbol{x}^*) = \sum_{i \in I(\boldsymbol{x}^*)} \lambda_i \boldsymbol{a}_i.$$

Lastly, defining $\lambda_i = 0$ for $i \notin I(\boldsymbol{x}^*)$ we get that $\lambda_i(\boldsymbol{a}_i^T\boldsymbol{x}^* - b_i) = 0$. So,

$$\nabla f(\boldsymbol{x}^*) + \sum_i^m \lambda_i \boldsymbol{a}_i = 0.$$

□

These KKT conditions are a necessary optimality condition, but when the objective function is convex, then KKT are both necessary and sufficient global optimality conditions.

We refer to the condition $\lambda_i(\boldsymbol{a}_i^T\boldsymbol{x} - b_i) = 0$ as *complementary slackness*.

Furthermore, we may generalize the theorem to encapsulate linear equality constraints. The proof of this variant relies of the fact that a linear equality can be decomposed into two linear inequalities. The proof then follows the same structure as before. We present the theorem below.

Theorem 9.2. *(**KKT conditions for linearly constrained problems**) Consider the minimization problem,*

$$\begin{array}{ll} \min & f(\boldsymbol{x}) \\ s.t. & \boldsymbol{a}_i^T \boldsymbol{x} \leq b_i, \quad i = 1, 2, \ldots, m \\ & \boldsymbol{c}_j^T \boldsymbol{x} = d_j, \quad j = 1, 2, \ldots, p \end{array} .$$

f is continuously differentiable over $\mathbb{R}^n$, $\boldsymbol{a}_i, ..., \boldsymbol{a}_m, \boldsymbol{c}_1, ..., \boldsymbol{c}_m \in \mathbb{R}^n$ and $c_1, ..., c_m, d_1, ..., d_m \in \mathbb{R}$. Then, we have the following,

(a) If $\boldsymbol{x}^*$ is a local minimum point of the problem, then there exist $\lambda_1, ..., \lambda_m \geq 0$ and $\mu_1, ..., \mu_p \in \mathbb{R}$ such that

$$\nabla f(\boldsymbol{x}^*) + \sum_1^m \lambda_i \boldsymbol{a}_i + \sum_1^p \mu_j \boldsymbol{c}_j = 0,$$

and

$$\lambda_i (\boldsymbol{a}_i^T \boldsymbol{x}^* - b_i) = 0.$$

(b) If in addition, f is convex over $\mathbb{R}^n$ and $\boldsymbol{x}^*$ is a feasible solution of the problem and the existence of the Lagrange multipliers is as above, then $\boldsymbol{x}^*$ is an optimal solution of the problem.

Observe that the Lagrange multipliers for linear equality constraints need not be non-negative.

Furthermore, $\boldsymbol{x}^*$ is called a KKT point of the problem if (a) is satisfied.

9.0.1 The Lagrangian

A popular and compact representation of the KKT conditions is via the Lagrangian function. We present this idea here within the setting of general non-linear programming problems. Consider the porblem,

$$\begin{array}{ll} \min & f(\boldsymbol{x}) \\ \text{s.t.} & g_i(\boldsymbol{x}) \leq 0, \quad i = 1, 2, \ldots, m \\ & h_j(\boldsymbol{x}) = 0, \quad j = 1, 2, \ldots, p \end{array} .$$

Here $f, g_1, ..., g_m, h_1, ..., h_m$ are continuously differentiable over $\mathbb{R}^n$. The associated Lagrangian takes the form,

$$L(\boldsymbol{x}, \lambda, \mu) = f(\boldsymbol{x}) + \sum_1^m \lambda_i g_i(\boldsymbol{x}) + \sum_1^p \mu_i h_i(\boldsymbol{x}).$$

Notice then, we gain back the stationarity condition by taking the gradient of the Lagrangian with respect to $\boldsymbol{x}$,

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \lambda, \mu) = \nabla f(\boldsymbol{x}) + \sum_1^m \lambda_i \nabla g_i(\boldsymbol{x}) + \sum_1^p \mu_i \nabla h_i(\boldsymbol{x}) = \boldsymbol{0}.$$

Now considering the problem as in the KKT Theorem, if we define matricies $\boldsymbol{A}$,$\boldsymbol{C}$ and vectors $\boldsymbol{b}\mathbb{R}^m$ $\boldsymbol{d}\mathbb{R}^p$ such that,

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{a}_1^T \\ \vdots \\ \boldsymbol{a}_m^T \end{pmatrix}, \quad \boldsymbol{C} = \begin{pmatrix} \boldsymbol{c}_1^T \\ \vdots \\ \boldsymbol{c}_p^T \end{pmatrix}.$$

Considering this, the constraints can be written as $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ and $\boldsymbol{C}\boldsymbol{x} = \boldsymbol{d}$. Then we may re-write the Lagrangian as,

$$L(\boldsymbol{x}, \lambda, \mu) = f(\boldsymbol{x}) + \lambda^T (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) + \mu(\boldsymbol{C}\mu - \boldsymbol{d}).$$

In turn making the stationary condtion, expressed in erms of the Lagrangian as,

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \lambda, \mu) = \nabla f(\boldsymbol{x}) + \boldsymbol{A}^T \lambda + \boldsymbol{C}^T \mu = 0.$$

Duality

In general, we know convex problems are easier to solve than non-convex problems. Furthemore, outside convexity, we know that unconstrained problems are easier to solve than constrained ones.

We want to put forward a theory that somehow can use an unconstrained version of a constrained optimization problem to provide an estimate of the original constrained problem.

This notion sparks the idea of duality within mathematical optimization. Which we now construct.

Consider a non-linear minimization problem P. We find that in general, the unconstrained version of the problem P_0, has optimal solution such that,

$$\text{val}(P_0) \leq \text{val}(P).$$

In words, the optimal value of the unconstrained problem acts as a lower bound on the optimal value of the constrained problem.

If we now optimize over $\text{val}(P_0)$ i.e. find the best lower bound, we may formulate another optimization problem D such that,

$$\max\{\text{val}(P_0)\}.$$

We call this problem the **dual problem** and the original problem P, the primal problem.

We find that,

$$\text{val}(D) \leq \text{val}(P).$$

Deriving the dual problem from the primal problem is not difficult but is non-trivial. Given the primal problem P. We denote the Lagrangian as $L(\boldsymbol{x}, \lambda, \mu)$, assuming P is in general non-linear and consists of both equality and inequality constraints.

Then, the **dual objective function** is defined as,

$$q(\lambda, \mu) = \inf_{\boldsymbol{x} \in X} L(\boldsymbol{x}, \lambda, \mu).$$

When the minimum is attained, the infimum can be written explicity as a minimum.

One may notice that $q(\lambda, \mu) = -\infty$ for some values λ and μ. We thereby define the domain of q to be,

$$\text{dom}(q) = \{(\lambda, \mu) \in \mathbb{R}^m_+ \times \mathbb{R}^p : q(\lambda, \mu) > -\infty\}.$$

We may then define the dual problem explicitly as,

$$q^* = \max q(\lambda, \mu), \text{ s.t. } (\lambda, \mu) \in \text{dom}(q).$$

We then gain the following powerful theorem.

Theorem 9.3. *(**Convexity of the Dual Problem**) Consider the general constrained problem P. Let the constraint functions be finite valued defined on the set $X \subset \mathbb{R}^n$. Let q be the function as defined before, then,*

(a) $\text{dom}(q)$ *is a convex set.*

(b) q is concave over $\text{dom}(q)$*, hence $-q$ is convex over* $\text{dom}(q)$*.*

Proof. □

Exploting this fact we can formulate the so called duality theorems.

Theorem 9.4. *(**Weak Duality**) Consider the primal and dual problem. Then,*

$$\text{val}(D) \leq \text{val}(P).$$

Straightforwardly, the weak duality theorem states that the optimal value of the dual problem is a lower bound on the optimal value of the primal problem.

Strong Duality in the Convex Case

We now turn to establishing in the convex case i.e. when the primal problem is convex, **strong duality** holds. As the name suggests, the strong duality is precisely when $\text{val}(D) = \text{val}(P)$.

To do this rigourously, we will need an idea of seperating a point in a convex set from another point not necessarily within the convex set via hyperplanes.

Theorem 9.5. *(**Supporting Hyperplane Theorem**) Let $C \subset \mathbb{R}^n$ be a convex set and let $\boldsymbol{y} \notin C$. Then there exists $\mathbf{0} \neq p \in \mathbb{R}^n$ such that,*

$$\boldsymbol{p^T x} \leq \boldsymbol{p^T y},$$

for any $\boldsymbol{x} \in C$.

From here we may also deduce a seperation theorem between disjoint convex sets.

Theorem 9.6. *(**Seperation of two convex sets**) Let $C_1, C_2 \subset \mathbb{R}^n$ be two non-empty, disjoint convex sets. Then there exists $\mathbf{0} \neq p \in \mathbb{R}^n$ such that,*

$$\boldsymbol{p^T x} \leq \boldsymbol{p^T y},$$

for any $\boldsymbol{x} \in C_1$ and $\boldsymbol{y} \in C_2$.

This in turn will help in developing a non-linear versino of Farkas Lemma, which will then be the key in proving the strong duality result.

Theorem 9.7. *(**Non-Linear Farkas Lemma**) Let $X \subset \mathbb{R}^n$ be a convex set and let $f, g_1, g_2, ..., g_m$ be convex functions over X. Assume that there exists a $\hat{\boldsymbol{x}} \in X$ such that, $g_1(\hat{\boldsymbol{x}}) < 0, ..., g_m(\hat{\boldsymbol{x}}) < 0$. Let $c \in \mathbb{R}$. Then the following two claims are equivalent.*

(a) The following implication holds:

$$\boldsymbol{x} \in X,\ g_i(\boldsymbol{x}) \leq 0,\ \implies\ f(\boldsymbol{x}) \geq c.$$

(b) There exists Lagrange multipliers $\lambda_1, ..., \lambda_2$ such that,

$$\min_{\boldsymbol{x} \in X} \left\{ f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{x}) \right\} \geq c.$$

We see that the non-linear Farkas lemma shows that the convex function f is bounded below by c and this is equivalent to the minimum over $\boldsymbol{x} \in X$ of the Lagrangian being bounded below by c.

We can then turn to establishing a strong duality result with a Slater-type condition.

Theorem 9.8. *(**Strong duality for convex problems with inequality constraints**) Consider the optimization problem,*

$$f^* = \min f(\boldsymbol{x}),\ s.t.\ g_i(\boldsymbol{x}) \leq 0,\ \boldsymbol{x} \in X.$$

Here X is a convex set and $f, g_i, ..., g_m$ are convex functions over X. Suppose that there exists $\hat{\boldsymbol{x}} \in X$ such that $g_i(\hat{\boldsymbol{x}}) < 0$. Suppose also that the problem has a finite optimal value. Then the optimal value of the dual problem,

$$q^* = \max\{q(\lambda) = \min_{\boldsymbol{x} \in X} L(\boldsymbol{x}, \lambda) : \lambda \in \mathrm{dom}(q)\}.$$

Then the optimal value of the primal and dual problem are,

$$q^* = f^*.$$