Auxiliary notes on Optimisation

Daniel Lin; Professor: Dante Kalise

April 18, 2023

The official notes on optimisation are almost self-contained, but some details are omitted. This is an additional note for extensional study only, mainly collected from *Introduction to Non-Linear Optimisation* and lectures.

The following table of contents is clickable.

Contents

1 Basics

1.1 Spectral decomposition theorem

Recall diagonalisation of an $n \times n$ real symmetric matrix has satisfying properties

Theorem 1.1 (Spectral decomposition)**.** *Given symmetric $A \in \mathbb{R}^{n\times n}$, there is orthogonal U and diagonal matrix D s.t.*

$$U^T A U = D$$

elements of D are Eigenvalues of A and usually we assume WOLG they are laid in descending order (i.e. $d_1 \geq d_2 \geq \cdots d_n$ where d_i are the diagonal elements of D)

Corollary 1. *For symmetric matrix A,*

$$Tr(A) = \sum_{i=1}^{n} \lambda_i(A), \quad \det A = \prod_{i=1}^{n} \lambda_i(A)$$

where $\lambda_i(A)$ is the i'th Eigenvalue of A.

Proof. Trace is invariant under cyclic permutation, so

$$\mathrm{Tr}(A) = \mathrm{Tr}(UDU^T) = \mathrm{Tr}(U^T UD) = \mathrm{Tr}(D) = \sum_{i=1}^{n} \lambda_i(A)$$

as for determinants, this function is multiplicative.

$$\det A = \det(UDU^T) = \det U \det D \det U^T = \det U \det U^{-1} \det D = \prod_{i=1}^{n} \lambda_i(A)$$

□

Corollary 1 gives that criteria of definiteness for 2×2 matrices: $\mathrm{Tr}(A), \det A > 0 \Leftrightarrow A$ is positive definite.

Spectral decomposition also enables an upper bound for the smallest Eigenvalue $\lambda_{\min}(A)$ and a lower bound for the largest Eigenvalue $\lambda_{\max}(A)$

Corollary 2. *If A is symmetric,*

$$\lambda_{\min}(A) \leq \frac{\boldsymbol{x}^T A \boldsymbol{x}}{\|\boldsymbol{x}\|^2} \leq \lambda_{\max}(A)$$

The middle term is also called Rayleigh quotient, denoted by $R_A(\boldsymbol{x})$. This quotient is closely related to the definiteness of A.

Proof. Perform spectral decomposition on A and define change of variable $\boldsymbol{x} = U\boldsymbol{y}$ so that $\boldsymbol{x}^T A \boldsymbol{x} = \boldsymbol{y}^T D \boldsymbol{y}$. Then

$$\max_{\boldsymbol{x}\neq 0} \frac{\boldsymbol{x}^T A \boldsymbol{x}}{\|\boldsymbol{x}\|^2} = \max_{\boldsymbol{y}\neq 0} \frac{\boldsymbol{y}^T D \boldsymbol{y}}{\|\boldsymbol{y}\|^2} = \max_{\boldsymbol{y}\neq 0} \frac{\sum d_i y_i^2}{\sum y_i^2}$$

the denominator $\sum d_i y_i^2 \leq \lambda_{\max}(A) \sum (y_i^2)$ as $\lambda_{\max}(A)$ is the maximal element of D. So

$$R_A(\boldsymbol{x}) \leq \max_{\boldsymbol{x}\neq 0} \frac{\boldsymbol{x}^T A \boldsymbol{x}}{\|\boldsymbol{x}\|^2} \leq \lambda_{\max}(A)$$

$R_A(\boldsymbol{x}) \geq \lambda_{\min}(A)$ can be deduced in a similar way. □

1.2 Definiteness

We only talk about the definiteness of symmetric matrices, but you can always get a symmetric matrix by $(A + A^T)/2$.

Proposition 1.2 (Properties for positive definite matrices)**.** *If A is positive (semi-)definite,*

- *then all symmetric submatrices are positive (semi-)definite.*
- *A^{-1} exists and it is also positive (semi-)definite*
- *diagonal entries of A are (non-negative) positive*
- *element of A with the largest absolute value must be on the main diagonal*
- $\det(A) > 0$
- *If A, B are positive (semi-)definite, $A + B$ is positive (semi-)definite*

If A is negative (semi-)definite, then $-A$ is positive (semi-)definite so the above properties apply to $-A$, reversing the sign gives properties of negative (semi-)definite matrices.

Among the three criterion in lecture notes for definiteness (principal minor, strict diagonal dominance, Eigenvalue), the strict diagonal dominance criteria is NOT necessary condition. So, for example,

$$\begin{pmatrix} 3/2 & 1 & 1 \\ 1 & 3/2 & 1 \\ 1 & 1 & 3/2 \end{pmatrix}$$

has Eigenvalues $0.5, 0.5, 3.5$, it is positive definite, but diagonal dominance is not satisfied. However, this is the most convenient criterion, so you may consider trying this criteria first.
Further, without strict diagonal dominance(or even absence of diagonal dominance), A may still be positive definite. It is always worth checking other criteria or simply find $\boldsymbol{x} \neq 0$ s.t. $\boldsymbol{x}^T A \boldsymbol{x} = 0$ to prove the matrix is not positive definite.

Actually, as long as $a \geq 1$, the matrix A defined by

$$A_{ij} = \begin{cases} a & i = j \\ 1 & i \neq j \end{cases}$$

is positive semi-definite.

Principal minors $D_i(A)$ used in lectures are called leading principal minors. Following is a proper definition for principal minors.

Definition 1 (Principal minor)**.** Principal minors of $A \in \mathbb{R}^{n \times n}$ are determinants of principal submatrices A, which are obtained by deleting any rows and the corresponding columns. If only the last $n - k$ rows and columns are deleted, the submatrix is called the leading principal submatrix, and its determinant is called the leading principal minor, denoted as $D_k(A)$.

Leading principal minor criterion CANNOT be used to detect semi-definiteness, i.e. even if $D_i(A) \geq 0 \Leftrightarrow$, A may be indefinite:

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

three principal minors are $1, 0, 0$, but the matrix is indefinite. (check the Eigenvalues) This counter-example is provided by Jorge Catarecha Otero Saavedra. Even if all the diagonal entries are positive, consider the following matrix

$$A := \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & a \end{pmatrix}$$

where $0 < a < 1$. The three leading principal minors are $1, 0, 0$ (independent of a!), all non-negative. But consider $\boldsymbol{x}$ s.t. $x_1 + x_2 + x_3 = 0$

$$\boldsymbol{x}^T A \boldsymbol{x} = (x_1 + x_2 + x_3)^2 + (a-1)z^2 = (a-1)z^2 < 0$$

Though leading principal minors do not work, there is a criterion for semi-definiteness: If all principal minors(i.e. determinants of all possible combinations of rows and corresponding columns) are non-negative, then A is semi-definite.

There is another iff condition for positive definiteness, though not useful in practice: $A \succ 0 \Leftrightarrow$ exists invertible M s.t. $A = MM^T$. You can assemble M in another order as shown in the proposition below. There is also a version for the semi-definite matrices: $A \succeq 0 \Leftrightarrow$ there is a square matrix M s.t. $A = MM^T$. (dropping the requirement for M to be invertible) This is also related to the fact that any positive semi-definite matrix have a square root. The following two tricks are possible with square root:

- If A is positive semi-definite, $\boldsymbol{x}^T A \boldsymbol{y} = \langle A^{1/2}\boldsymbol{x}, A^{1/2}\boldsymbol{y}\rangle$.
- $\boldsymbol{x}^T A \boldsymbol{x} = \|A^{1/2}\boldsymbol{x}\|^2$

Proposition 1.3. *For any matrix A(may not be square), $A^T A$ is symmetric and positive semi-definite. If given further that A has full column rank, $A^T A \succ 0$.*

Proof.

$$\boldsymbol{x}^T A^T A \boldsymbol{x} = \|A\boldsymbol{x}\|^2 \geq 0$$

so $A^T A$ is always positive semi-definite. If A has full rank, $A\boldsymbol{x} = 0$ iff $\boldsymbol{x} = 0$, so $\boldsymbol{x}^T A^T A \boldsymbol{x} > 0$ for non-zero $\boldsymbol{x}$, i.e. $A \succ 0$. □

Outer product of a vector: $\boldsymbol{x}\boldsymbol{x}^T$ is positive semi-definite but NOT definite, because you can always find a vector $\boldsymbol{y} \neq 0$ orthogonal to $\boldsymbol{x}$. So $\boldsymbol{y}^T \boldsymbol{x}\boldsymbol{x}^T \boldsymbol{y} = 0$. (unless you are working on one-dimensional space)

There are only two criteria for the indefiniteness of the matrix in this course:

- There is a positive and a negative Eigenvalue.
- There are $\boldsymbol{x}, \boldsymbol{y}$ s.t. $\boldsymbol{x}^T A \boldsymbol{x} > 0$ and $\boldsymbol{y}^T A \boldsymbol{y} < 0$
- Or for 2×2 matrix only, proving $\det(A) < 0$ is enough. (use Corollary 1)

Another criterion not proved in this course is: if there is one positive and one negative element on the diagonal, A is indefinite.

Proposition 1.4. *Covariance matrix of any vector of random variables $\boldsymbol{X}$ is positive semi-definite.*

Proof. This becomes obvious using the formulation of covariance matrix $Q = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^T$ where n is the length of random vector $\boldsymbol{x}$, $\overline{x}$ is the mean value of $\boldsymbol{x}$. □

1.2.1 Summary of criterion for definiteness and semi-definiteness

- Diagonal dominance(strict):
  - for positive definiteness/semi-definiteness:

    $$|A_{ii}| \geq \sum_{j \neq i} |A_{ij}|$$

    and $A_{ii} \geq 0$ for all i.
  - for negative definiteness/semi-definiteness:

    $$|A_{ii}| \geq \sum_{j \neq i} |A_{ij}|$$

    and $A_{ii} \leq 0$ for all i.

- Diagonal dominance(strict):
  - for positive definiteness/semi-definiteness:

$$|A_{ii}| > \sum_{j \neq i} |A_{ij}|$$

    and $A_{ii} > 0$ for all i.
  - for negative definiteness/semi-definiteness: $D_k(A) < 0$ for odd k, $D_k(A) \geq 0$ for even k

$$|A_{ii}| > \sum_{j \neq i} |A_{ij}|$$

    and $A_{ii} < 0$ for all i.
- Principal minor(non-strict):
  - for positive definiteness/semi-definiteness: $D_k(A) \geq 0$ for all k where D_k is the kth leading principal minor
  - for negative definiteness/semi-definiteness: $D_k(A) \leq 0$ for odd k, $D_k(A) \geq 0$ for even k
- Principal minor(strict):
  - for positive definiteness/semi-definiteness: $D_k(A) > 0$ for all k where D_k is the kth leading principal minor
  - for negative definiteness/semi-definiteness: $D_k(A) < 0$ for odd k, $D_k(A) \geq 0$ for even k
- Eigenvalue(non-strict):
  - for positive definiteness/semi-definiteness: all Eigenvalues are non-negative
  - for negative definiteness/semi-definiteness: all Eigenvalues are non-positive
- Eigenvalue(strict):
  - for positive definiteness/semi-definiteness: all Eigenvalues are positive
  - for negative definiteness/semi-definiteness: all Eigenvalues are negative

Criterion	necessary for semi-def.	sufficient for semi-def.	necessary for def.	sufficient for def.
Diagonal dominance(non-strict)		✓		
Diagonal dominance(strict)		✓		✓
Principal minor(non-strict)	✓		✓	
Principal minor(strict)		✓	✓	✓
Eigenvalue(non-strict)	✓	✓		
Eigenvalue(strict)		✓	✓	✓

1.3 Transformation

Given $f : \mathbb{R}^n \to \mathbb{R}$, you may find the transformation $g(x) := f(Ax + b)$ where $x, b \in \mathbb{R}^n$ useful. Gradient and Hessian of g are given below:

$$\nabla g(x) = A(\nabla f(Ax + b))^T, \quad \nabla^2 g(x) = A^T \nabla^2 f(Ax + b) A$$

these can be proved by chain rule and writing out terms element-wise.

1.4 Derivatives

Derivatives of linear, quadratic functions.

$f(\boldsymbol{x})$	$\nabla f(\boldsymbol{x})$
$\boldsymbol{a}^T\boldsymbol{x} + \beta$	$\boldsymbol{a}$
$\boldsymbol{x}^T A \boldsymbol{x}$	$(A^T + A)\boldsymbol{x}$

2 Unconstrained optimisation

Rule of thumb to remember: minimising f is equivalent to maximising $-f$. You can use this to prove a version of maxima for theorems on minima for free. e.g. f attains global minimum point if f is continuous and coercive. But if we know

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

(for this note, I will call this property *anti-coercive*) then $-f$ is coercive, so $-f$ attains global minima. i.e. f attains global maxima. Of course, in an exam, you only have the minimum version from the official note, so you have to use the theorem on $-f$.

Regarding second-order optimality conditions, they all assume the following two conditions

- f is twice continuously differentiable on the required domain. So you should be careful with functions involving $\|x\|$.
- x^* is a stationary point.

Tricky cases of second order optimality conditions are $\nabla^2 f(x^*) \succeq 0, \nabla^2 f(x^*) \preceq 0$. x^* may be local extrema or saddle point. You must use definitions in these cases. i.e. find an open ball around x^* s.t. $f(x^*)$ is the smallest/largest in that ball, or find two trajectories $x = x(t), y = y(t)$ starting at x^* s.t. x^* is minima on one trajectory but maxima on the other to prove x^* is saddle point.

Optimising on balls: When the domain is $S = \{\boldsymbol{x} \in \mathbb{R}^n \,|\, \|x\| \leq r\}$, i.e. a ball, we need to play tricks on functions. Either differentiate for the interior of S and consider the boundary separately or write the function as inner product form $a^T\boldsymbol{x}$. Then maxima is obtained at $\boldsymbol{x} = ra/\|a\|$ and maximal value is $r\|a\|$. If you have an oval $\{\boldsymbol{x} = (x, y) \in \mathbb{R}^2 \,|\, x^2/a + y^2/b = 1 \leq r\}$ instead, change of variable $u = x/\sqrt{a}, v = y/\sqrt{b}$ should be performed first.

2.1 Global Optimality

Pushing local extrema to global: Assume S is a non-empty closed set

- If f is coercive, and you only found one stationary point on S, it must be global minima (on S).
- If f(defined) is anti-coercive and you only found one stationary point on S, it must be global maxima (on S).

If f is coercive/anti-coercive, but you cannot find any stationary point for the interior of S, the extrema must be on the boundary of S.

Proving local extrema are not global:

- If x^* is local minima, either prove f is not bounded below or find a smaller local minimum to show x^* is not global minima.
- If x^* is the local maxima, either prove f is not bounded above or find a larger local maximum to show x^* is not global maxima.

2.2 Quadratic functions

A quadratic function is in the following form

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

where A is usually assumed to be symmetric.

Quadratic functions are important as, for example, the linear least square problem of finding x s.t. $\|Ax - b\|^2$ is minimised and is essentially a quadratic after expanding it.

$$\|Ax - b\|^2 = (Ax - b)^T(Ax - b) = x^T A^T Ax - b^T Ax - x^T A^T b + b^T b = x^T A^T Ax - 2b^T Ax + b^T b$$

So gradient, Hessian, and optimality conditions for quadratic functions are studied in depth.

Roles of $A, \boldsymbol{b}, \boldsymbol{c}$ in quadratic functions

- A determines the main behaviours of the function, including whether maxima/minima exists
- $\boldsymbol{b}$ determines location of extrema together with A, namely $\boldsymbol{x}$ s.t. $A\boldsymbol{x} = -b$
- $\boldsymbol{c}$ only determines f's value at extrema.

The Coerciveness of quadratic functions is also determined by A

Proposition 2.1. *Quadratic function is coercive iff* $A \succ 0$.

Proof. By corollary 2, $\boldsymbol{x}^T A\boldsymbol{x} \geq \lambda_{\min}(A)\|x\|^2$, so LHS blows to ∞ if $\|x\| \to \infty$. And since $\boldsymbol{x}^T A\boldsymbol{x}$ is the leading term of quadratic function, f blows to ∞. (You need to use Cauchy-Schwarz inequality or equivalent to prove this rigorously)

The converse is proved by contradiction, assume A has negative eigenvalue λ i.e. $A\boldsymbol{v} = \lambda\boldsymbol{v}$ for some vector $\boldsymbol{v}$.

$$f(\alpha\boldsymbol{v}) = \lambda\|\boldsymbol{v}\|^2\alpha^2 + 2(\boldsymbol{b}^T\boldsymbol{v})\alpha + c \to -\infty$$

as $\alpha \to -\infty$. Again you need some inequalities to prove this rigorously, but intuitively $\lambda\|\boldsymbol{v}\|^2\alpha^2 < 0$ is the leading term. You also need contradictions for A having 0 Eigenvalue, left as an exercise. □

3 Gradient Descent

After having a scratch of gradient descent algorithm, or any other iterative algorithm, there is always something to consider

- **Initial values**: this will affect the convergence and even where the iteration arrives. Especially for functions with multiple local minima, different initial values lead to different minima.
- **Details of each iteration**: this depends on the algorithm considered. For gradient descent, the details are the descent direction (given by $-\nabla f$ as the gradient is essentially the steepest ascent. See the video Khan Academy for explanation), and step size/learning rate
- **Convergence criteria**: convergence of which variable would you use as criteria for convergence? Criterion is usually in the form $\|...\| < \epsilon$ where ϵ is called tolerance. How small tolerance should be is another topic. For gradient descent, criterion can be either $\|f(x_{k+1}) - f(x_k)\|$ or $\|\nabla f(x_{k+1})\|$. The second criterion is more promising as later we'll prove $\nabla f(x_k) \to 0$.

There are three ways to choose a step size

1. **constant step size**: choosing descent direction $d_k = -\nabla f(x_k)$ only ensures you will descent in this direction in a SMALL neighbourhood of x_k. So picking a large step size cause failure of decrease, but a small step size will give you slow convergence. Overall, the constant step size is difficult to choose.
2. **Exact line search**: each iteration gradient descent is only focused on one direction d_k, so it is essentially a 1-dimensional problem. Therefore optimising on this line and finding the step size giving the greatest descent is acceptable.
3. But since greatest descent is not required for convergence, a numerical method called **Backtracking line search**(BLS) can be used instead to find step size giving a "relatively" large descent. The idea is to first pick a larger step-size s, so you get the point $x_k + sd_k$ on the line $x_k + td_k$ $(t > 0)$. Let $t_0 = s$. Then you gradually shrink this step size by multiplication of a constant $\beta \in (0, 1)$. In Armijo's original paper (Armijo, 1966), he used $\beta = 1/2$. The value of step sizes while shrinking are $t_j = \beta^j t_0$.
   Descent rate is measured by $f(x_k) - f(x_k + td_k)$ (Note: k is considered as fixed when shrinking, we are working in one iteration of gradient descent) Using Maclaurin expansion,

   $$f(x_k) - f(x_k + td_k) = -t\nabla f(x_k)^T d_k - o(t\|d_k\|)$$

   $-t\nabla f(x_k)^T d_k > 0$ as d_k is descent direction. $o(t\|d_k\|) > 0$, so LHS $< -t\nabla f(x_k)^T d_k$. But we want inequality in the other direction, i.e. lower bound for the descent rate. Multiplying constant $\alpha \in (0, 1)$ to $t\nabla f(x_k)^T d_k$ gives a milder condition.

   $$f(x_k) - f(x_k + td_k) = -\alpha t\nabla f(x_k)^T d_k - (1 - \alpha)t\nabla f(x_k)^T d_k - o(t\|d_k\|)$$

Now using limit

$$\lim_{t\downarrow 0} \frac{(1-\alpha)t\nabla f(x_k)^T d_k + o(\|d_k\|t)}{t} = (1-\alpha)\nabla f(x_k)^T d_k < 0$$

therefore, for small enough t,

$$f(x_k) - f(x_k + td_k) \geq -\alpha t \nabla f(x_k)^T d_k$$

This will be used as a criterion to stop shrinking t_j, i.e. when t_j satisfies

$$f(x_k) - f(x_k + t_j d_k) \geq -\alpha t_j \nabla f(x_k)^T d_k$$

when $d_k := -\nabla f(x_k)$, this is equivalent to

$$f(x_k) - f(x_k + t_j d_k) \geq \alpha t_j \|\nabla f(x_k)\|^2$$

If your initial choice s already satisfies this, maybe your s is too small.
The choice of α, β for backtracking line search is another topic to study, out of the scope of this course.

Exact line search for quadratic functions is relative easy to find: assume A is positive definite, $f(x) = x^T Ax + 2b^T x + c$ where $b \in \mathbb{R}^n, c \in \mathbb{R}$. Then the mission is find $\lim_{t\geq 0} f(x + td)$ where $d = -\nabla f = -(2Ax + b)$. After long and boring expansion of $f(x + td)$,

$$g(t) := f(x + td) = (d^T Ad)t^2 + 2(d^T Ax + d^T b)t + f(x)$$

now $g'(t) = 2(d^T Ad)t + 2d^T(Ax + b)$ so stationary point is

$$t = -\frac{d^T 2(Ax + b)}{2d^T Ad} = -\frac{d^T \nabla f(x)}{2d^T Ad}$$

$g''(t) = 2(d^T Ad) > 0$ for all t as A is positive definite. So the answer above gives minima, and $t \geq 0$.
If your descent direction happens to be $d = -\nabla f(x)$,

$$t = \frac{d^T d}{2d^T Ad}$$

but you need to be careful in cases like Gauss-Newton, Newton's method and scaled gradient descent where the descent direction is NOT $-\nabla f(x)$.

3.1 Kantorovich inequality

We can define a partial order for symmetric matrices:

Definition 2 (Loewner Order)**.** For symmetric matrices A, B, $A \prec B$ if $B - A \succ 0$, and $A \preceq B$ if $B - A \succeq 0$. Note positive definiteness $A \succ 0$, and semi-definite $A \succeq 0$ are special cases of this partial order.

Proposition 3.1 (Equivalent condition)**.** *$A \prec B$ iff all Eigenvalues of A are smaller than that of B i.e. $\lambda_i(A) < \lambda_j(B)$ for all i, j, and $A \preceq B$ iff $\lambda_i(A) \leq \lambda_j(B)$ for all i, j.*

Proposition 3.2. *If $A \prec B$, then $x^T Ax \leq x^T Bx$ for any vector x.*

Proof. Left as an easy exercise. □

We will use matrix order later. For now, consider optimisation of $f(x) = x^T Ax$ where $A \succ 0$, which is quadratic, using the step size we discussed in the above session,

$$t_k = \frac{d_k^T d_k}{2d_k^T Ad_k}$$

where $d_k = -2Ax_k$. Let us find a relation between $f(x_{k+1})$ and $f(x_k)$ to study convergence,

$$f(x_{k+1}) = (x_k + t_k d_k)^T A(x_k + t_k d_k)^T$$

after expansion, you get $x_k^T Ax_k - t_k d_k^T d_k + t_k^2 d_k^T Ad_k$. Then plug in t_k above gives you

$$x_k^T Ax_k \left(1 - \frac{1}{4}\frac{(d_k^T d_k)^2}{4(d_k^T Ad_k)(x_k^T Ax_k)}\right)$$

but $x_k^T Ax_k = x_k^T AA^{-1}Ax_k = (d_k^T A^{-1}d_k)/4$, so we can remove x_k from the above equation

$$= f(x_k)\left(1 - \frac{1}{4}\frac{(d_k^T d_k)^2}{(d_k^T Ad_k)(d_k^T A^{-1}d_k)}\right)$$

To study convergence, we need an upper bound for the ratio $1-$the crazy fraction, i.e. lower bound for the fraction given by Kantorovich inequality.

Lemma 3.3. *For $x \neq 0$ and matrix $A \succ 0$,*

$$\frac{(x^T x)^2}{(x^T Ax)(x^T A^{-1}x)} \geq \frac{4Mm}{(M+m)^2}$$

where $M := \lambda_{\max}(A), m = \lambda_{\min}(A)$

Proof. Eigenvalues for $A + MmA^{-1}$ are considered, and the spectrum is exactly

$$\left\{\lambda_i(A) + \frac{Mm}{\lambda_i(A)} : \lambda_i(A) \in \mathrm{Spec}(A)\right\}$$

using calculus, $\phi(t) = t + Mm/t$ defined on $[m, M]$ has maximum attained at $t = M, m$ with value $M + m$, so Eigenvalues of $A + MmA^{-1}$ are smaller than $M + m$, i.e.

$$A + MmA^{-1} \preceq (M+m)I$$

so by Proposition 3.2,

$$x^T Ax + Mm(x^T A^{-1}x) \leq (M+m)(x^T x)$$

using the fact that $\alpha\beta \leq (\alpha+\beta)^2/4$,

$$x^T Ax[Mm(x^T A^{-1}x)] \leq \frac{1}{4}\left[x^T Ax + Mm(x^T A^{-1}x) \leq (M+m)(x^T x)\right]^2 \leq \frac{(M+m)^2}{4}(x^T x)^2$$

rearranging yields the required inequality. □

Then you can prove the theorem on the notes that strongly suggest the behaviour of gradient descent on this quadratic function:

Theorem 3.4.

$$f(x_{k+1}) \leq \left(\frac{M-m}{M+m}\right)^2 f(x_k)$$

Note dividing by m, the ratio becomes

$$\left(\frac{\kappa(A)-1}{\kappa(A)+1}\right)^2$$

where $\kappa(A) := M/m = \lambda_{\max}(A)/\lambda_{\min}(A)$ is called condition number. This upper bound becomes small if $\kappa(A) \approx 1$, which leads to faster convergence. But larger condition numbers(ill-conditioned) give slower convergence. Similar behaviours appears for general quadratic function as A is the Hessian matrix for any $f(x) = x^T Ax + 2b^T x + c$.

A modified gradient descent is to do transformation $x = S_k y$ so $g(y) := f(S_k y)$ becomes a new function to optimise (new S_k can be chosen for each iteration). The advantage is that Hessian of g, according to transformation rules in section 1.3,

$$\nabla^2 g(y) = S_k \nabla^2 f(S_k y) S_k$$

can be adjusted to have a better condition number. The new descent direction is $-D_k \nabla f(x_k)$ where $D_k := S_k S_k^T$ is called *scaling matrix*. Choosing $D_k = (\nabla^2 f(x_k))^{-1}$ makes sure $\nabla^2 g(y_k) = S_k \nabla^2 f(S_k y) S_k = I$ as $S_k = D_k^{1/2}$. $\kappa(I) = 1$, the best conditioned matrix ever. But matrix inversion is very time-consuming. Simpler D_k, such as a diagonal matrix, works the best. Note with, $D_{ii} := (\nabla^2 f(x_k))_{ii}^{-1}$, diagonal elements of $\nabla^2 g(y_k)$ are 1, so the Hessian will also be well-conditioned.

3.2 Fermat-Weber Problem

This problem is called the "facility location problem" imagine you are a land agent trying to plan the location of a new airport or train station $\boldsymbol{x}$, and there are many cities/towns $\boldsymbol{a}_i$ around. You seek to ensure the facility is not distanced too far from someone, i.e. minimise

$$f(\boldsymbol{x}) = \sum_{i=1}^{m} \|\boldsymbol{x} - \boldsymbol{a}_i\|$$

But maybe people in some cities are wealthier, or the population is larger; you want to maximise your benefit by adding weights ω_i to the distances.

$$f(\boldsymbol{x}) = \sum_{i=1}^{m} \omega_i \|\boldsymbol{x} - \boldsymbol{a}_i\|$$

From the official notes, you already know an iterative process (Weizfeld's method) can be used to find the stationary point of f i.e.

$$\boldsymbol{x}_{k+1} = T(\boldsymbol{x}_k), \quad \text{where } T \text{ is operator } T(\boldsymbol{x}) = \frac{1}{\sum_i \omega_i / \|\boldsymbol{x} - \boldsymbol{a}_i\|} \sum_i \frac{\omega_i \boldsymbol{a}_i}{\|\boldsymbol{x} - \boldsymbol{a}_i\|}$$

But there are several things to consider here:

- The assumption $\boldsymbol{x} \neq \boldsymbol{a}_i$ for any i is made, how to ensure $\boldsymbol{x}_k$ does not become one of $\boldsymbol{a}_i$ during iteration.
- Will this iterative process converge?

For the rest of this section, we will study these problems. Though these are not required for exams, the proofs involve useful techniques and provide you with the basic routes of proving convergence.

The first thing is to show $f(\boldsymbol{x}_k)$ is non-increasing. We can start by using the fact that the value of $T(x)$ can be viewed as an optimisation problem of the auxiliary function

$$h(y, x) = \sum_{i=1}^{m} \omega_i \frac{\|\boldsymbol{y} - \boldsymbol{a}_i\|^2}{\|\boldsymbol{x} - \boldsymbol{a}_i\|}$$

Lemma 3.5. *$T(x) = Argmin_{\boldsymbol{y}}\{h(\boldsymbol{y}, \boldsymbol{x})\}$*

$h(y, x)$ function has some sweet properties connecting to $f(\boldsymbol{x})$

Lemma 3.6. *$h(\boldsymbol{x}, \boldsymbol{x}) = f(\boldsymbol{x})$ and $h(\boldsymbol{y}, \boldsymbol{x}) \geq 2f(\boldsymbol{y}) - f(\boldsymbol{x})$*

Proof. Left as an exercise. (Hint: for the target inequality, use the inequality $a^2/b \geq 2a - b$ for any $a \geq 0, b > 0$) □

Let's prove Lemma 3.5,

Proof. Fix $\boldsymbol{x}$, $h(\cdot, \boldsymbol{x})$ can be viewed as quadratic function. After expanding, you would find the leading term is

$$\boldsymbol{y}^T \underbrace{\left(\sum_i \frac{\omega_i}{\|\boldsymbol{x} - \boldsymbol{a}_i\|} \right) I}_{=:A} \boldsymbol{y}$$

note matrix $A \succ 0$. Hence, by proposition on page 27 of lecture notes, $y^* := -A^{-1}b$, is the unique strict global minimum point. So

$$\nabla_y h(\boldsymbol{y}^*, \boldsymbol{x}) = 2 \sum_i \omega_i \frac{\boldsymbol{y}^* - \boldsymbol{a}_i}{\|\boldsymbol{x} - \boldsymbol{a}_i\|} = 0$$

extracting $\boldsymbol{y}^*$ gives $\boldsymbol{y}^* = T(\boldsymbol{x})$ □

Lemma 3.7. *Let $\boldsymbol{x}_k$ be the sequence taken from Weizfeld's method and $\boldsymbol{x}_k \neq \boldsymbol{a}_i$. Then*

- *sequence $\{f(\boldsymbol{x}_k)\}$ is non-increasing.*

- $f(\boldsymbol{x}_k) = f(\boldsymbol{x}_{k+1})$ *iff* $\nabla f(\boldsymbol{x}_k) = 0$. *If decrement is stopped, we have found a stationary point of* f.

Proof. For the first statement, just have to prove $f(T(\boldsymbol{x})) \leq f(\boldsymbol{x})$. By Lemma 3.5,

$$h(T(\boldsymbol{x}), \boldsymbol{x}) \leq h(\boldsymbol{x}, \boldsymbol{x}) = f(\boldsymbol{x})$$

On the other hand, by the inequality in Lemma 3.6,

$$h(T(\boldsymbol{x}), \boldsymbol{x}) \geq 2f(T(\boldsymbol{x})) - f(T(\boldsymbol{x}))$$

so

$$2f(T(\boldsymbol{x})) - f(T(\boldsymbol{x})) \leq h(T(\boldsymbol{x}), \boldsymbol{x}) \leq f(\boldsymbol{x}) \qquad (*)$$

which yields the required inequality.

For the second statement, $\nabla f(\boldsymbol{x}) = 0 \Leftrightarrow \boldsymbol{x} = T(\boldsymbol{x})$ so we have to prove $\boldsymbol{x} = T(\boldsymbol{x}) \Leftrightarrow f(\boldsymbol{x}) = f(T(\boldsymbol{x}))$. By $(*)$, if $f(\boldsymbol{x}) = f(T(\boldsymbol{x}))$, then

$$h(T(\boldsymbol{x}), \boldsymbol{x}) = h(\boldsymbol{x}, \boldsymbol{x}) = f(\boldsymbol{x})$$

by uniqueness of the minimiser $T(\boldsymbol{x})$ for $h(\cdot, \boldsymbol{x})$, $\boldsymbol{x} = T(\boldsymbol{x})$. □

With the above lemma, we can pick $\boldsymbol{x}_0$ s.t.

$$f(\boldsymbol{x}_0) < \min_i \{f(\boldsymbol{a}_i)\}$$

then any $\boldsymbol{x}_k$ will not be in the set $\{\boldsymbol{a}_i\}$. With this assumption, we can state the main theorem:

Theorem 3.8. *Let* $\{\boldsymbol{x}_k\}$ *be the sequence taken from Weizfeld's method and* $\boldsymbol{x}_0$ *satisfies*

$$f(\boldsymbol{x}_0) < \min_i \{f(\boldsymbol{a}_i)\}$$

then any limit point of $\{\boldsymbol{x}_k\}$ *is stationary point of* f.

Proof. If subsequence $\{\boldsymbol{x}_{k_n}\}$ converges to $\boldsymbol{x}^*$, want to show $\nabla f(\boldsymbol{x}^*) = 0$. By monotonicity of sequence $\{f(\boldsymbol{x}_k)\}$,

$$f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}_0) < f(\boldsymbol{a}_i \ \forall i$$

so $\boldsymbol{x}^* \neq \boldsymbol{a}_i$ for any i, $\nabla f(\boldsymbol{x}^*)$ is well-defined.

Note $\boldsymbol{x}_{k_n+1} = T(\boldsymbol{x}_{k_n}) \to T(\boldsymbol{x}^*)$. Sequence $\{f(\boldsymbol{x}_k)\}$ is decreasing and bounded below by 0 so it converges to some f^*. So all subsequences, including $\{f(\boldsymbol{x}_{k_n})\}, \{f(\boldsymbol{x}_{k_n+1})\}$, converges to f^*. By continuity of f, $f(T(x^*)) = f(x^*) = f^*$. Then since $f(T(\boldsymbol{x})) = f(\boldsymbol{x}) \Rightarrow T(\boldsymbol{x}) = \boldsymbol{x} \Rightarrow \nabla f(\boldsymbol{x}) = 0$, x^* is the stationary point. □

The above theorem can be improved: all limit points of $\{\boldsymbol{x}_k\}$ are the global minimum and gives the same value $f(x^*)$ where x^* is the limit point. (Proof beyond scope)

3.3 Convergence of gradient descent

The concept of Lipschitz continuity is fundamental. The official notes have already shown that the existence of a Lipschitz constant L of the gradient is equivalent to an upper bound of Hessian. A small remark on the notation $C^{1,1}$: C^1 is the space of a continuously differentiable function, and it guarantees the existence of a gradient. The extra 1 means the gradient is Lipschitz continuous. Functions in $C_L^{1,1}$ have another vital property assisting theorems on convergences:

Lemma 3.9 (Descent lemma)**.**

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

Proof. By the fundamental theorem of calculus,

$$f(y) - f(x) = \int_0^1 \langle \nabla f(x + t(y - x)), (y - x) \rangle \, \mathrm{dt}$$

then by the trick of adding 0,

$$f(y) - f(x) = \langle \nabla f(x), (y - x) \rangle + \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), (y - x) \rangle \, \mathrm{dt}$$

Only have to show

$$\left| \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), (y - x) \rangle \, \mathrm{dt} \right| \leq \frac{L}{2} \|x - y\|^2$$

This follows a similar process to the proof of the theorem (Equivalence to Boundedness of the Hessian) in official notes. □

Descent lemma can be used on gradient descent, i.e. substitute $y = x - t\nabla f(x)$, we get the following result (*sufficient decrease inequality*),

$$f(x) - f(x - t\nabla f(x)) \geq t \left(1 - \frac{Lt}{2}\right) \|\nabla f(x)\|^2$$

So for constant step size $t_k \equiv t$, the descent rate is sufficient by the following inequality

$$f(x_k) - f(x_{k+1}) \geq t \left(1 - \frac{Lt}{2}\right) \|\nabla f(x_k)\|^2$$

Of course, we require $t \in (0, 2/L)$ for this lower bound to be positive. The maximiser for this lower bound is $t^* = 1/L$. And the descent rate bound becomes

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L} \|\nabla f(x_k)\|^2$$

which is the bound for exact linear search as exact linear search looks for the fastest descent.

For backtracking, recall an initial step size $t_0 = s$ is picked and then t_k is shrunk by factor β each time, until reaching criterion

$$f(x_k) - f(x_k - t_k \nabla f(x_k)) \geq \alpha t_k \|\nabla f(x_k)\|^2$$

Let's find lower bound for t_k, assuming that it is the first t_i satisfying above criterion, so $t_{k-1} = t_k/\beta$ does not satisfy the criterion, i.e.

$$f(x_k) - f(x_k - \frac{t_k}{\beta} \nabla f(x_k)) < \alpha \frac{t_k}{\beta} \|\nabla f(x_k)\|^2$$

but according to sufficient decrease inequality,

$$f(x_k) - f\left(x_k - \frac{t_k}{\beta} \nabla f(x_k)\right) \geq \frac{t_k}{\beta} \left(1 - \frac{Lt_k}{2\beta}\right) \|\nabla f(x_k)\|^2$$

So

$$\frac{t_k}{\beta} \left(1 - \frac{Lt_k}{2\beta}\right) < \alpha \frac{t_k}{\beta}$$

$$\Rightarrow t_k > \frac{2(1 - \alpha)\beta}{L}$$

considering $t_k \leq s$, we can get a general lower bound for the descent rate shown below

$$f(x_k) - f(x_k - t_k \nabla f(x_k)) \geq \alpha \min \left\{ s, \frac{2(1 - \alpha)\beta}{L} \right\} \|\nabla f(x_k)\|^2$$

Collecting the results above together:

Lemma 3.10 (Sufficient Decrease Rate)**.** *The descent rate for three step-size choices is sufficient in the sense of inequality below,*

$$f(x_k) - f(x_k - t_k \nabla f(x_k)) \geq M \|\nabla f(x_k)\|^2$$

where

$$M = \begin{cases} t\left(1 - \frac{tL}{2}\right) & \text{constant size} \\ \frac{1}{2L} & \text{exact line search} \\ \alpha \min\left\{s, \frac{2(1-\alpha)\beta}{L}\right\} & \text{backtracking} \end{cases}$$

You can see the importance of the Lipschitz constant L again; it almost determines the convergence rate. Conclusions like $f(x_{k+1}) = f(x_k)$ iff $\nabla f(x_k) = 0$ and $\nabla f(x_k) \to 0$ follows easily from above lemma. Further, the convergence rate of $\nabla f(x_k)$ is given by the theorem below

Theorem 3.11. *If $f(x_k) \to f^*$, then for any n,*

$$\min_{0 \leq k \leq n} \|\nabla f(x_k)\| \leq \sqrt{\frac{f(x_0) - f^*}{M(n+1)}}$$

where M is the same as that in lemma 3.10.

Proof. Left as an exercise. □

3.4 Newton's method

We have seen that matrix A is essential for a quadratic function $f(x) = x^T Ax + 2b^T x + c$, and when $A \succ 0$, $x^* = -A^{-1}b$ which is extremely convenient. What about general functions? We can use quadratic approximation! Assume we start with x_0

$$f(x) \approx f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{1}{2}(x - x_0)^T \nabla^2 f(x_0)(x - x_0)$$

then approximately the minimum satisfies (need to use a change of variable $y = x - x_0$) $\nabla^2 f(x_0) y = -\nabla f(x_0)$ according to properties of quadratic functions.
If luckily $\nabla^2 f(x) \succ 0$ for all x (such functions are said to be *concave up*), then the following equation leads x_0 to x_1, a point closer to the minimum:

$$x_1 = x_0 - (\nabla^2 f(x_0))^{-1} \nabla f(x_0)$$

iterating this process gives pure Newton's method, i.e.

$$x_{n+1} = x_n - (\nabla^2 f(x_n))^{-1} \nabla f(x_n)$$

Note Newton's method for finding the zero of a function $g(x)$ is

$$x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)}$$

if $g(x) = \nabla f(x)$, this is exactly the iteration we derived above.

3.5 Kaczmarz method

Suppose given a system of linear equations, $Ax = b$, where $A \in \mathbb{R}^{m \times n}$ i.e. given equations $a_i^T x = b_i$ where a_i^T is the i'th row of A. Kaczmarz gives an iterative method which moves x along one row a_i^T at a time (Kaczmarz, 1937). Geometrically, his idea is a sequence of projections. Taking $\mathbb{R}^2$ as an example, we have two lines $a_1^T x = b_1, a_2^T x = b_2$ and want to find the intersection x^* as shown in figure 1.

Pick x_0 arbitrarily and project it to one line getting x_1. Continue this process until convergence. From the graph, you can see the convergence is quite fast.

In general, such lines are defined as hyperplanes in the form $\mathcal{H} = \{x \in \mathbb{R}^n : \langle a, x \rangle = b\}$ where $a \in \mathbb{R}^n, b \in \mathbb{R}$. We must find a projection from arbitrary $x \in \mathbb{R}^n$ to the hyperplane. Note if $b = 0$, a is the orthogonal direction of the hyperplane, so the projection point should be in the form $x^* := x - sa/\|a\|$ (here, a is normalised) for some

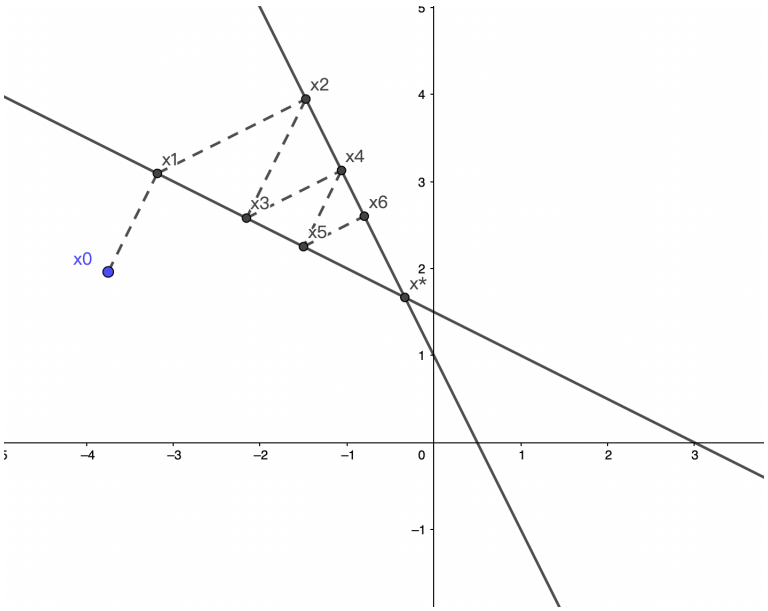

Figure 1: Image representation of Kaczmarz method

constant $s \in \mathbb{R}$. This coefficient s can be found by the inner product $\langle a, x \rangle$. This inner product is the projection length from x to a, but the scaled by $\|a\|$. So we need to divide $\|a\|$ here.

$$x^* = x - \frac{\langle a, x \rangle}{\|a\|} \frac{a}{\|a\|}$$

Now if $b \neq 0$, simply pick any $x_0 \in \mathcal{H}$, i.e. $\langle a, x_0 \rangle = b$. That means $\langle a, x - x_0 \rangle = 0$. So using change of variable $u = x - x_0$ (now $u^* = x^* - x_0$)

$$u^* = u - \frac{\langle a, u \rangle}{\|a\|^2} a$$

which is equivalent to

$$x^* = x - \frac{\langle a, x \rangle - b}{\|a\|^2} a$$

using this to build iterations:

$$x_{k+1} = x_k - \frac{\langle a, x_k \rangle - b}{\|a\|^2} a$$

will be the mathematical realisation of the algorithm shown in figure 1.

If you let i go through 1 to m, i.e. take rows of A one by one, this algorithm would work fine if m is small. But what if you are given a matrix of size $1e6 \times 1e5$ with many 0 entries?

A simple solution is to use the stochastic method, i.e. pick i randomly according to the uniform distribution on $\{i\}_{i=1,2,\ldots,m}$. Stochastic methods are beneficial for big data analysis as you can always randomly pick a data point or small data group for each iteration. This would improve efficiency and, surprisingly, usually converges to the desired solution!

Of course, you don't have to stick to the uniform distribution. In the case of the Kaczmarz method, weights can be given to each row according to $\|a_i\|^2$ as a larger norm means a more significant impact on the solution.

Similarly in gradient descent, if you have an objective function

$$\frac{1}{m} \sum_{i=1}^{m} Q_i(x)$$

where $1/m$ is to ensure $g(x) = E_i[Q_i(x)]$ if uniform distribution is used. If m is extremely large, you can randomly select i and let the descent direction be $-\nabla Q_i(x_k)$. Or you can select small group $K \subseteq \{1, \cdots m\}$ and let descent direction be $-\sum_{i\in K} \nabla Q_i(x_k)$.

4 Convexity

4.1 Visual Explanations of some Geometric Shapes

Hyperplane:

$$H := \{\boldsymbol{x} \in \mathbb{R}^n \,:\, \boldsymbol{a}^T\boldsymbol{x} = b\} \quad b \in \mathbb{R},\, \boldsymbol{a} \in \mathbb{R}^n$$

it is also mentioned in the Kaczmarz method that you can always pick any point x_0 on the hyperplane and transform the equation into $\boldsymbol{a}^T(\boldsymbol{x} - \boldsymbol{x}_0) = 0$, which means vector $\boldsymbol{a}$ is the normal direction to the plane.

Using any hyperplane, the space $\mathbb{R}^n$ can be divided into two half-spaces:

$$H^- := \{\boldsymbol{x} \in \mathbb{R}^n \,:\, \boldsymbol{a}^T\boldsymbol{x} \le b\}, \quad H^+ := \{\boldsymbol{x} \in \mathbb{R}^n \,:\, \boldsymbol{a}^T\boldsymbol{x} \ge b\}$$

It is easy to prove that H, H^-, H^+ are all convex.

Ellipsoid: Our most familiar form of the ellipsoid is

$$\frac{x_1^2}{a_1^2} + \cdots + \frac{x_n^2}{a_n^2} = 1$$

you can change any x_i^2 to $(x_i - c)^2$ to translate the ellipsoid. All ellipsoids can be written in the matrix form

$$\{\boldsymbol{x} \in \mathbb{R}^n \,:\, \boldsymbol{x}^T Q\boldsymbol{x} + 2\boldsymbol{b}^T\boldsymbol{x} + c \le 0\}$$

where Q must be positive semi-definite, because by Theorem 1.1, any symmetric matrix Q can be decomposed into UDU^T where D is diagonal. So $\boldsymbol{x}^T Q\boldsymbol{x} = \boldsymbol{x}^T UDU^T\boldsymbol{x} = \boldsymbol{y}^T D\boldsymbol{y}$ where $\boldsymbol{y} := U^T\boldsymbol{x}$. But the orthogonal matrix U^T is a rigid transformation, i.e. shape of objects will not be changed. So the shape of $\boldsymbol{x}^T Q\boldsymbol{x} - 1 = 0$ will be the same as $\boldsymbol{y}^T D\boldsymbol{y} - 1 = 0$ which can be written as $\sum_i d_i y_i^2 = 1$. Hence, if the object is ellipsoid, then $di \ge 0$, i.e. Q is positive semi-definite. The term $2\boldsymbol{b}^T\boldsymbol{x}$ will not affect the shape but will move the centre of the ellipsoid away from the origin.

4.2 Properties of Convex set

The intersection of arbitrary convex sets is convex by the nature of the definition. But the union of convex sets may not be convex. The easiest example is the union of two distinct lines.

The image and pre-image of linear transformation on a convex set are still convex, i.e. if C is convex, A is any matrix, the following two sets are convex:

$$A(C) := \{Ax + b \,:\, x \in C\}, \quad A^{-1}(C) := \{y \,:\, Ay + b \in C\}$$

And any linear combination and Cartesian product of convex sets are convex:

$$\mu_1 C_1 + \cdots + \mu_k C_k, \quad C_1 \times \cdots \times C_k$$

are convex.

The Interior and closure of a convex set are convex, as formulated in the following theorems

Theorem 4.1 (Preservation under Closure)**.** *If C is convex set, closure $\overline{C}$ is convex*

Proof. Easy proof using the sequential definition of closure. i.e. if $x \in \overline{C}$, exists $(x_k)_{k\ge1} \subseteq C$ s.t. $x_k \to x$ as $k \to \infty$. □

Theorem 4.2 (Preservation under Interior)**.** *If C is convex, then interior int(C) is convex.*

Proof. In the case, $\text{int}(C) = \emptyset$, the theorem is trivially true.
If $\text{int}(C) \neq \emptyset$, it is easier to prove another version: if $x \in \text{int}(C), y \in \overline{C}$, then $(1-\lambda)x+\lambda y \in \text{int}(C)$ for all $\lambda \in (0,1)$. This is called *line segment principle*, see proof in (Beck 2014, p.109). □

Using the line segment principle, it can be proved that for the convex set, $\overline{\text{int}(C)} = \overline{C}$ and $\text{int}(\overline{C}) = \text{int}(C)$.

Recall Carathéodory theorem says that elements in a convex hull in $\mathbb{R}^n$ can be written as a linear combination of at most $n+1$ vectors in S. The additional one dimension is used to amend for requirement on the coefficients: $\boldsymbol{\lambda} \in \Delta_{n+1}$ (this means the sum of entries of $\boldsymbol{\lambda}$ is 1). In fact, on the conic hull, where this requirement on coefficients is dropped and they can be any positive numbers, any element in the conic hull of S can be written as a linear combination of at most n linearly independent vectors x_i in S. This is called the *conic representation theorem.*

The convex hull of the closed set may not be closed, e.g.

$$S = \{(0,0)\} \cup \{(x,y) \,:\, xy \geq 1, x \geq 0, y \geq 0\}$$

S is closed, but its hull is not closed:

$$\text{conv}(S) = \{(0,0)\} \cup \{(x,y) \,:\, x > 0, y > 0\}$$

But compactness can be preserved

Theorem 4.3 (Preservation of compactness)**.** *If S is compact, then conv(S) is compact.*

Proof. .
Step 1. Boundedness
S is compact so there is $M > 0$ s.t. $\|x\| \leq M \forall x \in S$. We want to find a bound for $\text{conv}(S)$, and Carathéodory theorem becomes useful now: for all $y \in \text{conv}(S)$, exists $x_1, \cdots, x_{n+1} \in S$ and $\lambda \in \Delta_{n+1}$(i.e. $\sum_{i=1}^{n+1} \lambda_i = 1$) s.t. $y = \sum_{i=1}^{n+1} \lambda_i x_i$. Therefore,

$$\|y\| \leq \sum_{i=1}^{n+1} \lambda_i \|x_i\| \leq M \sum_{i=1}^{n+1} \lambda_i = M$$

it turns out that $\text{conv}(S)$ is bounded by the same bound of S.

Step 2. Closedness
Given sequence $(y^k)_{k\geq 1} \subseteq \text{conv}(S)$ converging to $y \in \mathbb{R}^n$, the target is to prove $y \in \text{conv}(S)$. Again by Carathéodory theorem, for any k, there are $x_1^k, \cdots x_{n+1}^k \in S$ and $\lambda^k \in \Delta_{n+1}$ s.t.

$$y^k = \sum_{i=1}^{n+1} \lambda_i^k x_i^k$$

S is compact so $\left((\lambda^k, x_1^k, \cdots x_{n+1}^k)\right)_{k\geq 1}$ have convergent subsequence, say the limit is $(\lambda, x_1, \cdots, x_{n+1})$. Taking limit of the equation $y^k = \sum_{i=1}^{n+1} \lambda_i^k x_i^k$ on the convergent subsequence (i.e. letting $k = k_j$ where k_j are indices of the subsequence, and send $j \to \infty$) yields

$$y = \sum_{i=1}^{n+1} \lambda_i x_i$$

□

4.3 Convex Functions

Convex functions are defined by the *Fundamental inequality*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y), \quad \forall x, y \in C, \lambda \in [0,1]$$

where C is a convex set. This can be visually understood in figure 2 taken from (page 117, Beck, 2014)

If the fundamental inequality becomes a strict inequality in the above definition, the function is called a *strictly convex* function.

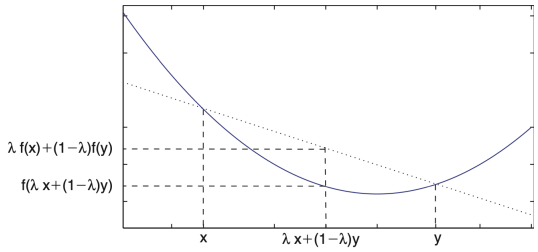

Figure 2: Graph representation of the fundamental inequality

An important property of convex functions, Jensen's inequality, basically says that this inequality generalises to any linear combination of vectors in C as long as coefficients add up to 1. The proof is done by induction on the number of vectors in linear combination and proceeds in a similar way to proving if C is convex, then any linear combination of m vectors in C is still in C if $\lambda \in \Delta_m$.

Exercise 1. *Using the convex function $f(x) = -\ln x$ defined on $\{x \in \mathbb{R} : x > 0\}$(it is convex because second derivative always positive) and Jensen's inequality, prove the AGM inequality*

$$\frac{\sum_i x_i}{n} \geq \sqrt[n]{\prod_i x_i}$$

4.4 First, second order characterisations of convex functions

In this section, equivalent conditions(characterisations) for the convexity of a continuously differentiable function are considered, based on the first and second derivatives. Continuously differentiability ensures the existence of derivatives.

Staring at figure 2, you may notice that if we draw the tangent line(or tangent hyperplane in higher dimension), then the tangent line is always below the function.

Theorem 4.4 (Gradient inequality)**.** *If f is continuously differentiable function, then f is convex over C*

$$\Leftrightarrow f(x) + \nabla f(x)^T (y - x) \leq f(y) \quad \forall x, y \in C$$

$f(x) + \nabla f(x)^T (y - x)$ is exactly the tangent hyperplane taken at $f(x)$.

There is another characterisation for convexity of continuously differentiable functions, from Figure 2, it can be seen that for convex functions on $\mathbb{R}$, the derivative is always non-decreasing. In a higher dimension, this characterisation is stated below

Theorem 4.5 (Monotonicity of gradient)**.** *If $f : C \to \mathbb{R}$ is continuously differentiable function defined on convex set $C \subseteq \mathbb{R}^n$, then f is convex iff*

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq 0 \quad \forall x, y \in C$$

Proof. ($\Rightarrow$) First assume f is convex, then by gradient inequality,

$$f(x) \geq f(y) + \nabla f(y)^T (x - y)$$

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

And adding two inequalities yields the required inequality.

($\Leftarrow$) Assume f satisfies the inequality (i.e. gradient is monotonic), we aim to prove the gradient inequality. Construct function $g(t) := f(x + t(y - x))$ which are the values of f on the line joining x, y. Then

$$f(y) = g(1) = g(0) + \int_0^1 g'(t)\, dt$$

then by chain rule, $g'(t) = (y-x)^T \nabla f(x + t(y-x))$, so

$$f(y) = f(x) + \int_0^1 (y-x)^T \nabla f(x + t(y-x))\, dt$$

$$= f(x) + \nabla f(x)^T (y-x) + \int_0^1 (y-x)^T \left(\nabla f(x + t(y-x)) - \nabla f(x)\right)\, dt$$

The integrand is positive as it can be rearranged in the following way

$$(y-x)^T \left(\nabla f(x + t(y-x)) - \nabla f(x)\right) = \frac{1}{t}(\nabla f(x + t(y-x)) - \nabla f(x))^T (x + t(y-x) - x) \geq 0$$

that yields

$$f(y) \geq f(x) + \nabla f(x)^T (y-x)$$

□

For continuously differentiable convex functions defined over a convex set, any stationary point is a global minimiser. (the example $f(x) = x^2$ illustrates this point)

Proposition 4.6 (Sufficient global optimality condition of convex function)**.** *If f defined on $C \subseteq \mathbb{R}^n$ is convex and continuously differentiable. Assume $\nabla f(x^*) = 0$, then x^* is global minimiser of f over C.*

Proof. Given $z \in C$, using the gradient inequality,

$$f(z) \geq f(x^*) + \nabla f(x^*)^T (z - x^*)$$

which implies $f(z) \geq f(x^*)$. So x^* must be a global minimiser. □

However, the condition above is necessary only when $C = \mathbb{R}^n$, the entire space. Because when C is not the entire space, the minimiser may appear at the boundaries.

Recall that in 1-D case, the gradient inequality (gradient is non-decreasing) is equivalent to $f''(x) > 0$ for all x. This is a more handy characterisation for checking whether a function is convex. There is a similar characterisation in higher dimension

Theorem 4.7 (Second-order characterisation)**.** *If $f : C \to \mathbb{R}$ ($C \subseteq \mathbb{R}^n$ is open convex) is twice continuously differentiable, then f convex $\Leftrightarrow \nabla^2 f(x) \succeq 0\ \forall x \in C$.*

Proof. Gradient inequality will be the bridge to prove this theorem.

Assume $\nabla^2 f(x) \succeq 0$, using linear approximation of f, given $x, y \in C$ there is $z \in [x, y] \cap C$ s.t.

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

the last term is non-negative as $\nabla^2 f(x) \succeq 0$, so gradient inequality is proved.

For the opposite direction, assuming f is convex, aim to show the positive semi-definiteness of Hessian. The openness of C is required, as we need to do quadratic approximation in a neighbourhood of x and this is only possible with C open. So given $x \in C$ and $y \in \mathbb{R}^n$, there is small enough ϵ s.t. $\forall \lambda \in (0.\epsilon)$, $x + \lambda y \in C$. (Treat y as a direction here, instead of a point) Then quadratic approximation gives

$$f(x + \lambda y) = f(x) + \nabla f(x)^T \lambda y + \frac{\lambda^2}{2} y^T \nabla f^2(x) y + o(\lambda^2 \|y\|^2)$$

Using gradient inequality,

$$f(x + \lambda y) \geq f(x) + \nabla f(x)^T \lambda y$$

so combining the above two results,

$$\frac{\lambda^2}{2} y^T \nabla f^2(x) y + o(\lambda^2 \|y\|^2) \geq 0$$

dividing by λ^2 and sending $\lambda \to 0^+$ yields

$$y^T \nabla^2 f(x) y \geq 0$$

y was chosen arbitrarily, so $\nabla^2 f(x) \succeq 0$. □

You can easily see that a small modification to the above proof gives

Theorem 4.8. *If $f : C \to \mathbb{R}$ ($C \subseteq \mathbb{R}^n$ is open convex) is twice continuously differentiable, then f strictly convex $\Leftrightarrow \nabla^2 f(x) \succ 0 \; \forall x \in C$.*

4.5 Properties of Convex Functions

It is straightforward to prove with definitions that linear combinations of convex functions are still convex as long as the coefficients are positive. And the linear transformation of independent variable $y := Ax + b$ will preserve convexity i.e. if f is convex over convex set C, then $g(x) := f(y) = f(Ax+b)$ is convex over $D := \{x : Ax+b \in C\}$. Note D is convex because it is pre-image of a convex set under linear transformation.

Using the above properties, it can be proved that

$$g(x) := \frac{\|A\boldsymbol{x} + \boldsymbol{b}\|^2}{\boldsymbol{c}^T\boldsymbol{x} + d}$$

is convex over $\{\boldsymbol{x} : \boldsymbol{c}^T\boldsymbol{x} + d > 0\}$ because $h(y,t) := \|\boldsymbol{y}\|^2/t$ is convex over $\{(y\ t)^T : t > 0\}$. g can be obtained from h by two consecutive linear transformations $\boldsymbol{y} \mapsto A\boldsymbol{x} + b$ and $t \mapsto \boldsymbol{c}^T\boldsymbol{x} + d$. As for the convexity of $h(y,t)$, note $h = \sum_i y_i^2/t$. The function $\phi(x,z) := x^2/z$ defined on $\{(x\ z)^T : x \in \mathbb{R}, z > 0\}$ is convex. (Left as an exercise)

Composition of convex functions may not be convex

Example 1. $g(x) = x^2$, $f(x) = x^2 - 4$ are both convex but $g(f(x)) = (x^2-4)^2$ is not convex. (check the second derivatives of these functions)

But if $g : I \to \mathbb{R}$ (where I is interval on $\mathbb{R}$ with $f(C) \subseteq I$) is non-decreasing convex function, $h(x) := g(f(x))$ is convex for any convex function $f : C \to \mathbb{R}$. (Proof left as an exercise) Note in fact, g only has to be non-decreasing on $f(C)$.

Using this composition rule, one can show that $e^{\|x\|^2}$ is convex.

.

Using the famous inequality $\max_i(a_i + b_i) \le \max_i a_i + \max_i b_i$, one can prove the following

Theorem 4.9 (Pointwise Maximum)**.** *If $f_i : C \to \mathbb{R}$ (for $i \in I$, arbitrary index set) are convex functions, then $f(x) := \max_i f_i(x)$ is convex.*

Warning: the function $h(\boldsymbol{x}) = x_{[i]}$ where $x_{[i]}$ is the i-th largest entry in $\boldsymbol{x}$ is NOT convex. But the following function is convex:

$$h_k(\boldsymbol{x}) := x_{[1]} + x_{[2]} + \cdots x_{[k]} := \max\{x_{i_1} + \cdots x_{i_k} : i_j \text{ are distinct indices}\}$$

Convexity is also preserved when some entries of input $\boldsymbol{x}$ are minimised over a convex set

Theorem 4.10 (Partial Minimisation)**.** *If $f : C \times D \to \mathbb{R}$ is convex function where C, D are convex,*

$$g(x) := \min_{y \in D} f(x,y)$$

is convex. Note min *should really be* inf *here, and we assume infimum exists and is finite.*

Proof. Given $x_1, x_2 \in C, \lambda \in [0,1]$. Minimum may not be attained by some y so we need to give a room of ϵ: given $\epsilon > 0$, there are $y_1, y_2 \in D$ s.t.

$$f(x_1, y_1) \le g(x_1) + \epsilon \quad f(x_2, y_2) \le g(x_2) + \epsilon$$

By convexity of f,

$$g(\lambda x_1 + (1-\lambda)x_2) \le f(\lambda x_1 + (1-\lambda)x_2, \lambda y_1 + (1-\lambda)y_2) \le \lambda f(x_1, y_1) + (1-\lambda) f(x_2, y_2)$$

$$\le \lambda(g(x_1) + \epsilon) + (1-\lambda)(g(x_2) + \epsilon) = \lambda g(x_1) + (1-\lambda) g(x_2) + \epsilon$$

The above inequality holds for all ϵ, so sending ϵ to 0 yields convexity of g. □

Remark. Note if f is convex,

$$g(x) := \max_{y \in D} f(x, y)$$

g will also be convex over C. Because we can treat y as index set and apply Theorem 4.9. Both partial and full maximisation of convex function is convex, but only partial minimisation(over convex set) of convex function is convex. This may not hold for arbitrary minimisation.

.

Continuity and differentiability:
Convex functions are continuous on open convex sets, but may not be continuous otherwise. However, convex functions are locally Lipschitz continuous for points in the interior of C. i.e. if $x_0 \in \text{int}(C)$, then exists $\epsilon > 0, L > 0$ with $B_\epsilon(x_0) \subseteq C$ and

$$|f(x) - f(x_0)| \leq L\|x - x_0\| \quad \forall x \in B_\epsilon(x_0)$$

Convex functions may not be differentiable, but all directional derivatives exist at interior points, which is enough for most optimisation problems.

See detailed, but rather long proofs of the above results in (Beck 2014, Section 7.6).

4.6 Level sets

Definition 3. Level set of f at level α is

$$\text{Lev}(f, \alpha) = \{x \,:\, f(x) \leq \alpha\}$$

Theorem 4.11. *If $f : C \to \mathbb{R}$ is convex function, for all $\alpha \in \mathbb{R}$, Lev(f, α) is convex set.*

Proof. Given $x, y \in \text{Lev}(f, \alpha)$, i.e. $f(x), f(y) \leq \alpha$. Using convexity,

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y) \leq \alpha$$

that means $\lambda x + (1 - \lambda)y \in \text{Lev}(f, \alpha)$, so $\text{Lev}(f, \alpha)$ is convex. □

There is a special class of functions

Definition 4 (Quasi-convex functions)**.** $f : C \to \mathbb{R}$ is quasi-convex if for all α, $\text{Lev}(f, \alpha)$ is convex set.

Convex functions are all quasi-convex, but there are non-convex functions in this class, e.g. $f(x) = \sqrt{|x|}$ and

$$f(x) = \frac{a^T x + b}{c^T x + d} \quad c \neq 0, \text{ defined on } \{x \,:\, c^T x + d > 0\}$$

4.7 Maxima of convex functions

We have seen that finding minima for convex functions is very convenient, as any stationary point is a global minimiser. Maxima are even easier to deal with, given the fact that they cannot exist in the interior:

Theorem 4.12. *If $f : C \to \mathbb{R}$ is non-constant convex function, f does not attain maximum in int(C)*

Proof. If $x^* \in \text{int}(C)$ is global maximiser, by non-constant property, there is $y \in C$ s.t. $f(y) < f(x^*)$. There is $\epsilon > 0$ s.t. $z := x^* + \epsilon(x^* - y) \in C$. Note $x^* = \epsilon/(\epsilon + 1)y + 1/(1 + \epsilon)z$, so by convexity,

$$f(x^*) \leq \frac{\epsilon}{\epsilon + 1} f(y) + \frac{1}{\epsilon + 1} f(z)$$

$$\Rightarrow f(z) \geq \epsilon(f(x^*) - f(y)) + f(x^*) > f(x^*)$$

Contradiction as x^* is a global maximiser. □

As you may have guessed, maximisers of convex functions will only appear on the boundary of convex sets, and surprisingly, one of them must be an extreme point.

Theorem 4.13. *If $f : C \to \mathbb{R}$ is convex, continuous and C is convex, compact set, then there is at least one maximiser that is an extreme point.*

Proof. By the Weierstrass theorem, there must be a maximiser x^* of f over C as C is compact. Assume x^* is not extreme point, by Krein-Milman, $C = \text{conv}(\text{ext}(C))$, by definition, there are $x_1, \cdots, x_k \in \text{ext}(C), \lambda \in \Delta_k$ s.t.

$$x^* = \sum_{i=1}^{k} \lambda_i x_i$$

So by the convexity of f,

$$f(x^*) \leq \sum_{i=1}^{k} \lambda_i f(x_i) \Rightarrow \sum_i \lambda_i (f(x_i) - f(x^*)) \geq 0$$

but x^* is maximiser, $f(x_i) \leq f(x^*)$, so the sum

$$\sum_i \lambda_i (f(x_i) - f(x^*)) \leq 0$$

that means the sum is indeed 0 and $f(x_i) = f(x^*)$. So x_i are all maximisers. □

The theorem also tells us that if there is only one maximiser, then it must be an extreme point.

Example 2. Consider optimisation problem

$$\max\{\|Ax\| : \|x\|_1 \leq 1\}$$

where A is a matrix. The function $x \mapsto \|Ax\|_1$ is convex as 1-norm and matrix-multiplication is convex. Further, $\{x : \|x\|_1 \leq 1\}$ is compact convex set. So there must be a maximiser among extreme points: $\pm e_i$ for $i = 1, \cdots, n$. Comparing the values at these extreme points, noting $\|Ae_j\| = \|A(-e_j)\|$, we have

$$\max\{\|Ax\| : \|x\|_1 \leq 1\} = \max_j \|Ae_j\|_1$$

which means the matrix norm $\|A\|_{1,1}$ is the maximum column sum of absolute values of the entries.

5 Convex Optimisation

Convex optimisation problems are problems of the form

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

where f, g_i are convex, h_j are affine functions.

Note h_j being convex is not enough for the set $\{x \in \mathbb{R}^n : h_j(x) = 0\}$ to be convex. e.g. imagine $h(x) = \|x\|_2^2 - 1$, then the set above is the unit circle, not convex. But if h is affine, the set above is convex. (Prove it)

The feasible set of this problem is

$$\left(\bigcap_{i=1}^{m} \text{Lev}(g_i, 0) \right) \bigcap \left(\bigcap_{j=1}^{p} \{x \in \mathbb{R}^n : h_j(x) = 0\} \right)$$

which is a convex set.

The local minimum point of convex functions on a convex set must be a global minimum, graphically this is due to the shape of convex functions: if x^* is minimum the shape of f around $x = x^*$ would look like a bowl. But the convex function has this shape everywhere, so $f(x^*)$ will be smaller than the value at any other point.

Theorem 5.1. *For convex function f, convex set C, the set of optimal solutions*

$$X^* := \min\{f(\boldsymbol{x}) \,:\, \boldsymbol{x} \in C\}$$

is convex and if f is strictly convex, there is at most one optimal solution.

Proof. The first part is easy. Now assume f is strictly convex and $x, y \in X^*$ $(x \neq y)$ and let $f^* := f(x) = f(y)$, then

$$f(\frac{1}{2}x + \frac{1}{2}y) < \frac{1}{2}f(x) + \frac{1}{2}f(y) = f^*$$

contradicts the minimality of f^*. □

Linear programming is a special case of convex optimisation where objective function $f(\boldsymbol{x})$ is linear, $g_i(\boldsymbol{x}) = \boldsymbol{a}_i^T\boldsymbol{x} - c_i$ and $h_j(\boldsymbol{x}) = \boldsymbol{b}_j^T\boldsymbol{x} - d_j$, so the constraints can be summarised to

$$A\boldsymbol{x} \leq \boldsymbol{c}, \quad B\boldsymbol{x} = \boldsymbol{d}$$

Linear functions are convex and concave, so theorems on convex minimisation problems apply to both minimisation and maximisation linear problems. (Maximisation of concave function f over convex set C is equivalent to minimisation of convex function $-f$ over C)

When the objective is a quadratic function, the problem is called a quadratic problem. And if the constraint is also quadratic, we call it QCQP(quadratically constrained quadratic problem). It is not necessarily convex unless all matrices are positive and semi-definite and all constraints are inequality.

5.1 Classification using linear separator

One should be able to convert a practical problem into a convex optimisation problem, e.g. suppose given two types of data $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m$(type 1) and $\boldsymbol{x}_{m+1}, \boldsymbol{x}_{m+2}, \cdots, \boldsymbol{x}_{m+p}$(type 2) and we wish to find a hyperplane $\{\boldsymbol{x} \,:\, \boldsymbol{w}^T\boldsymbol{x}+\beta = 0\}$ that separates the two sets of data. i.e.

$$\boldsymbol{w}^T\boldsymbol{x}_i + \beta < 0 \text{ for type 1}, \quad \boldsymbol{w}^T\boldsymbol{x}_i + \beta > 0 \text{ for type 2}$$

and that the hyperplane is kept as far from all points as possible, i.e. the problem is

$$\begin{aligned} \max \left\{ \min_{i=1,2,\cdots,m+p} \frac{|\boldsymbol{w}^T\boldsymbol{x}_i|+\beta}{\|\boldsymbol{w}\|} \right\} \\ \boldsymbol{w}^T\boldsymbol{x}_i + \beta < 0, \quad i = 1, \cdots, m \\ \boldsymbol{w}^T\boldsymbol{x}_i + \beta > 0, \quad i = m+1, \cdots, m+p \end{aligned}$$

the expression

$$\frac{|\boldsymbol{w}^T\boldsymbol{x}_i + \beta|}{\|\boldsymbol{w}\|}$$

represents the minimal distance from the hyperplane to point $\boldsymbol{x}_i$, so the first line means "maximising the minimum distance from $\boldsymbol{x}_i$ to the hyperplane". All optimal solutions can be scaled, i.e. if $\alpha \neq 0$, $(\alpha\boldsymbol{w}, \alpha\beta)$ is optimal if $(\boldsymbol{w}, \beta)$ is optimal.(because $\{\boldsymbol{x} \,:\, \alpha\boldsymbol{w}^T\boldsymbol{x} + \alpha\beta\} = 0$ still represents the same hyperplane) So the constraint

$$\min_i |\boldsymbol{w}^T\boldsymbol{x}_i + \beta| = 1$$

can be added to the problem without changing the result. But this new constraint is equivalent to

$$\boldsymbol{w}^T\boldsymbol{x}_i + \beta \leq -1, \quad i = 1, \cdots, m$$

$$\boldsymbol{w}^T\boldsymbol{x}_i + \beta \geq 1, \quad i = m+1, \cdots, m+p$$

and objective f becomes $1/\|\boldsymbol{w}\|$. Maximising this is equivalent to minimising $\frac{1}{2}\|\boldsymbol{w}\|^2$. Now the problem

$$\begin{aligned}\min\ & \frac{1}{2}\|\boldsymbol{w}\|^2 \\ & \boldsymbol{w}^T\boldsymbol{x}_i + \beta \le -1, \quad i = 1, \cdots, m \\ & \boldsymbol{w}^T\boldsymbol{x}_i + \beta \ge 1, \quad i = m+1, \cdots, m+p\end{aligned}$$

is a convex optimisation problem.

5.2 Stationary points

For continuously differentiable f defined over closed convex function C, $x^* \in C$ is stationary point if for all other $\boldsymbol{x} \in C$, the directional derivative along $x - x^*$

$$\nabla f(x^*)^T(x - x^*) \ge 0$$

that means along any direction from x^*, f has a positive gradient. Stationary points are important because if given f is continuously differentiable convex function, C is non-empty closed and convex set, for convex optimisation problem

$$\begin{aligned}\min\ & f(\boldsymbol{x}) \\ \text{s.t.}\ & \boldsymbol{x} \in C\end{aligned}$$

$\boldsymbol{x}^*$ is optimal solution iff $\boldsymbol{x}^*$ is stationary point.

As shown in lectures, when $C = \mathbb{R}^n$, $\boldsymbol{x}$ is stationary point iff $\nabla f(\boldsymbol{x}) = 0$. Let us consider three other cases

Example 3. Let $C := \{\boldsymbol{x} : x_i \ge 0\ \ \forall i\}$, stationary condition can be written as

$$\nabla f(\boldsymbol{x}^*)^T\boldsymbol{x} - (\nabla f(\boldsymbol{x}^*)^T\boldsymbol{x}^*) \ge 0 \quad \forall \boldsymbol{x} \ge 0$$

which is equivalent to the problem $\boldsymbol{a}^T\boldsymbol{x} + b \ge 0$ for all $\boldsymbol{x} \ge 0$ where $\boldsymbol{a} = \nabla f(\boldsymbol{x}^*)$, $b = -(\nabla f(\boldsymbol{x}^*)^T\boldsymbol{x}^*)$. Let $\boldsymbol{x} = 0$, we can show $b \ge 0$. Let $\boldsymbol{x} = (1 - \frac{b}{a_i})\boldsymbol{e}_i \ge 0$ (it is always positive as b, a_i has opposing signs), then

$$\boldsymbol{a}^T\boldsymbol{x} + b = a_i - b + b = a_i \ge 0$$

so $a_i \ge 0$ for all i. Therefore, the stationary condition is equivalent to

$$\nabla f(\boldsymbol{x}^*) \ge 0, \quad \nabla f(\boldsymbol{x}^*)^T\boldsymbol{x}^* \le 0$$

but $\boldsymbol{x}^* \ge 0$, so the latter condition is in fact equality and cleaning up, we have

$$\frac{\partial f}{\partial x_i}(\boldsymbol{x}^*) \begin{cases} = 0 & \text{if} x_i^* > 0 \\ \ge 0 & \text{if} x_i^* = 0 \end{cases}$$

Example 4. Let $C := \{\boldsymbol{x} : \boldsymbol{e}^T\boldsymbol{x} = \sum_i x_i = 1\}$(*unit-sum set*) where $\boldsymbol{e}$ is the vector with all entries being 1. Stationary condition on $\boldsymbol{x}^* \in C$ is equivalent to

$$\frac{\partial f}{\partial x_i}(\boldsymbol{x}^*) = \frac{\partial f}{\partial x_j}(\boldsymbol{x}^*) \quad \forall i, j$$

Proof: If the above hold, for all $\boldsymbol{x} \in C$,

$$\nabla f(\boldsymbol{x}^*)^T(\boldsymbol{x} - \boldsymbol{x}^*) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\boldsymbol{x}^*)(x_i - x_i^*) = \frac{\partial f}{\partial x_1}\left(\sum_i x_i - \sum_i x_i^*\right) = 0$$

Conversely, assume stationary condition holds but for some i, j

$$\frac{\partial f}{\partial x_i}(\boldsymbol{x}^*) > \frac{\partial f}{\partial x_j}(\boldsymbol{x}^*)$$

then define a new vector $\boldsymbol{x}$ as follows

$$x_i = x_i^* + 1,\ x_j = x_j^* - 1,\ x_k = x_k^* \text{ if } k \notin \{i, j\}$$

if $\boldsymbol{x}^* \in C$ then $\boldsymbol{x} \in C$. And

$$\begin{aligned}\nabla f(\boldsymbol{x}^*)^T(\boldsymbol{x} - \boldsymbol{x}^*) &= \frac{\partial f}{\partial x_i}(\boldsymbol{x}^*)(x_i - x_i^*) + \frac{\partial f}{\partial x_j}(\boldsymbol{x}^*)(x_j - x_j^*) \\ &= -\frac{\partial f}{\partial x_i}(\boldsymbol{x}^*) + \frac{\partial f}{\partial x_j}(\boldsymbol{x}^*) < 0\end{aligned}$$

this is a contradiction.

So stationary condition is that all partial derivatives are the same.

Example 5. $C = B(0, 1)$ the unit ball. Stationary condition is equivalent to

$$\nabla f(\boldsymbol{x}^*) = 0 \text{ or } \|\boldsymbol{x}^*\| = 1 \text{ and } \exists \lambda \leq 0 \text{ s.t. } \nabla f(\boldsymbol{x}^*) = \lambda \boldsymbol{x}^*$$

this means if the stationary point is in the interior of the unit ball, the gradient must be 0. But if it is on the boundary, the gradient should be 0 or point towards the centre of the unit ball. For example, consider optimisation of $f(\boldsymbol{x}) = (x_1 - 2)^2 + x_2^2 - 1$ over the unit ball, the minimum is at $(1, 0)$ (see Figure 3) with gradient $(-2, 0) = -2 * (1, 0)$

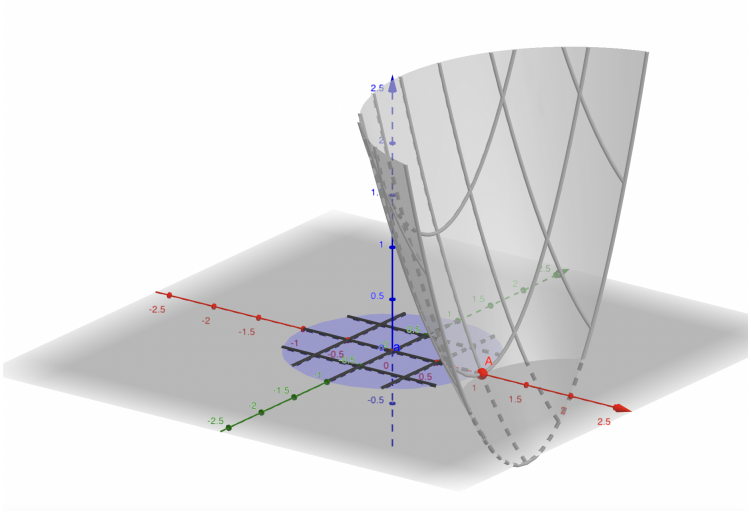

Figure 3: Optimisation of $(x_1 - 2)^2 + x_2^2 - 1$ over unit ball

5.3 Orthogonal Projection

Projection of $\boldsymbol{x}$ to convex set, i.e. the point in convex set C with closest distance to $\boldsymbol{x}$, is

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

and note this is a minimisation problem of convex quadratic function $\|\boldsymbol{y} - \boldsymbol{x}\|^2$ over convex set C. This is well-defined (first projection theorem), i.e. optimal point exists and is unique.

Proof.

$$f(\boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{x}\|^2 = \boldsymbol{y}^T I \boldsymbol{y} - 2\boldsymbol{x}^T \boldsymbol{y} + \|\boldsymbol{x}\|^2$$

it is quadratic which is coercive (because I is positive definite matrix), so at least one optimal solution exists. Further, f is strictly convex, so there is at most one optimal solution by Theorem 5.1. That means $P_c(\boldsymbol{x})$ is uniquely defined. □

Distance from $\boldsymbol{x}$ to convex set C is

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

Following are examples of orthogonal projection

Example 6. Let $C = \mathbb{R}^n_+ := \{\boldsymbol{x} \in \mathbb{R}^n \,:\, x_i \geq 0\}$. $P_C(\boldsymbol{x})$ is an optimisation problem where the object function $\sum_{i=1}^n (y_i - x_i)^2$ and constraints $\forall i,\, y_i \geq 0$ are separable. So consider the individual optimisation problems

$$\min\{(y_i - x_i)^2 \,:\, y_i \geq 0\}$$

The solution is

$$y_i^* = [x_i]_+ := \begin{cases} x_i & x_i \geq 0 \\ 0 & x_i < 0 \end{cases}$$

So $P_{\mathbb{R}^n_+}(\boldsymbol{x}) = [\boldsymbol{x}]_+$ where $[\boldsymbol{x}]_+ := ([x_1]_+, [x_2]_+, \cdots, [x_n]_+)^T$.

Example 7 (Box projection)**.** Let $C = [l_1, u_1] \times [l_2, u_2], \times \cdots \times [l_n, u_n]$ where $l_i \geq u_i$
Note the optimisation problem of $\text{Argmin}_y\{\|\boldsymbol{y} - \boldsymbol{x}\| : \boldsymbol{y} \in C\}$ is actually separable as long as constraint $\boldsymbol{y} \in C$. So in this case, like the previous example, consider the individual optimisation problems

$$\min\{(y_i - x_i)^2 \,:\, l_i \leq y_i \leq u_i\}$$

the optimal solution is

$$y_i = \begin{cases} u_i & x_i \geq u_i \\ x_i & l_i < x_i < u_i \\ l_i & x_i \leq l_i \end{cases}$$

Example 8 (Projection to balls)**.** Let $C = B(0, r)$. This time the optimisation problem $\text{Argmin}_y\{\|\boldsymbol{y} - \boldsymbol{x}\| : \boldsymbol{y} \in C\}$ is no longer separable. When $\|x\| \leq r$, $\boldsymbol{y} = \boldsymbol{x}$ is optimal solution. Otherwise, $\boldsymbol{y}$ must be on the boundary of the ball. Assume the contrary, it is a stationary point in the interior of the unit ball, i.e. $2(\boldsymbol{y} - \boldsymbol{x}) = 0 \Rightarrow \boldsymbol{y} = \boldsymbol{x}$. But $\boldsymbol{x} \notin C$, contradiction.

So the minimisation problem is equivalent to

$$\text{Argmin}_y\{-2\boldsymbol{x}^T\boldsymbol{y} + r^2 + \|\boldsymbol{x}\|^2 : \|\boldsymbol{y}\| = r\} = \text{Argmin}_y\{-2\boldsymbol{x}^T\boldsymbol{y} \,:\, \|\boldsymbol{y}\| = r\}$$

where the norm is squared, expanded using the inner product, and constant terms are removed. Using Cauchy-Schwarz,

$$-2\boldsymbol{x}^T\boldsymbol{y} \geq -2\|\boldsymbol{x}\|\|\boldsymbol{y}\| = -2r\|\boldsymbol{x}\|$$

and this bound is attained at $\boldsymbol{y} = r\boldsymbol{x}/\|\boldsymbol{x}\|$. So

$$P_{B(0,r)} = \begin{cases} \boldsymbol{x} & \|\boldsymbol{x}\| \leq r, \\ r\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} & \|\boldsymbol{x}\| > r \end{cases}$$

Theorem 5.2 (Second Projection Theorem)**.** *Given convex set C, any point $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{y} \in C$, the angle between $\boldsymbol{x} - P_C(\boldsymbol{x}), \boldsymbol{y} - P_C(\boldsymbol{x})$ is greater than or equal to 90 degrees. (Try to draw it to see why this is true!) i.e.*

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

Proof. Use the fact that $P_C(\boldsymbol{x})$ is solution to optimisation

$$\begin{aligned} \min \quad & g(\boldsymbol{y}) := \|y - x\|^2 \\ \text{s.t.} \quad & \boldsymbol{y} \in C \end{aligned}$$

which means $P_C(\boldsymbol{x})$ is a stationary point, i.e.

$$\nabla g(P_C(\boldsymbol{x}))(\boldsymbol{y} - P_C(\boldsymbol{x})) \geq 0 \quad \forall \boldsymbol{y} \in C$$

which is equivalent to the required inequality. □

Orthogonal projection gives an equivalent condition to stationarity

Theorem 5.3 (Representation of Stationarity)**.** *For any continuously differentiable f, closed convex set C, $s > 0$, $\boldsymbol{x}^*$ is stationary iff*

$$\boldsymbol{x}^* = P_C(\boldsymbol{x}^* - s\nabla f(x^*))$$

note $\boldsymbol{x}^* - s\nabla f(x^*)$ is a step in gradient descent with step size s. So the equation means the projection of the point after gradient descent back to the convex set is $\boldsymbol{x}^*$ itself.

Proof. Use second projection theorem on the pair $\boldsymbol{x}^* - s\nabla f(x^*), \boldsymbol{x}$ where $\boldsymbol{x} \in C$. The second theorem is useful for proving properties with inequality. □

5.4 Gradient Projection

Theorem 5.3 gives an iterative formulae that numerically finds minimal value of f over C, i.e.

$$x_{k+1} = P_C(x_k - t_k \nabla f(x_k))$$

Similar to gradient descent, there are many strategies to choose the step sizes t_k. But before that, we need to prove gradient projection method has sufficient decrease (find a lower bound for $f(\boldsymbol{x}) - f(P_C(\boldsymbol{x} - t\nabla f(\boldsymbol{x})))$)

Lemma 5.4 (Sufficient Decrease of gradient projection)**.** *Given $f \in C^{1,1}(C)$ with Lipschitz constant L where C is closed, convex. For any $\boldsymbol{x} \in C, t \in (0, 2/L)$,*

$$f(\boldsymbol{x}) - f(P_C(\boldsymbol{x} - t\nabla f(\boldsymbol{x}))) \geq t\left(1 - \frac{Lt}{2}\right)\left\|\frac{1}{t}\left(\boldsymbol{x} - P_C(\boldsymbol{x} - t\nabla f(\boldsymbol{x}))\right)\right\|^2$$

This is very similar to the sufficient decrease lemma for gradient descent (lemma 3.10)

Proof. Let $\boldsymbol{x}^+ := P_C(\boldsymbol{x} - t\nabla f(\boldsymbol{x}))$, using descent lemma (lemma 3.9),

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

using the second projection theorem,

$$(\boldsymbol{x} - t\nabla f(\boldsymbol{x}) - \boldsymbol{x}^+)^T(\boldsymbol{x} - \boldsymbol{x}^+) \leq 0$$

so

$$\nabla f(\boldsymbol{x})^T(\boldsymbol{x}^+ - \boldsymbol{x}) \leq -\frac{1}{t}\|\boldsymbol{x}^+ - \boldsymbol{x}\|^2$$

plugging this to the first inequality,

$$f(\boldsymbol{x}^+) \leq f(\boldsymbol{x}) + (\frac{L}{2} - \frac{1}{t})\|\boldsymbol{x}^+ - \boldsymbol{x}\|^2$$

$$\Rightarrow f(\boldsymbol{x}) - f(\boldsymbol{x}^+) \geq \frac{1}{t}(1 - \frac{Lt}{2})\|\boldsymbol{x}^+ - \boldsymbol{x}\|^2$$

which is the same as required inequality. □

So an ideal choice for constant step size is $1/L$ where L is the Lipschitz constant of objective f. To simplify notations, let

$$G_L(\boldsymbol{x}) := L\left[\boldsymbol{x} - P_C\left(\boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x})\right)\right] \quad L > 0$$

note if $C = \mathbb{R}^n$, $G_L(\boldsymbol{x}) = \nabla f(\boldsymbol{x})$. And by representation theorem of stationarity, $G_M(\boldsymbol{x}) = 0 \Leftrightarrow \boldsymbol{x}$ is a stationary point. The sufficient decrease now can be restated as

$$f(\boldsymbol{x}) - f(P_C(\boldsymbol{x} - t\nabla f(\boldsymbol{x}))) \geq t\left(1 - \frac{Lt}{2}\right)\|G_{1/t}(\boldsymbol{x})\|^2$$

This new general version of the gradient function has the following property

Lemma 5.5 (Monotonicity)**.** *If f is continuously differentiable function defined on closed and convex set C, and $L_1 \geq L_2$, then*

$$\|G_{L_1}(\boldsymbol{x})\| \geq \|G_{L_2}(\boldsymbol{x})\|$$

$$\frac{\|G_{L_1}(\boldsymbol{x})\|}{L_1} \leq \frac{\|G_{L_2}(\boldsymbol{x})\|}{L_2}$$

5.4.1 Backtracking

Similar to gradient descent, we need to choose initial step size $s > 0$, tracking rates $\alpha, \beta \in (0, 1)$. At iteration k of gradient projection, let $t_k := s$. Continue update by $t_k = \beta t_k$ if

$$f(\boldsymbol{x}_k) - f(P_C(\boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k))) < \alpha t_k \|G_{1/t_k}(\boldsymbol{x}_k)\|^2$$

process stops when

$$f(\boldsymbol{x}_k) - f(P_C(\boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k))) \geq \alpha t_k \|G_{1/t_k}(\boldsymbol{x}_k)\|^2 \qquad (*)$$

again this is very similar to the formulae for gradient descent.

If $f \in C^{1,1}(C)$, backtracking eventually stops. Using sufficient decrease lemma for gradient projection,

$$f(\boldsymbol{x}_k) - f(P_C(\boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k))) \geq \left(1 - \frac{L t_k}{2}\right) t_k \|G_{1/t_k}(\boldsymbol{x}_k)\|^2$$

when $1 - Lt_k/2 \geq \alpha$, the stopping inequality (*) is satisfied. This is equivalent to $t_k \leq 2(1-\alpha)/L$. So after t_k is shrunk below this boundary, backtracking stops.

Lower bound on t: t_k is either s, or backtracking invoked. If backtracking is invoked, the step size in the last iteration of backtracking $t_k/\beta > 2(1-\alpha)/L$, i.e. $t_k > 2(1-\alpha)\beta/L$, so

$$t_k \geq \min\left\{s, \frac{2(1-\alpha)\beta}{L}\right\}$$

Lemma 5.6 (Sufficient decrease of backtracking)**.** *If $(\boldsymbol{x}_k)_{k\geq 0}$ is the sequence of $\boldsymbol{x}$ generated by gradient projection,*

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq M \|G_d(\boldsymbol{x}_k)\|^2$$

where

$$M = \alpha \min\left\{s, \frac{2(1-\alpha)\beta}{L}\right\}, \quad d = \frac{1}{s}$$

Proof. By definition of backtracking,

$$f(\boldsymbol{x}_k) - f(P_C(\boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k))) \geq \alpha t_k \|G_{1/t_k}(\boldsymbol{x}_k)\|^2$$

so the value of M follows from lower bound of t_k derived above. Note $t_k \leq s$, so by lemma 5.5, $\|G_{1/t_k}(\boldsymbol{x}_k)\| \geq \|G_{1/s}(\boldsymbol{x}_k)\|$ which yields the result above. □

Finally, for convergence of the gradient projection method, we need to verify $f(\boldsymbol{x}_k)$ converges.

Theorem 5.7 (Convergence of gradient projection)**.** *Assume $f \in \mathbb{C}^{1,1}(C)$ is bounded below, $\boldsymbol{x}_k$ is sequence generated by gradient projection, then*
(1) $f(\boldsymbol{x}_k)$ converges
(2) $G_d(\boldsymbol{x}_k) \to 0$, i.e. the lower bound in sufficient decrease lemma does not block the convergence of $\boldsymbol{x}_k$.

Proof.

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq M \|G_d(\boldsymbol{x}_k)\|^2 \qquad (*)$$

for some $M > 0$ by the previous lemma and sufficient decrease lemma. So directly we have $f(\boldsymbol{x}_k) \geq f(\boldsymbol{x}_{k+1})$ with equality holds iff $\boldsymbol{x}_k$ is stationary point.

$f(\boldsymbol{x}_k)$ must converge as it is bounded below and non-increasing. So $f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \to 0$, combined with the above inequality (*) gives $\|G_d(\boldsymbol{x}_k)\| \to 0$. □

5.5 Sparsity constraint

Sparsity in data means that most entries are 0 while only few important entries are kept. Given continuously differentiable f which is bounded below, sparsity of optimal solution $\boldsymbol{x}$ is realised by the following NON-convex optimisation problem (called *Sparsity constrained problem*)

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & \|\boldsymbol{x}\|_0 \leq s \quad s \in \mathbb{N}(s \neq 0) \end{aligned}$$

where the 0-norm (not actually a norm) counts number of non-zero components

$$\|\boldsymbol{x}\|_0 := |\{i \,:\, x_i \neq 0\}|$$

Since this problem is non-convex, theorems derived in previous sections do not apply.

Define the set $C_s := \{\boldsymbol{x} \,:\, \|\boldsymbol{x}\|_0 \leq s\}$.
Motivated by Theorem 5.3, we say $\boldsymbol{x}^*$ is L-stationary point if

$$\boldsymbol{x}^* \in P_{C_s}\left(\boldsymbol{x}^* - \frac{1}{L}\nabla f(\boldsymbol{x}^*)\right)$$

Equality is not used, because for non-convex set C_s, projection to C_s is not well-defined, i.e. it could give several points. For example,

$$P_{C_2}((3,1,1)) = \{(3,1,0),(3,0,1)\}$$

this projection sends the smallest $n-s$ entries to 0.
This stationarity condition depends on L, smaller L makes it stricter. A natural question that arises is for what L, L-stationarity is related to optimality.

Given $f \in C^{1,1}_{L_f}(\mathbb{R}^n)$, assume $L > L_f$, the L-stationarity condition

$$\boldsymbol{x}^* \in P_{C_s}\left(\boldsymbol{x}^* - \frac{1}{L}\nabla f(\boldsymbol{x}^*)\right)$$

is equivalent to

$$\boldsymbol{x}^* \in \text{Argmin}_{\boldsymbol{y}\in C_s}\left\|\boldsymbol{y} - \left(\boldsymbol{x}^* - \frac{1}{L}\nabla f(\boldsymbol{x}^*)\right)\right\|^2$$

Recall in descent lemma for Lipschitz continuous function,

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}^*) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x}^*) + \frac{L}{2}\|\boldsymbol{x}^*-\boldsymbol{y}\|^2 =: h_L(\boldsymbol{y},\boldsymbol{x}^*)$$

rearranging $h_L(\boldsymbol{y},\boldsymbol{x}^*)$ yields

$$= \frac{L}{2}\left\|\boldsymbol{y} - \left(\boldsymbol{x}^* - \frac{1}{L}\nabla f(\boldsymbol{x}^*)\right)\right\|^2 + f(\boldsymbol{x}^*) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}^*)\|^2$$

where the last two terms are constant w.r.t. $\boldsymbol{y}$. So the optimisation problem is equivalent to

$$\boldsymbol{x}^* \in \text{Argmin}_{\boldsymbol{y}\in C_s} h_L(\boldsymbol{y},\boldsymbol{x}^*)$$

It can be proved that if $L > L_f$(the Lipschitz constant of f), then L-stationarity is the necessary condition for $\boldsymbol{x}^*$ to be optimal solution of the sparsity constrained problem.

A generalisation of gradient projection algorithm known as *iterative hard-threshold*(IHT) method, begins with constant $L > L_f$ and $\boldsymbol{x}_0 \in C_s$. Then the iterative step is given by

$$\boldsymbol{x}^{k+1} \in P_{C_s}\left(\boldsymbol{x}^k - \frac{1}{L}\nabla f(\boldsymbol{x}^k)\right)$$

and this is equivalent to

$$\boldsymbol{x}^{k+1} \in \text{Argmin}_{\boldsymbol{x}\in C_s} h_L(\boldsymbol{x},\boldsymbol{x}^k)$$

as derived above. It has been proved in (Sun and Cheng, 2016) that IHT converges.

6 KKT conditions

Optimality condition we studied:

$$\nabla f(x^*)^T(x - x^*) \geq 0 \text{ for all } \boldsymbol{x} \in C$$

may not be easy to verify. But alternative formulation KKT can be used, which is easier to solve. We will study a specific family of problems: linearly constrained problems (LCP) first.

LCP takes the form

$$\begin{aligned} (LCP)\min \quad & f(x) \\ \text{s.t.} \quad & \boldsymbol{a}_i^T \boldsymbol{x} \leq b_i \text{ for } i = 1, 2, \cdots, m \end{aligned}$$

with $\boldsymbol{a}_i \in \mathbb{R}^n,\ b_i \in \mathbb{R}$. We always assume f is continuously differentiable.

The KKT condition is like a generalisation of Lagrange multiplier,

Theorem 6.1 (KKT condition for LCP(necessary condition))**.** *If $\boldsymbol{x}^*$ is local minimum point of (LCP), exists $\lambda_1, \cdots, \lambda_m \geq 0$ s.t.*

$$\nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \lambda_i \boldsymbol{a}_i = 0 \qquad \textit{(KKT1)}$$

and

$$\lambda_i(\boldsymbol{a}_i^T \boldsymbol{x}^* - b_i) = 0 \qquad \textit{(KKT2)}$$

Remark. (KKT1) is a modification to $\nabla f(\boldsymbol{x}^*) = 0$ by adding terms $\lambda_i \nabla c_i(\boldsymbol{x}^*)$ where $c_i(\boldsymbol{x}) = \boldsymbol{a}_i^T \boldsymbol{x} - b_i$ are constraints, $\lambda_i \geq 0$ ensures that the direction of inequality $c_i(\boldsymbol{x}) \leq 0$ is not changed. Later you will see that multipliers of constraints like $h_i(\boldsymbol{x}) = 0$ do not have to be non-negative. Set of m equations in (KKT2), also called *complementary slackness conditions*, are satisfied iff $\lambda_i = 0$ or $\boldsymbol{a}_i^T \boldsymbol{x}^* = b_i$. Recall the examples given in the last chapter for optimality conditions on unit-sum set or unit ball, either some(or all) partial derivative(s) to be 0, or $\boldsymbol{x}^*$ sits at the boundary. In this case, the boundary is $\{\boldsymbol{x} : \boldsymbol{a}_i^T \boldsymbol{x} = b_i\}$, and we will call the set $\{i : \boldsymbol{a}_i^T \boldsymbol{x} = b_i\} =: I(\boldsymbol{x})$ the set of active constraints.

If given f is convex, then the KKT condition becomes sufficient. In practice, even if f is not convex, you can solve the equations given by KKT conditions, find all possible candidates of $\boldsymbol{x}^*$, and pick the one giving smallest $f(\boldsymbol{x})$ value.

KKT condition for LCP is built on Farkas' lemma, which requires the separation theorem to prove. Separation theorem simply means there is always a hyperplane separating a point from a closed, convex set.

Theorem 6.2. *Given non-empty, closed, convex C and $\boldsymbol{y} \notin C$, exists hyperplane*

$$\{\boldsymbol{x} : \boldsymbol{p}^T \boldsymbol{x} = \alpha\} \qquad \boldsymbol{p} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\},\ \textit{and } \alpha \in \mathbb{R}$$

that separates $\boldsymbol{y}$ and C, i.e.

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

Proof. And let $\boldsymbol{p}$ be $\boldsymbol{y} - P_C(\boldsymbol{y})$, the vector connecting $\boldsymbol{y}$ to C which is orthogonal to C, and $\alpha := (\boldsymbol{y} - P_C(\boldsymbol{y}))^T P_C(\boldsymbol{y})$. The choice of α ensures $\boldsymbol{p}^T P_C(\boldsymbol{y}) = \alpha$. Then second orthogonal projection theorem ensures $\boldsymbol{p}^T \boldsymbol{x} \leq \alpha$ for all other $\boldsymbol{x} \in C$. Closeness of C is to ensure $P_C(\boldsymbol{y}) \in C$. See detailed proof in lecture notes(page 94). □

Farkas' lemma is stated as below

Lemma 6.3. *Given $A \in \mathbb{R}^{m \times n},\ \boldsymbol{c} \in \mathbb{R}^n$, exactly one of the following hold*
I. exists $\boldsymbol{y} \geq 0$ s.t. $A^T \boldsymbol{y} = \boldsymbol{c}$
II. there exists $\boldsymbol{x}$ s.t. $A\boldsymbol{x} \leq \boldsymbol{0},\ \boldsymbol{c}^T \boldsymbol{x} > 0$.

It means that given convex cone $C_0 := \{A^T\boldsymbol{y} : \boldsymbol{y} \in \mathbb{R}^n_+\} = \{\sum y_i \boldsymbol{r}_i^T\}$ where $\boldsymbol{r}_i$ are rows of A (i.e. set of linear combinations of rows of A with non-negative coefficients). For any vector $\boldsymbol{c}$, if it is not in C_0 (case I not satisfied), then you can find hyperplane $\{\boldsymbol{y} : \boldsymbol{x}^T\boldsymbol{y} = 0\}$ s.t. separates cone C_0 and $\boldsymbol{c}$(condition II). To see this,

$$A\boldsymbol{x} \leq \boldsymbol{0} \Leftrightarrow \begin{pmatrix} \boldsymbol{r}_1^T\boldsymbol{x} \\ \vdots \\ \boldsymbol{r}_m^T\boldsymbol{x} \end{pmatrix} \leq \boldsymbol{0} \Leftrightarrow \boldsymbol{r}_i^T\boldsymbol{x} < 0 \Rightarrow \boldsymbol{y}^T\boldsymbol{x} < 0 \, \forall \boldsymbol{y} \in C_0$$

Farkas' lemma is formulated in the way "either A or B", which can be treated as B iff not A.

Lemma 6.4 (Farkas' Lemma)**.** *Given $A \in \mathbb{R}^{m\times n}$, $\boldsymbol{c} \in \mathbb{R}^n$, the following are equivalent:*
(A) $A\boldsymbol{x} \leq \boldsymbol{0} \Rightarrow \boldsymbol{c}^T\boldsymbol{x} \leq 0$
(B) exists $\boldsymbol{y} \in \mathbb{R}^m_+$ s.t. $A^T\boldsymbol{y} = \boldsymbol{c}$

Proof. (B) $\Rightarrow$ (A) is easy because you are using an inequality $A\boldsymbol{x} \leq \boldsymbol{0}$ and an equality $A^T\boldsymbol{y} = \boldsymbol{c}$ to derive an inequality, a simple substitution will deal with this case. (see detailed proof on page 95 of lecture notes)

The other direction is easier with contradiction. Assume (A) is true, (B) is false, aim to find $\boldsymbol{x}$ s.t. $(A\boldsymbol{x})^T = \boldsymbol{x}^TA^T \leq 0$ but $\boldsymbol{c}^T\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{c} > 0$. This is exactly a separation problem using hyperplane $\{\boldsymbol{y} : \boldsymbol{x}^T\boldsymbol{y} = 0\}$ where $\boldsymbol{x}$ is the vector to be found. So we need to justify the cone

$$C_0 := \{A^T\boldsymbol{y} : \boldsymbol{y} \in \mathbb{R}^n_+\}$$

is convex, closed, non-empty. It is closed because it is a conic hull of a finite set of vectors $\{\boldsymbol{r}_i\}_{i=1,\cdots,m}$ (search for proof this result if you are interested). C_0 is clearly convex, non-empty, so separation theorem applies. □

A different theorem is required for KKT condition of general optimisation problem,

Theorem 6.5 (Gordan's alternative theorem)**.** *Given $A \in \mathbb{R}^{m\times n}$, exactly one of the following hold:*
(A) $\exists \boldsymbol{x}$ s.t. $A\boldsymbol{x} < 0$
(B) $\exists \boldsymbol{p} \geq 0 (\boldsymbol{p} \neq 0)$ s.t. $A^T\boldsymbol{p} = 0$

Proof. If (A) holds, and for contradiction assume (B) also holds, then

$$\boldsymbol{x}^TA^T\boldsymbol{p} = 0 \Rightarrow (A\boldsymbol{x})^T\boldsymbol{p} = 0$$

but $A\boldsymbol{x} < 0, \boldsymbol{p} \geq 0$ and $\boldsymbol{p} \neq 0$, so this is impossible. (if you are not convinced, write it element-wise)

If (A) fails, note (A) is equivalent to

$$A\boldsymbol{x} + s\boldsymbol{e} \leq 0, \quad s > 0$$

for small enough $s > 0$, and $\boldsymbol{e} = (1, 1, \cdots, 1)^T$. Defining $\boldsymbol{c} := e_{n+1}$ Writing these in matrix form:

$$\underbrace{\begin{pmatrix} A & \boldsymbol{e} \end{pmatrix}}_{=:\tilde{A}} \underbrace{\begin{pmatrix} \boldsymbol{x} \\ s \end{pmatrix}}_{=:\boldsymbol{w}} \leq \boldsymbol{0}, \quad \boldsymbol{c}^T \begin{pmatrix} \boldsymbol{x} & s \end{pmatrix} > 0$$

So if (A) does not have solution,

$$\tilde{A}\boldsymbol{w} \leq \boldsymbol{0}, \ \boldsymbol{c}^T\boldsymbol{w} > 0$$

has no solution for $\boldsymbol{w} \in \mathbb{R}^{n+1}$, so by Farkas' lemma, exists $\boldsymbol{z} \in \mathbb{R}^m_+$ s.t.

$$\tilde{A}^T\boldsymbol{z} = \begin{pmatrix} A^T \\ \boldsymbol{e}^T \end{pmatrix} \boldsymbol{z} = \boldsymbol{c}$$

$$\Rightarrow A^T\boldsymbol{z} = 0, \ \boldsymbol{e}^T\boldsymbol{z} = 1$$

clearly $\boldsymbol{z} \neq 0$, so $\boldsymbol{z}$ satisfies condition (B). □

Now we are ready to prove the KKT condition, Theorem 6.1.

Proof. If $\boldsymbol{x}^*$ is local minimum, then $\boldsymbol{x}^*$ is stationary point. Let $\boldsymbol{y} := \boldsymbol{x} - \boldsymbol{x}^*$, stationarity condition becomes $\nabla(\boldsymbol{x}^*)^T\boldsymbol{y} \geq 0$ for any $\boldsymbol{y}$ satisfying the constraint. (i.e. $\boldsymbol{a}_i^T(\boldsymbol{y} + \boldsymbol{x}^*) \leq b_i$) Recall we defined the set of active constraints $I(\boldsymbol{x})$ (i.e. indices of constrains i s.t. $\boldsymbol{x}$ falls on the boundary) by $I(\boldsymbol{x}) := \{i : \boldsymbol{a}_i^T\boldsymbol{x} = b_i\}$, using these we can partition the inequalities into two categories. Note the constraints are equivalent to $\boldsymbol{a}_i^T\boldsymbol{y} \leq b_i - \boldsymbol{a}_i^T\boldsymbol{x}^*$, so

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

constraints not in $I(\boldsymbol{x}^*)$ can be removed: assume $\boldsymbol{a}_i^T\boldsymbol{y} \leq 0$ for $i \in I(\boldsymbol{x}^*)$. For $i \notin I(\boldsymbol{x}^*)$, $b_i - \boldsymbol{a}_i^T\boldsymbol{x}^* > 0$, so exists small enough $\alpha > 0$ s.t. $\boldsymbol{a}_i^T(\alpha\boldsymbol{y}) \leq b_i - \boldsymbol{a}_i^T\boldsymbol{x}^*$. Together with the fact that $\boldsymbol{a}_i^T(\alpha\boldsymbol{y}) \leq 0$ for $i \in I(\boldsymbol{x}^*)$(because $\boldsymbol{a}_i^T\boldsymbol{y} \leq 0$), stationary condition for $\nabla f(\boldsymbol{x}^*)^T(\alpha\boldsymbol{y}) \geq 0$ is satisfied. That yields $\nabla f(\boldsymbol{x}^*)^T\boldsymbol{y} \geq 0$. So indeed the condition on $i \notin I(\boldsymbol{x}^*)$ is not required, i.e.

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

You may recognise this as condition (A) of Farkas' lemma (Lemma 6.4, correspondence: $A = (\boldsymbol{a}_i^T)_{i\in I(\boldsymbol{x}^*)}$, $\boldsymbol{c} = -\nabla f(\boldsymbol{x}^*)$ where negative sign ensures inequality is in the direction ≤ 0). So we can find $\lambda_i \geq 0$ (corresponds to $\boldsymbol{y}_i$ in condition (B) of Farkas' lemma) s.t.

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

For $i \notin I(\boldsymbol{x}^*)$, let $\lambda_i = 0$. Now we have $\lambda_i(\boldsymbol{a}_i^T\boldsymbol{x}^* - b_i) = 0$ for all i and

$$\nabla f(\boldsymbol{x}^*) + \sum_{i=1}^m \lambda_i \boldsymbol{a}_i = 0$$

Further Assumption If f is convex, we want to prove $\boldsymbol{x}^*$ is optimal solution iff $\boldsymbol{x}^*$ satisfies KKT condition. Only have to prove sufficiency (necessity is proved above, because $\boldsymbol{x}^*$ is optimal solution means it is local minimum point) Suppose $\boldsymbol{x}^*$ satisfies KKT conditions, define

$$L(\boldsymbol{x}) := f(\boldsymbol{x}) + \sum_{i=1}^m \lambda_i(\boldsymbol{a}_i^T\boldsymbol{x} - b_i)$$

(KKT1) means $\nabla L(\boldsymbol{x}) = \boldsymbol{0}$. L is convex, so gradient being 0 is sufficient for $\boldsymbol{x}^*$ to be global minimiser of L. Combine with (KKT2), for any $\boldsymbol{x}$ satisfying the constraints $\boldsymbol{a}_i^T\boldsymbol{x} \leq b_i$,

$$f(\boldsymbol{x}^*) + 0 = f(\boldsymbol{x}^*) + \sum_{i=1}^m \lambda_i(\boldsymbol{a}_i^T\boldsymbol{x}^* - b_i) = L(\boldsymbol{x}^*) \leq L(\boldsymbol{x}) = f(\boldsymbol{x}) + \sum_{i=1}^m \lambda_i(\boldsymbol{a}_i^T\boldsymbol{x} - b_i) \leq f(\boldsymbol{x})$$

where last inequality follows by constraints. So $\boldsymbol{x}^*$ is also global optimal point for minimising f with constraints $\boldsymbol{a}_i^T\boldsymbol{x} \leq b_i$. □

Further constraints $\boldsymbol{a}^T\boldsymbol{x} = b$ can be added to LCP, and observe that it is equivalent to $\boldsymbol{a}^T\boldsymbol{x} \leq b$, $-\boldsymbol{a}^T\boldsymbol{x} \leq -b$. So KKT conditions can also be imposed.

Theorem 6.6 (KKT condition for LCP with equalities)**.** *Given the following optimisation problem*

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

where f is continuously differentiable, $\boldsymbol{a}_i, \boldsymbol{c}_j \in \mathbb{R}^n$, $b_i, d_j \in \mathbb{R}$.

(a) If $\boldsymbol{x}^*$ is local minimum point, exists $\lambda_i \geq 0$, $\mu_j \in \mathbb{R}$ s.t.

$$\nabla f(\boldsymbol{x}^*) + \sum_{i=1}^m \lambda_i \boldsymbol{a}_i + \sum_{j=1}^p \mu_j \boldsymbol{c}_j = 0 \qquad \textit{(KKT1)}$$

and

$$\lambda_i(\boldsymbol{a}_i^T\boldsymbol{x}^* - b_i) = 0 \qquad \textit{(KKT2)}$$

note actually there are other p hidden equations: the constraints!

(b) If f is convex and $\boldsymbol{x}^*$ is feasible solution (i.e. satisfies the constraints), if there are $\lambda_i \geq 0,\ \mu_j \in \mathbb{R}$ satisfying (KKT1), (KKT2), then $\boldsymbol{x}^*$ is optimal solution.

Proof. The proof basically uses previous KKT theorem, in the spirit of

$$\boldsymbol{a}^T\boldsymbol{x} \leq b,\ -\boldsymbol{a}^T\boldsymbol{x} \leq -b \Leftrightarrow \boldsymbol{a}^T\boldsymbol{x} = b$$

though a small trick is required: let $\mu^+ := \max\{\mu, 0\} \geq 0,\ \mu^- := \max\{-\mu, 0\} \geq 0$, then $\mu = \mu^+ - \mu^-$.
Proof left as exercise. □

By treating orthogonal projections as optimisation problems, one can find the formulae for projections to affine space $C = \{\boldsymbol{x} : A\boldsymbol{x} = \boldsymbol{b}\}$ (assume rows of A are independent), hyperplane $H := \{\boldsymbol{x} : \boldsymbol{a}^T\boldsymbol{x} = b\}$ and half-space $H^- := \{\boldsymbol{x} : \boldsymbol{a}^T\boldsymbol{x} \leq b\}$ are

$$P_C(\boldsymbol{y}) = \boldsymbol{y} - A^T(AA^T)^{-1}(A\boldsymbol{y} - \boldsymbol{b})$$

$$P_H(\boldsymbol{y}) = \boldsymbol{y} - \frac{\boldsymbol{a}^T\boldsymbol{y} - b}{\|\boldsymbol{a}\|^2}\boldsymbol{a}$$

$$P_{H^-}(\boldsymbol{y}) = \boldsymbol{y} - \frac{[\boldsymbol{a}^T\boldsymbol{y} - b]_+}{\|\boldsymbol{a}\|^2}\boldsymbol{a}$$

note AA^T is guaranteed to be invertible because A has full row rank by our assumption.

6.1 General KKT Condition

In general, optimisation problem with constraints takes the form

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

where f, g_i, h_j are continuously differentiable.

In this general case, (KKT1) is often written as $\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = 0$ where

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) := f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{x}) + \sum_{j=1}^{p} \mu_j h_j(\boldsymbol{x})$$

For the LCP case, with the following definitions

$$A := \begin{pmatrix} \boldsymbol{a}_1^T \\ \vdots \\ \boldsymbol{a}_m^T \end{pmatrix},\ C := \begin{pmatrix} \boldsymbol{c}_1^T \\ \vdots \\ \boldsymbol{c}_p^T \end{pmatrix},\ \boldsymbol{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix},\ \boldsymbol{d} = \begin{pmatrix} d_1 \\ \vdots \\ d_p \end{pmatrix}$$

constraints become $A\boldsymbol{x} \leq \boldsymbol{b}, C\boldsymbol{x} = \boldsymbol{d}$ and Lagrangian becomes

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

with KKT condition

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

A definition makes life easier when dealing with general cases:

Definition 5. When considering problem

$$\begin{aligned} \min \quad & h(\boldsymbol{x}) \\ \text{s.t.} \quad & \boldsymbol{x} \in C \end{aligned}$$

where h is continuously differentiable defined over closed and convex C. $\boldsymbol{d} \neq \boldsymbol{0}$ is called feasible descent direction at $\boldsymbol{x} \in C$ if $\nabla h(\boldsymbol{x})^T \boldsymbol{d} < 0$ and exists $\epsilon > 0$ s.t. $\boldsymbol{x} + t\boldsymbol{d} \in C$ for all $t \in [0, \epsilon]$.

It can be shown that x^* is local optimal solution iff no feasible descent direction at $\boldsymbol{x}^*$ exists. (either you cannot descend anymore, or descend direction is outside the constraints)

Now we begin to state theorems for KKT conditions in non-linear case, but will not prove them.

The original KKT condition for problem

$$\begin{aligned} \min \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & g_i(x) \leq 0,\ i = 1, 2, \cdots, m \end{aligned}$$

(where f, g_i are continuously differentiable) is given by Fritz-John, proved using Gordan's theorem. Fritz-John condition: exists $\lambda_0, \cdots, \lambda_m \geq 0$ which are not all zero s.t.

$$\lambda_0 \nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\boldsymbol{x}^*) = \boldsymbol{0} \quad \text{(FJ1)}$$

$$\lambda_i g_i(\boldsymbol{x}^*) = 0 \quad \text{for } i = 1, \cdots m \quad \text{(FJ2)}$$

The problem is if $\lambda_0 = 0$, (FJ1) simply becomes

$$\sum_{i=1}^{m} \lambda_i \nabla g_i(\boldsymbol{x}^*) = \boldsymbol{0}$$

which means $\{\nabla g_i(\boldsymbol{x}^*)\}_{i \in I(\boldsymbol{x}^*)}$ (similar to before, we define $I(\boldsymbol{x}) := \{i : g_i(\boldsymbol{x}) = 0\}$) is linearly dependent set. Adding the constraint that this set is linearly independent, and setting $\lambda_0 = 1$ gives us a useful kKT condition

Theorem 6.7 (KKT for inequality constraints)**.** *If $\boldsymbol{x}^*$ is local minimum of*

$$\begin{aligned} \min \quad & f(\boldsymbol{x}) \\ \textit{s.t.} \quad & g_i(x) \leq 0,\ i = 1, 2, \cdots, m \end{aligned}$$

where f, g_i are continuously differentiable. If $\{\nabla g_i(\boldsymbol{x}^*)\}_{i \in I(\boldsymbol{x}^*)}$ is linearly independent, then exists $\lambda_1, \cdots, \lambda_m \geq 0$ s.t.

$$\nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\boldsymbol{x}^*) = \boldsymbol{0}$$

$$\lambda_i g_i(\boldsymbol{x}^*) = 0 \quad \textit{for } i = 1, \cdots m$$

Now back to the full problem (P), with equality constratins,

Theorem 6.8 (KKT full theorem)**.** *If $\boldsymbol{x}^*$ is local minimum of (P), where f, g_i, h_j are continuously differentiable. If*

$$\{\nabla g_i(\boldsymbol{x}^*)\}_{i \in I(\boldsymbol{x}^*)} \cup \{\nabla h_j(\boldsymbol{x}^*)\}_{j = 1, \cdots, p}$$

is linearly independent, then exists $\lambda_1, \cdots, \lambda_m \geq 0,\ \mu_1, \cdots, \mu_p \in \mathbb{R}$ s.t.

$$\nabla L(\boldsymbol{x}^*, \boldsymbol{\lambda}, \boldsymbol{\mu}) = 0 \quad \textit{(KKT1)}$$

$$\lambda_i g_i(\boldsymbol{x}^*) = 0 \quad \textit{for } i = 1, \cdots m \quad \textit{(KKT2)}$$

Remark. Convexity of f, g_i, h_j are not required here.

The additional assumption is defined as regularity

Definition 6 (Regularity)**.** For optimisation problem (P), feasible point $\boldsymbol{x}^*$ is regular if set of gradients of active constraints of inequality constraints and equality constraints,

$$\{\nabla g_i(\boldsymbol{x}^*)\}_{i\in I(\boldsymbol{x}^*)} \cup \{\nabla h_j(\boldsymbol{x}^*)\}_{j=1,\cdots,p}$$

is linearly independent.

With this definition, Theorem 6.8 means necessary optimality condition for a regular point to be local minima is KKT condition. If there is no irregular point, every local minima satisfies KKT condition. The regularity condition is not required for LCP. (If gradients of two linear constraints are linearly dependent, then essentially they are the same constraint!)

Regularity can also be used to derive necessary second-order condition, when f, g_i, h_j are all twice continuously differentiable, if $\boldsymbol{x}^*$ is local minimum and regular, then KKT conditions are satisfied, and

$$\boldsymbol{d}^T \left[\nabla^2 f(\boldsymbol{x}^*) + \sum_{i=1}^m \lambda_i \nabla^2 g_i(\boldsymbol{x}^*) + \sum_{j=1}^p \mu_j \nabla^2 h_j(\boldsymbol{x}^*)\right] \boldsymbol{d} \geq 0 \quad \forall \boldsymbol{d} \in \Lambda(\boldsymbol{x}^*)$$

where $\Lambda(\boldsymbol{x}^*)$ is the set of vectors perpendicular to gradient of active inequality constraints and all equality constraints

$$\Lambda(\boldsymbol{x}^*) := \{\boldsymbol{d} \in \mathbb{R}^n \,:\, \nabla g_i(\boldsymbol{x}^*)^T \boldsymbol{d} = 0 \; \forall i \in I(\boldsymbol{x}^*),\; \nabla h_j(\boldsymbol{x}^*)^T \boldsymbol{d} = 0 \text{ for } j = 1, 2, \cdots, p\}$$

6.2 KKT for Convex Problems

Things become easier when convexity assumption is added. KKT conditions are always sufficient for $\boldsymbol{x}^*$ to be optimal condition as long as f, g_i are continuously differentiable convex functions, and h_j are affine functions. Recall these conditions mean feasible set C is convex. The proof is surprisingly simple:

Proof. Note

$$s(\boldsymbol{x}) =: s(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \sum_{i=1}^m \lambda_i g_i(\boldsymbol{x}) + \sum_{j=1}^p \mu_j h_j(\boldsymbol{x})$$

is convex. By the (KKT1),

$$\nabla s(\boldsymbol{x}^*) = \nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{0}$$

so $\boldsymbol{x}^*$ must be minimiser of s over $\mathbb{R}^n$. So for any $\boldsymbol{x} \in \mathbb{R}^n$,

$$\begin{aligned}
f(\boldsymbol{x}^*) =& f(\boldsymbol{x}^*) + \sum_{i=1}^m \lambda_i g_i(\boldsymbol{x}^*) + \sum_{j=1}^p \mu_j h_j(\boldsymbol{x}^*) \quad \text{derived by (KKT2) and constraints } h_j(\boldsymbol{x}^*) = 0 \\
&= s(\boldsymbol{x}^*) \leq s(\boldsymbol{x}) \quad \text{because } \boldsymbol{x}^* \text{ is minimiser of } s \\
&= f(\boldsymbol{x}) + \sum_{i=1}^m \lambda_i g_i(\boldsymbol{x}) + \sum_{j=1}^p \mu_j h_j(\boldsymbol{x}) \\
&\leq f(\boldsymbol{x}) \quad \text{because } \lambda_i \geq 0,\; g_i(\boldsymbol{x}) \leq 0,\; h_j(\boldsymbol{x}) = 0
\end{aligned}$$

so $\boldsymbol{x}^*$ is optimal solution. □

Further, with convexity an very convenient *Slater's condition* to replace regularity condition, it only requires existence of one single point in feasible set to be away from the boundary of constraints $g_i(\boldsymbol{x}) \leq 0$, i.e. $\exists \hat{\boldsymbol{x}} \in C$(feasible set) s.t. $g_i(\hat{\boldsymbol{x}}) < 0$ for all i. Note this condition is automatically satisfied if g_i are affine(linear) maps. So consider the following problem:

$$\begin{aligned}
\text{(CP) min} \quad & f(\boldsymbol{x}) \\
\text{s.t.} \quad & g_i(x) \leq 0,\; i = 1, 2, \cdots, m \\
& h_j(\boldsymbol{x}) \leq 0,\; j = 1, 2, \cdots, p \\
& s_k(\boldsymbol{x}) = 0,\; k = 1, 2, \cdots, q
\end{aligned}$$

where h_j, s_k are affine, f, g_i are continuously differentiable convex functions, (generalised) Slater's condition is exists $\hat{\boldsymbol{x}} \in \mathbb{R}^n$ s.t.

$$\begin{aligned} g_i(\hat{\boldsymbol{x}}) &< 0, \quad i = 1, 2, \cdots, m \\ h_j(\hat{\boldsymbol{x}}) &\leq 0, \quad i = 1, 2, \cdots, p \\ s_k(\hat{\boldsymbol{x}}) &= 0, \quad i = 1, 2, \cdots, q \end{aligned}$$

Theorem 6.9 (Necessity of KKT for convex problem with constraints)**.** *If problem (CP) satisfies generalised Slater's condition, then exists* $\lambda_1, \cdots, \lambda_m, \eta_1, \eta_2, \cdots, \eta_p \geq 0,\ \mu_1, \mu_2 \cdots, \mu_q \in \mathbb{R}$ *s.t.*

$$\nabla L(\boldsymbol{x}^*, \boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\mu}) = 0$$

and

$$\lambda_i g_i(\boldsymbol{x}^*) = 0, \quad \eta_j h_j(\boldsymbol{x}^*) = 0 \quad \forall i, j$$

Below is a summary about relation of KKT conditions with optimality in various conditions on various optimisation problems

f(objective)	g_i(inequality)	h_j(equality)	condition	result
cont. diff.	affine	affine		optimal $\Rightarrow$ KKT
convex	affine	affine		optimal $\Leftrightarrow$ KKT
convex	convex	affine		optimal $\Leftarrow$ KKT
convex	convex	affine	Slater	optimal $\Leftrightarrow$ KKT
cont. diff.	cont. diff.	cont. diff.	regularity	optimal $\Rightarrow$ KKT

Note: cont. diff. means continuously differentiable. When the "condition" is empty, it means there is no extra condition required.

7 Duality

Given the following optimisation problem

$$\begin{aligned} (P)\min \quad & x_1^2 + x_2^2 + 2x_1 \\ \text{s.t.} \quad & x_1 + x_2 = 0 \end{aligned}$$

using basic techniques it is easy to prove that $(-1/2, 1/2)$ is a global minimiser and minimal value $f^* = -1/2$. Suppose the constraint is not this easy, we aim to find a lower bound for the minimal value f^*. One can turn (P) into unconstrained problem by

$$(UP)\min \quad x_1^2 + x_2^2 + 2x_1 + \mu(x_1 + x_2)$$

where $\mu \in \mathbb{R}$. Since on the feasible set of (P), $x_1 + x_2 = 0$, the optimal value of (UP) is lower bound of f^*. (UP) is another convex optimisation problem, so finding stationary point yields $(-1 - \mu/2, -\mu/2)$. The corresponding optimal value, denote by $q(\mu)$, is $q(\mu) = -\mu^2/2 - \mu - 1$. Now, we aim to find the best lower bound, i.e. find solution to the following problem

$$(\mathrm{D}) \quad \max_{\mu \in \mathbb{R}} \{q(\mu)\}$$

and the value of q at optimal point of (D) is denoted as q^*. It is clear that $q^* \leq f^*$ (called weak duality).

In this particular case, optimal value of (D) is given by $\mu = -1$, and $q^* = -1/2 = f^*$, so solving the *dual problem* (D) gives the solution to the primal(original) problem (P). This very useful property is called *Strong duality*. (Conditions for strong duality to hold will be studied later).

You may worry that similar techniques cannot be used on inequality constraints, because after adding the term with multiplier, you are not able to remove the constraint without changing the solutions. New primal problem:

$$\begin{aligned}(P)\min \quad & x_1^2 + x_2^2 + 2x_1 \\ \text{s.t.} \quad & x_1 + x_2 \le 0\end{aligned}$$

after adding Lagrange multiplier:

$$\begin{aligned}\max \quad & x_1^2 + x_2^2 + 2x_1 + \mu(x_1 + x_2) \\ & x_1 + x_2 \le 0\end{aligned}$$

But weak duality still holds! Suppose feasible set of (P) is S.

$$\begin{aligned}q(\mu) &= \min_{x\in\mathbb{R}^2} x_1^2 + x_2^2 + 2x_1 + \mu(x_1 + x_2) \\ &\le \min_{x\in S} x_1^2 + x_2^2 + 2x_1 + \mu(x_1 + x_2) \\ &\le \min_{x\in S} x_1^2 + x_2^2 + 2x_1 \quad \text{because } x_1 + x_2 \le 0 \text{ for } x \in S \\ &= f^*\end{aligned}$$

Now we rigorously state the definition of dual problem: Given general model (called primal problem in this section)

$$\begin{aligned}(P)\min \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & g_i(\boldsymbol{x}) \le 0, \quad i = 1, 2, \cdots, m \\ & h_j(\boldsymbol{x}) = 0, \quad i = 1, 2, \cdots, p \\ & \boldsymbol{x} \in X\end{aligned}$$

where f, g_i, h_j are functions with NO assumption (for now). Recall the Lagrangian $L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ defined in the chapter 6, define dual objective function

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

note all constraints are dropped for this minimisation. You may notice that there may not be a value $\boldsymbol{x}$ that attains this infimum, and value of q can be $-\infty$. So an additional requirement that $q(\boldsymbol{\lambda}, \boldsymbol{\mu}) > -\infty$ is required.

$$\begin{aligned}(D)\min \quad & q(\boldsymbol{\lambda}, \boldsymbol{\mu}) \\ \text{s.t.} \quad & (\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \operatorname{dom}(q)\end{aligned}$$

where

$$\operatorname{dom}(q) := \{(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathbb{R}_+^m \times \mathbb{R}^p \,:\, q(\boldsymbol{\lambda}, \boldsymbol{\mu}) > -\infty\}$$

Assume f, g_i, h_j in (P) are finite-valued. It can be shown that $\operatorname{dom}(q)$ is convex

Proof. Given $(\boldsymbol{\lambda}_1, \boldsymbol{\mu}_1), (\boldsymbol{\lambda}_2, \boldsymbol{\mu}_2) \in \operatorname{dom}(q)$, since Lagrangian $L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ is affine w.r.t. $\boldsymbol{\lambda}, \boldsymbol{\mu}$,

$$\begin{aligned}q(\alpha\boldsymbol{\lambda}_1 + (1-\alpha)\boldsymbol{\lambda}_2, \alpha\boldsymbol{\mu}_1 + (1-\alpha)\boldsymbol{\mu}_2) &= \min_{\boldsymbol{x}\in X} L(\boldsymbol{x}, \alpha\boldsymbol{\lambda}_1 + (1-\alpha)\boldsymbol{\lambda}_2, \alpha\boldsymbol{\mu}_1 + (1-\alpha)\boldsymbol{\mu}_2) \\ &= \min_{\boldsymbol{x}\in X} [\alpha L(\boldsymbol{x}, \boldsymbol{\lambda}_1, \boldsymbol{\mu}_1) + (1-\alpha) L(\boldsymbol{x}, \boldsymbol{\lambda}_2, \boldsymbol{\mu}_2)] \\ &\ge \alpha \min_{\boldsymbol{x}\in X} L(\boldsymbol{x}, \boldsymbol{\lambda}_1, \boldsymbol{\mu}_1) + (1-\alpha) \min_{\boldsymbol{x}\in X} L(\boldsymbol{x}, \boldsymbol{\lambda}_2, \boldsymbol{\mu}_2) \\ &= aq(\boldsymbol{\lambda}_1, \boldsymbol{\mu}_1) + (1-\alpha) q(\boldsymbol{\lambda}_2, \boldsymbol{\mu}_2) \\ &> -\infty \quad \text{because } (\boldsymbol{\lambda}_1, \boldsymbol{\mu}_1), (\boldsymbol{\lambda}_2, \boldsymbol{\mu}_2) \in \operatorname{dom}(q)\end{aligned}$$

□

From the above proof, you may have also noticed that since $L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ is affine, $q(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is minimum of concave functions, so q is concave. (follows from Theorem 4.9) Then (D) is maximisation problem of concave q over convex set, which is essentially minimisation of $-q$. So (D) is a convex problem.

As mentioned before, dual problem is used to find lower bound for f^*, the optimal value of f,

Theorem 7.1 (Weak Duality Theorem)**.** *If q^* is optimal solution of (D), f^* is optimal solution of (P), then $q^* \leq f^*$.*

Proof. Similar to the proof used for minimisation of $x_1^2 + x_2^2 + 2x_1$ discussed before. See page 102 of Lecture notes for detailed proof. □

Duality problem may give useless lower bound,

Example 9. Consider

$$\begin{aligned}(P)\min \quad & x_1^2 - 3x_2^2 \\ \text{s.t. } & x_1 = x_2^3\end{aligned}$$

This function is not convex, and the equality constraint is not affine.
Using substitution, one can find that $(1,1), (-1,-1)$ are optimal solutions giving $f^* = -2$. But the Lagrangian is

$$L(x_1, x_2, \mu) = x_1^2 + \mu x_1 - 3x_2^2 - \mu x_2^3$$

if $\mu \geq 0$, plug in $(0, x_2)$ and send $x_2 \to \infty$ gives $-\infty$, or if $\mu < 0$, plug in $(x_1, 0)$ and send $x_2 \to \infty$ gives $-\infty$. So for any $\mu \in \mathbb{R}$, minimisation of L is $-\infty$, and that means dual optimal value is $q^* = -\infty$.

7.1 Strong Duality

Convexity leads us to strong duality property, where optimal value of primal and dual problem coincides. Proof of strong duality theorem requires a variation of separation theorem, where assumption "closed" of the convex set is removed. Now instead of requiring "separation", we allow the point to sit on boundary of the convex set

Theorem 7.2 (Supporting Hyperplane Theorem)**.** *Given convex set C and point $\boldsymbol{y} \notin C$. Exists $\boldsymbol{p} \neq \mathbf{0}$ s.t.*

$$\boldsymbol{p}^T \boldsymbol{x} \leq \boldsymbol{p}^T \boldsymbol{y} \; \forall \boldsymbol{x} \in C$$

Proof. We play some topological tricks to allow us to apply separation theorem on $\overline{C}$, the closure of C, which is closed by definition and convex. Unfortunately, we cannot guarantee $\boldsymbol{y} \notin \overline{C}$, but since $\boldsymbol{y} \notin \text{int}(C) = \text{int}(\overline{C})$, we can find sequence of $\boldsymbol{y}_k \notin \overline{C}$ s.t. $\boldsymbol{y}_k \to \boldsymbol{y}$. Then applying separation theorem, there exists $\boldsymbol{p}_k \neq \mathbf{0}$ s.t.

$$\boldsymbol{p}_k^T \boldsymbol{x} < \boldsymbol{p}_k^T \boldsymbol{y}_k \; \forall \boldsymbol{x} \in \overline{C}$$

already close to our objective. But $\boldsymbol{p}_k$ may not converge, and further, this sequence may not be bounded! But, we can derive the following inequality:

$$\frac{\boldsymbol{p}_k^T}{\|\boldsymbol{p}_k\|}(\boldsymbol{x} - \boldsymbol{y}_k) < 0 \; \forall \boldsymbol{x} \in \overline{C}$$

the sequence $(\frac{\boldsymbol{p}_k^T}{\|\boldsymbol{p}_k\|})_{k \in \mathbb{N}}$ is bounded (norm of the whole sequence is just 1), so by Weierstrass theorem, there is a convergent subsequence, say the limit is $\boldsymbol{p}$. We have $\|\boldsymbol{p}\| = 1$ so $\boldsymbol{p} \neq \mathbf{0}$. Using this convergent subsequence on the inequality yields

$$\boldsymbol{p}^T(\boldsymbol{x} - \boldsymbol{y}) \leq 0 \; \forall x \in \overline{C}$$

the result follows as $C \subseteq \overline{C}$. □

Theorem 7.3 (Separation of convex sets)**.** *Given non-empty convex sets C_1, C_2 with $C_1 \cap C_2 = \emptyset$. Then exists $\mathbf{0} \neq \boldsymbol{p}$ s.t.*

$$\boldsymbol{p}^T \boldsymbol{x} \leq \boldsymbol{p}^T \boldsymbol{y} \; \forall \boldsymbol{x} \in C_1, \, \boldsymbol{y} \in C_2$$

Proof. Use Theorem 7.2 with $C_1 - C_2$ (prove it is convex) and $\mathbf{0}$. Vector $\mathbf{0} \notin C_1 - C_2$ as C_1, C_2 are disjoint. □

Farkas lemma can be generalised to non-linear cases, with Slater-type conditions.

Theorem 7.4 (Nonlinear Farkas Lemma)**.** *Given convex set X and convex functions $f, g_1, \cdots, g_m$ (not necessarily continuous), assume exists $\hat{\boldsymbol{x}} \in X$ s.t. $g_i(\hat{\boldsymbol{x}}) < 0 \, \forall i$. Given $c \in \mathbb{R}$, the following two are equivalent:*
(a) $\boldsymbol{x} \in X, g_i(\boldsymbol{x}) \leq 0 \Rightarrow f(\boldsymbol{x}) \geq c$
(b) exists $\lambda_1, \cdots, \lambda_m \geq 0$ s.t.

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

note the function inside is $L(\boldsymbol{x}, \boldsymbol{\lambda})$ when g_i are used as inequality constraints.

The proof is similar to that of Farkas lemma but rather long. See (page 242-243, Beck, 2014)

Theorem 7.5 (Strong Duality Property of convex problems)**.** *Given optimisation problem*

$$\begin{aligned}(P)\min \quad & f(\boldsymbol{x}) \\ s.t.\ & g_i(\boldsymbol{x}) \le 0, \quad i = 1, 2, \cdots, m\end{aligned}$$

with X convex, f, g_i convex functions. Assume exists $\hat{\boldsymbol{x}} \in X$ s.t. $g_i(\hat{\boldsymbol{x}}) < 0 \, \forall i$. (this is exactly Slater's condition). If the optimal value f^* of problem (P) is finite, then optimal value of dual problem is attained, and

$$q^* = f^*$$

where q^* is optimal value of the dual problem, i.e.

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

Proof. By optimality of f^*,

$$\boldsymbol{x} \in X, g_i(\boldsymbol{x}) \le 0 \Rightarrow f(\boldsymbol{x}) \ge f^* \, \forall \boldsymbol{x} \in X$$

So by nonlinear Farkas's lemma, there are $\lambda_i \ge 0$ s.t.

$$q(\boldsymbol{\lambda}) = \min_{\boldsymbol{x} \in X} L(\boldsymbol{x}, \boldsymbol{\lambda}) \ge f^*$$

So combining with weak duality theorem,

$$q^* \ge q(\boldsymbol{\lambda}) \ge f^* \ge q^*$$

which means $f^* = g^*$. □

When the Slater's condition is not satisfied, dual problem may still give the strong dual property, but optimal value of (D) is not attained.

Example 10.

$$\begin{aligned}\min \quad & x_1^2 - x_2 \\ \text{s.t.} \ & x_2^2 \le 0\end{aligned}$$

The problem is convex but not satisfy Slater's condition. Optimal solution is simply $(0, 0)$ giving $f^* = 0$. The Lagrangian is $L(x_1, x_2, \lambda) = x_1^2 - x_2 + \lambda x_2^2$ giving dual objective

$$q(\lambda) = \min_{x_1, x_2 \in \mathbb{R}} L(x_1, x_2, \lambda) = \begin{cases} -\infty, & \lambda = 0 \\ -\frac{1}{4\lambda}, & \lambda > 0 \end{cases}$$

So dual problem has optimal value $q^* = 0$, but not attained by any $\lambda \ge 0$.

Equality conditions can be added to Theorem 7.5

Theorem 7.6. *Consider*

$$\begin{aligned}(P)\min \quad & f(\boldsymbol{x}) \\ s.t.\ & g_i(\boldsymbol{x}) \le 0, \quad i = 1, 2, \cdots, m \\ & h_j(\boldsymbol{x}) \le 0, \quad j = 1, 2, \cdots, p \\ & s_k(\boldsymbol{x}) = 0, \quad k = 1, 2, \cdots, q \\ & \boldsymbol{x} \in X\end{aligned}$$

where X is convex, f, g_i are convex, h_j, s_k are affine. Suppose generalised Slater's condition is satisfied, then if optimal value f^* of (P) is finite, optimal of dual problem is attained and

$$f^* = q^*$$

where

$$\begin{aligned} q^* &:= \max\{q(\boldsymbol{\lambda}, \boldsymbol{\nu}, \boldsymbol{\mu}) \,:\, \boldsymbol{\lambda}, \boldsymbol{\nu}, \boldsymbol{\mu} \in dom(q)\} \\ q(\boldsymbol{\lambda}, \boldsymbol{\nu}, \boldsymbol{\mu}) &:= \min_{\boldsymbol{x} \in X} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \boldsymbol{\mu}) \end{aligned}$$

.

One can prove that complementary slackness condition (similar to that of KKT2) holds as long as $q^* = f^*$ (even if f is not convex)

Theorem 7.7 (Complementary Slackness)**.** *For problem*

$$\begin{aligned} \min \quad & x_1^2 - x_2 \\ s.t. \ & x_2^2 \leq 0 \end{aligned}$$

if $f^* = q^*$, then if $\boldsymbol{x}^*, \boldsymbol{\lambda}^*$ are optimal solutions of primal and dual problem respectively, then

$$\boldsymbol{x}^* \in ArgminL(\boldsymbol{x}, \boldsymbol{\lambda}^*)$$

$$\lambda_i^* g_i(\boldsymbol{x}^*) = 0 \ \forall i$$

References

[1] Beck, Amir. Introduction to Nonlinear Optimization. SIAM, 2014.

[2] Chen, Xuemei. Kaczmarz Algorithm, Row Action Methods, and Statistical Learning Algorithms. faculty.sites.iastate.edu/esweber/files/inline-files/KaczmarzSGDrevised.pdf.

[3] S. Kaczmarz. Angenäherte auflösung von systemen linearer gleichungen. Bulletin International de l'Académie Polonaise des Sciences et des Lettres A, 35:355–357, 1937.

[4] Weisstein, Eric W. "Ellipsoid." Mathworld.wolfram.com, mathworld.wolfram.com/Ellipsoid.html.

[5] Armijo, Larry. "Minimization of Functions Having Lipschitz Continuous First Partial Derivatives." Pacific Journal of Mathematics, vol. 16, no. 1, Jan. 1966, pp. 1–3, https://doi.org/10.2140/pjm.1966.16.1.

[6] Sun, T., Cheng, L. Convergence of iterative hard-thresholding algorithm with continuation. Optim Lett 11, 801–815 (2017). https://doi.org/10.1007/s11590-016-1062-0