A dissertation for the degree of
Doctor of Philosophy

Canonical Duality Theory for Global Optimization Problems and Applications

Yi Chen

Faculty of Science and Technology

June 2015

Abstract

The canonical duality theory is studied, through a discussion on a general global optimization problem and applications on fundamentally important problems. This general problem is a formulation of the minimization problem with inequality constraints, where the objective function and constraints are any convex or nonconvex functions satisfying certain decomposition conditions. It covers convex problems, mixed integer programming problems and many other nonlinear programming problems. The three main parts of the canonical duality theory are canonical dual transformation, complementary-dual principle and triality theory. The complementary-dual principle is further developed, which conventionally states that each critical point of the canonical dual problem is corresponding to a KKT point of the primal problem with their sharing the same function value. The new result emphasizes that there exists a one-to-one correspondence between KKT points of the dual problem and of the primal problem and each pair of the corresponding KKT points share the same function value, which implies that there is truly no duality gap between the canonical dual problem and the primal problem. The triality theory reveals insightful information about global and local solutions. It is shown that as long as the global optimality condition holds true, the primal problem is equivalent to a convex problem in the dual space, which can be solved efficiently by existing convex methods; even if the condition does not hold, the convex problem still provides a lower bound that is at least as good as that by the Lagrangian relaxation method. It is also shown that through examining the canonical dual problem, the hidden convexity of the primal problem is easily observable.

The canonical duality theory is then applied to dealing with three fundamentally important problems. The first one is the spherically constrained quadratic problem, also referred to as the trust region subproblem. The canonical dual problem is one-dimensional and it is proved that the primal problem, no matter with convex or nonconvex objective function, is equivalent to a convex problem in the dual space. Moreover, conditions are found which comprise the boundary that separates instances into "hard case" and "easy case". A canonical primal-dual algorithm is developed, which is able to efficiently solve the problem, including the "hard case", and can be used as a unified method for similar problems. The second one is the binary quadratic problem, a fundamental problem in discrete optimization. The discussion is focused on lower bounds and analytically solvable cases, which are obtained by analyzing the canonical dual problem with perturbation techniques. The third one

is a general nonconvex problem with log-sum-exp functions and quartic polynomials. It arises widely in engineering science and it can be used to approximate nonsmooth optimization problems. The work shows that problems can still be efficiently solved, via the canonical duality approach, even if they are nonconvex and nonsmooth.

To My Parents.

Acknowledgements

I would like to gratefully and sincerely thank my supervisors, Professor David Yang Gao and Professor John Yearwood, whose contributions of time and mentoring are of great importance to the completion of my PhD study and thesis. I would also like to thank the university and the staff for your excellent support.

Special thanks to my family and friends. It is your understanding and support that encouraged me to carry on with my doctoral journey.

Statement of Authorship

I, Yi Chen, hereby declare that: I am the sole author of this thesis; I have fully acknowledged and referenced the ideas and work of others, whether published or unpublished, in my thesis; I have prepared my thesis specifically for the degree of Doctor of Philosophy, while under supervision at Federation University Australia; My thesis does not contain work extracted from a thesis, dissertation or research paper previously presented for another degree or diploma at this or any other university.

Candidate's signature: Yichen Date: 18.06.2015

Contents

Glossary

$\mathbb{R}$	Real numbers.
$\mathbb{R}_+$	Nonnegative real numbers.
$\mathbb{R}_{++}$	Positive real numbers.
$\mathbb{R}^n$	n-dimensional Euclidean space.
$\mathbb{R}^{n\times m}$	$n \times m$-dimensional real matrices space.
$\mathbb{S}^n$	$\{A \in \mathbb{R}^{n\times n} \mid A^T = A\}$.
$\mathbb{S}^n_+$	all positive semidefinite matrices in $\mathbb{S}^n$.
$\mathbb{S}^n_{++}$	all positive definite matrices in $\mathbb{S}^n$.
$\{a, b\}^n$	Set of n-dimensional vectors whose components are a or b.
$\lvert S\rvert$	Cardinality of the set S.
$\lVert\boldsymbol{x}\rVert$	Euclidean norm, i.e., $\lVert\boldsymbol{x}\rVert_2$.
$\{x_i\}_{i=1}^n$	Column vector $(x_1, \ldots, x_n)^T$.
$\boldsymbol{e}$	All-ones vector.
I	Identity matrix.
A^T	Transpose of matrix A.
$A^\dagger$	Moore-Penrose or pseudo-inverse of matrix A.
$\mathrm{tr}(A)$	Trace of matrix A.
$\mathrm{rank}(A)$	Rank of matrix A.
$\mathrm{diag}(A)$	Diagonal vector of matrix A.
$\mathrm{diag}(\boldsymbol{x})$	Diagonal matrix with diagonal elements $x_1, \ldots, x_n$.
$A \circ B$	Hadamard product of matrices, i.e. $A \circ B = \{a_{ij}b_{ij}\}_{i,j=1}^n$.
$A \cdot B$	Inner product of matrices, i.e. $A \cdot B = \mathrm{tr}(AB)$.
$Q \succ 0$	Denotes Q is a positive definite matrix.
$Q \succeq 0$	Denotes Q is a positive semidefinite matrix.
f^*	Conjugate function of f.
$\mathrm{dom} f$	Domain of function f.
∇f	Gradient of function f.
$\nabla^2 f$	Hessian of function f.
$\partial f/\partial \boldsymbol{x}$	Partial derivative of function f with respect to $\boldsymbol{x}$.
$\exp(x)$	Exponential function e^x.
$\log(x)$	Logarithmic function $\log_e x$.

Chapter 1

Introduction

The *Canonical duality theory* was developed from Gao and Strang's original work for solving a general nonconvex/nonsmooth variational problem [54]:

$$\min\{\Pi(\boldsymbol{\chi}) = W(D\boldsymbol{\chi}) - F(\boldsymbol{\chi}) \mid \boldsymbol{\chi} \in \mathcal{X}_c\}. \tag{1.1}$$

The function $F(\boldsymbol{\chi})$ models the external energy and must be linear on its domain $\mathcal{X}_a$; while the function $W : \mathcal{W}_a \to \mathbb{R}$ models the internal energy and must possess "objectivity"[1]. Here the linear operator $D : \mathcal{X}_a \to \mathcal{W}_a$ assigns each configuration $\boldsymbol{\chi}$ to an internal variable $\boldsymbol{\epsilon} = D\boldsymbol{\chi}$. The feasible set $\mathcal{X}_c = \{\boldsymbol{\chi} \in \mathcal{X}_a \mid D\boldsymbol{\chi} \in \mathcal{W}_a\}$ is called the kinetically admissible space. Through the model (1.1), the canonical duality theory can be illustratively described, and more detailed explanations can be found in the recent paper [48].

The "objectivity" that the internal energy W must possess is mathematically defined in [38] (Definition 6.1.2). In Euclidean space, the definition is given as follows: let $\mathcal{X}_a \subseteq \mathbb{R}^n$, $\mathcal{W}_a \in \mathbb{R}^m$, and $\mathcal{R} = \{R \in \mathbb{R}^{n\times m} \mid R^T = R^{-1}\}$, i.e., the set of all orthogonal matrices; a real-valued function $W : \mathcal{W}_a \to \mathbb{R}$ is said to be "objective", if $\mathcal{W}_a$ and W satisfy $R\boldsymbol{\epsilon} \in \mathcal{W}_a, W(R\boldsymbol{\epsilon}) = W(\boldsymbol{\epsilon}),\ \forall \boldsymbol{\epsilon} \in \mathcal{W}_a,\ \forall R \in \mathcal{R}$. Geometrically, it means that the domain and the function are invariant under any rotations about the origin.

There are principally three parts which comprise the canonical duality theory:

1) a *canonical dual transformation*, which is used to reformulate nonconvex or discrete problems arising in different systems as a unified canonical dual problems;

2) a *complementary-dual principle*, which illustrates the perfect duality relation between the primal and dual problems (there are no duality gaps) and provides unified analytical solutions for the primal problem in terms of the canonical dual solutions;

[1]Here, the word means that the function only depends on certain measure of its variables [48], in contrast to the meaning of target or goal in "objective function" in mathematical optimization.

3) a *triality theory*, which reveals an intrinsic duality pattern, being composed of *canonical min-max duality*, *double-min duality* and *double-max duality*, which can be used to identify global optimal solutions and local solutions.

The canonical duality theory has been used successfully for solving a wide range of difficult problems, within a unified framework. The applications of canonical duality in global optimization include solving quadratic problems [32, 42, 46, 47, 51], polynomial optimization problems [44], transportation problems [52], location problems [49, 111], network optimization problems [114], geometry problems [115], fixed point problems [118], fractional programming problems [113], optimization problems in machine learning [112, 82, 81], and mixed integer programming problems [128, 110]. Recently, an open problem left in the triality theory has been solved [56] and some efficient algorithms have been developed [55].

As emphasized in the paper [48], the "objectivity" is a necessity for the canonical duality theory. However, many optimization problems arising in the real world do not possess the "objectivity", even if they are convex. For example, when the matrix Q is indefinite, the quadratic function $h(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x}$ can not be transformed into $W(D\boldsymbol{x})$ with D being a linear operator and W being "objective". The objective here is to investigate the theory and applications of the canonical duality for general convex/nonconvex optimization problems, which may not possess the "objectivity". In mainly two aspects, the thesis attempt to achieve positive results: (1) revealing insightful relations that might not otherwise be observed; (2) inspiring the development of unified solution methods.

In this thesis, the investigation of the canonical duality theory focuses on a general optimization problem, which is only assumed to satisfy certain decomposition conditions and where the "objectivity" is not a necessity. This general problem covers convex problems, mixed integer programming problems and many other nonlinear programming problems. The three main parts of the canonical duality theory are then developed. The complementary-dual principle, which conventionally only says that each critical point of the canonical dual problem is corresponding to a KKT point of the primal problem with they sharing the same function value, is further developed, and it truly reveals that there are no duality gaps between primal and dual problems. In the case where all operators are quadratic, the triality theory is proposed, and besides the global optimality condition, comprehensive relations among certain local solutions are also presented. Through examining the canonical dual problem, the hidden convexity of the primal problem is discussed.

Then the canonical duality theory is applied to dealing with three fundamentally important problems. The first one is the spherically constrained quadratic problem, also referred to as the trust region subproblem. The difficulty here is to efficiently solve problems in "hard case". By applying the canonical duality, a boundary that separates instances into "hard case" and "easy case" is discovered, and it inspires a canonical primal-dual algorithm. The second problem is the binary quadratic problem, a fundamental problem in discrete optimization. The discussion is focused on lower bounds and analytically solvable cases, which are obtained by analyzing the

canonical dual problem with perturbation techniques. The third one is a nonconvex problem with log-sum-exp functions and quartic polynomials. It arises widely in engineering science and it can be used to approximate nonsmooth optimization problems. The discussion attempts to show that the optimization problems, even if they are nonconvex and nonsmooth, can still be efficiently solved via the canonical duality approach.

The remaining chapters are organized as follows. In Chapter 2, the basic definitions and results of convex analysis, mathematical optimization, and Lagrangian duality are presented as a preparation for the later discussions. In Chapter 3, the canonical duality theory for the proposed optimization problem is developed; both the general case and the case where operators are quadratic are discussed. Then, three important problems are provided to illustrate the application of the canonical duality theory: the spherically constrained quadratic problem in Chapter 4, the binary quadratic problem in Chapter 5, and a nonconvex problem with log-sum-exp functions and quartic polynomials in Chapter 6. In the end, the key findings are summarized in Chapter 7, with some remarks on the future directions.

The work in Chapter 4 has been presented in The 3rd World Congress of Global Optimization (July 8-12, 2013, The Yellow Mountains, China) and published in [24] and [25]. The work in Chapter 6 has been presented in The 5th International Conference on Optimization and Control with Applications (December 4-8, 2012, Beijing, China) and published in [23].

Chapter 2

Preliminaries

2.1 Convex analysis

In this section, convex sets, convex functions and their properties, and the theory of conjugate are introduced. For more comprehensive descriptions and proofs, refer to [21, 12].

2.1.1 Convex sets

A set $\mathcal{C} \in \mathbb{R}^n$ is said to be *convex* if the line segment joining any two points in the set also belongs to the set, i.e., if for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{C}$ and any $\lambda \in [0, 1]$, we have

$$\lambda \boldsymbol{x}_1 + (1 - \lambda)\boldsymbol{x}_2 \in \mathcal{C}.$$

The line segment, $\lambda \boldsymbol{x}_1 + (1 - \lambda)\boldsymbol{x}_2$ with $\lambda \in [0, 1]$, consists of all points between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ on the line that passes through. The form $\sum_{i=1}^n \lambda_i \boldsymbol{x}_i$ with $\sum_{i=1}^n \lambda_i = 1, \lambda_i \geq 0, i = 1, \ldots, n$ is called a *convex combination* of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. If the nonnegativity conditions, $\lambda_i \geq 0, i = 1, \ldots, n$, are dropped, the combination is known as an *affine combination*, and if the multipliers $\lambda_i, i = 1, \ldots, n$ are simply required to be in $\mathbb{R}$, the form is known as a *linear combination*.

For any $\mathcal{C} \in \mathbb{R}^n$, the smallest convex set containing $\mathcal{C}$ is called *convex hull*, which is the set of all convex combinations of points in $\mathcal{C}$:

$$\operatorname{conv}(\mathcal{C}) = \left\{ \sum_{i=1}^n \lambda_i \boldsymbol{x}_i \mid \boldsymbol{x}_i \in \mathcal{C}, \lambda_i \geq 0, i = 1, \ldots, n, \sum_{i=1}^n \lambda_i = 1 \right\}.$$

Hence, if $\mathcal{C}$ is a convex set, its convex hull is $\mathcal{C}$ itself.

The following foundamental sets are convex: (1) hyperplanes, $\{\boldsymbol{x} \mid \boldsymbol{a}^T\boldsymbol{x} = b\}$; (2) halfspaces, $\{\boldsymbol{x} \mid \boldsymbol{a}^T\boldsymbol{x} \leq b\}$ and $\{\boldsymbol{x} \mid \boldsymbol{a}^T\boldsymbol{x} < b\}$; (3) Euclidean balls, $\{\boldsymbol{x} \mid \|\boldsymbol{x} - \boldsymbol{x}_0\| \leq r\}$; (4) polyhedra, $\{\boldsymbol{x} \mid \boldsymbol{a}_j^T\boldsymbol{x} \leq b_j, j = 1, \ldots, m, \boldsymbol{c}_k^T\boldsymbol{x} = d_k, k = 1, \ldots, p\}$.

Convexity is preserved under some operations, which appears to be useful in convex analysis. If $\mathcal{C}_i, i = 1, \ldots, n$ are convex, then $\cap_{i=1}^n \mathcal{C}_i$ is convex. The property

still hold if n is infinity. The affine mapping also preserves the convexity: if $\mathcal{C}$ is convex and $f : \mathbb{R}^n \to \mathbb{R}^m$ is an affine function, then the image of $\mathcal{C}$ under f, $f(\mathcal{C}) = \{f(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{C}\}$, is convex.

A set $\mathcal{K} \subseteq \mathbb{R}^n$ is called a *cone* if $\alpha\boldsymbol{x} \in \mathcal{K}$ for every $\boldsymbol{x} \in \mathcal{K}$ and $\alpha \geq 0$. If, in addition, $\mathcal{K}$ is convex, $\mathcal{K}$ is called a *convex cone*. From its definition, we know that a cone always contains the zero point. A cone $\mathcal{K} \subseteq \mathbb{R}^n$ is called a *proper cone* if it is convex, closed, solid and pointed, where *solid* means $\mathcal{K}$ has nonempty interior and *pointed* means $\mathcal{K}$ contains no line.

The following convex cones are well-known and important in mathematical optimization:

- the *nonnegative orthant*

$$\mathbb{R}^n_+ = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x} \geq 0\},$$

- the set of *symmetric matrices*

$$\mathbb{S}^n = \{X \in \mathbb{R}^{n\times n} \mid X = X^T\},$$

- the set of symmetric *positive semidefinite matrices*

$$\mathbb{S}^n_+ = \{X \in \mathbb{S}^n \mid X \succeq 0\},$$
$$\mathbb{S}^n_{++} = \{X \in \mathbb{S}^n \mid X \succ 0\},$$

- the *second-order cone* (*quadratic cone*, *Lorentz cone* or *ice-cream cone*)

$$\mathcal{SOC}^n = \{(\boldsymbol{x}, t) \in \mathbb{R}^{n+1} \mid \|\boldsymbol{x}\| \leq t\},$$

- the set of symmetric *copositive matrices*

$$\mathcal{COP}^n = \{X \in \mathbb{S}^n \mid \boldsymbol{y}^T X \boldsymbol{y} \geq 0, \forall \boldsymbol{y} \in \mathbb{R}^n_+\},$$

- the set of symmetric *completely positive matrices*

$$\mathcal{CP}^n = \{Y = \sum_{i=1}^{k} \boldsymbol{y}_i \boldsymbol{y}_i^T \mid k > 0, \boldsymbol{y}_i \in \mathbb{R}^n_+, i = 1, \ldots, k\}.$$

Among them, $\mathbb{R}^n_+$, $\mathbb{S}^n_+$, $\mathcal{SOC}^n$, $\mathcal{COP}^n$ and $\mathcal{CP}^n$ are proper cones. Moreover, it is true that

$$\mathcal{CP}^n \subseteq \mathbb{S}^n_+ \subseteq \mathcal{COP}^n.$$

Let $\mathcal{K}$ be a cone in $\mathbb{R}^n$. An associated cone which is called *dual cone* is defined by

$$\mathcal{K}^* = \{\boldsymbol{y} \in \mathbb{R}^n \mid \boldsymbol{x}^T \boldsymbol{y} \geq 0 \text{ for all } \boldsymbol{x} \in \mathcal{K}\}.$$

The dual cone $\mathcal{K}^*$ is always a convex cone, even when the original cone $\mathcal{K}$ not. We have the following properties: (1) $\mathcal{K}^*$ is closed and convex; (2) $\mathcal{K}_1 \subseteq \mathcal{K}_2$ implies $\mathcal{K}_2^* \subseteq \mathcal{K}_1^*$; (3) If $\mathcal{K}$ has nonempty interior, then $\mathcal{K}^*$ is pointed; (4) If the closure of $\mathcal{K}$ is pointed then $\mathcal{K}^*$ has nonempty interior; (5) $\mathcal{K}^{**}$ is the closure of the convex hull of $\mathcal{K}$.

If $\mathcal{K}^* = \mathcal{K}$, then $\mathcal{K}$ is called a *self-dual* cone. Among the convex cones presented above, we have

$$(\mathbb{R}^n_+)^* = \mathbb{R}^n_+, \ (\mathbb{S}^n_+)^* = \mathbb{S}^n_+, \text{ and } (\mathcal{SOC}^n)^* = \mathcal{SOC}^n.$$

Whereas, the symmetric copositive matrices cone and the symmetric completely positive matrices cone are dual to each other, i.e.,

$$(\mathcal{CP}^n)^* = \mathcal{COP}^n \text{ and } (\mathcal{COP}^n)^* = \mathcal{CP}^n.$$

Here, only the proof for $\mathcal{S}^n_+$ is presented. By the definition, the dual cone for $\mathcal{S}^n_+$ is

$$(\mathbb{S}^n_+)^* = \{Y \in \mathbb{S}^{n\times n} \mid X \cdot Y \geq 0, \forall X \in \mathbb{S}^n_+\},$$

where $X \cdot Y = \text{tr}(XY)$, denoting the inner product of two matrices. Suppose $Y \in \mathbb{S}^n_+$. Then, for any $X \in \mathbb{S}^n_+$, we have

$$X \cdot Y = \sum_{i=1}^{n} \lambda_i (\boldsymbol{u}_i \boldsymbol{u}_i^T) \cdot Y = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i^T Y \boldsymbol{u}_i \geq 0,$$

where $\lambda_i \geq 0, i = 1, \ldots, n$ are the eigenvalues of X and $\boldsymbol{u}_i, i = 1, \ldots, n$ are the corresponding eigenvectors. This shows that $Y \in (\mathbb{S}^n_+)^*$. Now suppose $Y \in (\mathbb{S}^n_+)^*$. For any $\boldsymbol{x} \in \mathbb{R}^n$, as $\boldsymbol{x}\boldsymbol{x}^T \in \mathbb{S}^n_+$, we have $\boldsymbol{x}^T Y \boldsymbol{x} = \boldsymbol{x}\boldsymbol{x}^T \cdot Y \geq 0$, which is equivalent to $Y \in \mathbb{S}^n_+$.

At the end of this part, we present a very important result about the separation of two convex sets. Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be nonempty convex sets in $\mathbb{R}^n$ and suppose that $\mathcal{C}_1 \cap \mathcal{C}_2$ is empty. Then there exists a hyperplane that separates $\mathcal{C}_1$ and $\mathcal{C}_2$; that is, there exists a nonzero vector $\boldsymbol{p} \in \mathbb{R}^n$ such that

$$\inf\{\boldsymbol{p}^T \boldsymbol{x} \mid \boldsymbol{x} \in \mathcal{C}_1\} \geq \sup\{\boldsymbol{p}^T \boldsymbol{x} \mid \boldsymbol{x} \in \mathcal{C}_2\}.$$

2.1.2 Convex functions

Let $\mathcal{D}$ be a nonempty set in $\mathbb{R}^n$. The function $f : \mathcal{D} \to \mathbb{R}$ is said to be *convex* if the set $\mathcal{D}$ is convex and

$$f(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) \leq \lambda f(\boldsymbol{x}_1) + (1-\lambda) f(\boldsymbol{x}_2)$$

for each $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{D}$ and for each $\lambda \in (0, 1)$. The function f is called *strictly convex* on $\mathcal{D}$ if f satisfies the above inequality with $\leq$ being replaced by $<$. The function f is called *concave* (*strictly concave*) on $\mathcal{D}$ if $-f$ is convex (strictly convex) on $\mathcal{D}$. An

important property of convex and concave functions is that they are continuous on the interior of their domain.

Examples given below are some important convex functions that arise very often in practice.

1. (*Nonnegative weighted sums*) Let $f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$ be convex functions. Then $f(\boldsymbol{x}) = \sum_{j=1}^m \alpha_j f_j(\boldsymbol{x})$ with $\alpha_j \geq 0, j = 1, \ldots, m$ is a convex function.

2. (*Pointwise maximum*) Let $f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$ be convex functions. Then $f(\boldsymbol{x}) = \max\{f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})\}$ is also convex.

3. (*Composition with an affine mapping*) Let $g : \mathbb{R}^m \to \mathbb{R}$ be a convex function, and let $\boldsymbol{h} : \mathbb{R}^n \to \mathbb{R}^m$ be an affine function of the form $\boldsymbol{h}(\boldsymbol{x}) = A\boldsymbol{x} + \boldsymbol{b}$ with $A \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$. Then the composite function $f = g(\boldsymbol{h}(\boldsymbol{x})) : \mathbb{R}^n \to \mathbb{R}$ is a convex function.

4. (*Inverse*) Suppose that $g : \mathbb{R}^n \to \mathbb{R}$ is a concave function. Let $\mathcal{D} = \{\boldsymbol{x} \mid g(\boldsymbol{x}) > 0\}$. Then the function $f(\boldsymbol{x}) = 1/g(\boldsymbol{x}) : \mathcal{D} \to \mathbb{R}$ is convex on $\mathcal{D}$.

5. (*Supremum*) Suppose that $g(\boldsymbol{x}, \boldsymbol{y}) : \mathcal{C} \times \mathcal{D} \to \mathbb{R}$ is convex in $\boldsymbol{x}$ for each fixed $\boldsymbol{y} \in \mathcal{D}$. Then the function $f(\boldsymbol{x}) = \sup_{\boldsymbol{y} \in \mathcal{D}} g(\boldsymbol{x}, \boldsymbol{y})$ is convex in $\boldsymbol{x}$.

6. (*Minimization*) Suppose that $g(\boldsymbol{x}, \boldsymbol{y}) : \mathcal{C} \times \mathcal{D} \to \mathbb{R}$ is convex in $(\boldsymbol{x}, \boldsymbol{y})$. Then the function $f(\boldsymbol{x}) = \inf_{\boldsymbol{y} \in \mathcal{D}} g(\boldsymbol{x}, \boldsymbol{y})$ is convex in $\boldsymbol{x}$.

The *α-sublevel set* of a convex function $f : \mathcal{D} \to \mathbb{R}$ is defined by

$$\mathcal{D}_\alpha = \{\boldsymbol{x} \in \mathcal{D} \mid f(\boldsymbol{x}) \leq \alpha\}.$$

If f is concave, its *α-superlevel set* is defined by $\{\boldsymbol{x} \in \mathcal{D} \mid f(\boldsymbol{x}) \geq \alpha\}$. All sublevel sets of a convex function are convex, but the converse is not true.

The *epigraph* of a function $f : \mathcal{D} \to \mathbb{R}$ is a subset of $\mathbb{R}^{n+1}$ defined by

$$\text{epi} f = \{(\boldsymbol{x}, t) \mid \boldsymbol{x} \in \mathcal{D}, f(\boldsymbol{x}) \leq t\}.$$

The *hypograph* of f is defined by

$$\text{hyp } f = \{(\boldsymbol{x}, t) \mid \boldsymbol{x} \in \mathcal{D}, f(\boldsymbol{x}) \geq t\}.$$

It is true that a function being convex is equivalent to its epigraph being a convex set.

For differentiable functions, besides the definition and epigraph, there are other necessary and sufficient conditions that can be used to characterize convexity. Suppose that $\mathcal{D}$ is a nonempty open convex set in $\mathbb{R}^n$ and $f : \mathcal{D} \to \mathbb{R}$ is differentiable, i.e., its gradient ∇f exists at each point in $\mathcal{D}$. Then f is convex if and only if for any $\bar{\boldsymbol{x}} \in \mathcal{D}$, we have

$$f(\boldsymbol{x}) \geq f(\bar{\boldsymbol{x}}) + \nabla f(\bar{\boldsymbol{x}})^T (\boldsymbol{x} - \bar{\boldsymbol{x}}) \qquad \forall \boldsymbol{x} \in \mathcal{D}.$$

While, f is strictly convex if and only if the above inequality is strict for any $\boldsymbol{x} \neq \bar{\boldsymbol{x}}$. It is also true that f is convex if and only if for each $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{D}$ we have

$$(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1))^T(\boldsymbol{x}_2 - \boldsymbol{x}_1) \geq 0;$$

Similarly, f is strictly convex if and only if the above inequality is strict for any distinct $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{D}$.

From a computational standpoint, checking the above conditions is difficult. If the function is twice differentiable, a simple and more manageable characterization, at least for quadratic functions, can be obtained. Let $f : \mathcal{D} \to \mathbb{R}$ be twice differentiable on $\mathcal{D}$. Then f is convex if and only if the Hessian matrix is positive semidefinite,

$$\nabla^2 f(\boldsymbol{x}) \succeq 0,$$

at each point in $\mathcal{D}$. If the Hessian matrix $\nabla^2 f$ is positive definite at each point, then f is strictly convex. But, the converse for strictly convexity does not hold, i.e., for a general strictly convex function, its Hessian may not be positive definite at each point in $\mathcal{D}$. While if f is a quadratic function, the strict convexity is equivalent to positive definiteness.

There is an insightful connection between the univariate and multivariate cases. Consider a function $f : \mathbb{R}^n \to \mathbb{R}$, and for any point $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ and a nonzero direction $\boldsymbol{d} \in \mathbb{R}^n$, define $F_{(\bar{\boldsymbol{x}};\boldsymbol{d})}(\lambda) = f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d})$ as a function of $\lambda \in \mathbb{R}$. Then f is (strictly) convex if and only if $F_{(\bar{\boldsymbol{x}};\boldsymbol{d})}(\lambda)$ is (strictly) convex for all $\bar{\boldsymbol{x}}$ and $\boldsymbol{d} \neq 0$ in $\mathbb{R}^n$.

2.1.3 Conjugate functions

Fenchel's conjugate for convex functions plays an essential role in duality, including the canonical duality theory discussed in this study. Let $\mathcal{D}$ be a nonempty set in $\mathbb{R}^n$. The *Fenchel conjugate* (*conjugate* for short) of $f : \mathcal{D} \to \mathbb{R}^n$ is defined as

$$f^*(\boldsymbol{y}) = \sup_{\boldsymbol{x} \in \mathcal{D}} \{\boldsymbol{x}^T \boldsymbol{y} - f(\boldsymbol{x})\}.$$

The conjugate $f^*(\boldsymbol{y})$ is the point-wise supremum of a family of affine functions. Thus, it is a convex function, without regard to whether$f$ is convex or not. The domain of the conjugate function contains all $\boldsymbol{y} \in \mathbb{R}^n$ for which the supremum is finite. Passing from f to the conjugate f^* is called the *Legendre-Fenchel transformation*.

Immediately, we have the following inequality, which is called *Fenchel's inequality* (also known as *Fenchel-Young inequality*),

$$f(\boldsymbol{x}) + f^*(\boldsymbol{y}) \geq \boldsymbol{x}^T \boldsymbol{y}$$

for all $\boldsymbol{x}, \boldsymbol{y}$ in the domains.

Suppose that $\mathcal{D}$ is a nonempty open convex set and the convex function f is differentiable, thus continuously differentiable (or smooth) [106]. Given $\bar{\boldsymbol{y}}$, any $\bar{\boldsymbol{x}}$ that maximizes $\boldsymbol{x}^T \bar{\boldsymbol{y}} - f(\boldsymbol{x})$ satisfies $\bar{\boldsymbol{y}} = \nabla f(\bar{\boldsymbol{x}})$, and, conversely, if $\bar{\boldsymbol{x}}$ satisfies

$\bar{\boldsymbol{y}} = \nabla f(\bar{\boldsymbol{x}})$, then $\bar{\boldsymbol{x}}$ maximizes $\boldsymbol{x}^T\bar{\boldsymbol{y}} - f(\boldsymbol{x})$. Thus, the following equivalence always holds

$$\boldsymbol{y} = \nabla f(\boldsymbol{x}) \Longleftrightarrow f(\boldsymbol{x}) + f^*(\boldsymbol{y}) = \boldsymbol{x}^T\boldsymbol{y}. \tag{2.1}$$

In the case when f is differentiable, the conjugate is closely related to the Legendre conjugate. Let $(\nabla f)^{-1}$ denote the inverse mapping of ∇f, defined by

$$(\nabla f)^{-1}(\boldsymbol{y}) = \{\boldsymbol{x} \mid \boldsymbol{y} = \nabla f(\boldsymbol{x})\}.$$

Let $\mathcal{C}$ be the image of $\mathcal{D}$ under the gradient mapping ∇f. The *Legendre conjugate* is defined as

$$g(\boldsymbol{y}) = \boldsymbol{y}^T(\nabla f)^{-1}(\boldsymbol{y}) - f((\nabla f)^{-1}(\boldsymbol{y})).$$

For a general differentiable convex function, $(\nabla f)^{-1}(\boldsymbol{y})$ may contain more than one elements for some $\boldsymbol{y}$, and thus the gradient mapping ∇f is not one-to-one. However, here we can show that the Legendre conjugate $g(\boldsymbol{y})$ is always well-defined (i.e., single-valued) under the assumption above. For a given $\boldsymbol{y}$, no matter which $\boldsymbol{x}$ we choose in $(\nabla f)^{-1}(\boldsymbol{y})$, by the relationship (2.1) we get the same value $\boldsymbol{y}^T\boldsymbol{x} - f(\boldsymbol{x}) = f^*(\boldsymbol{y})$. Thus, g is the restriction of f^* to $\mathcal{C}$. The process of passing from f to the Legendre conjugate g is referred to as *Legendre transformation*.

In general, the conjugate function f^* need not be differentiable. One of the corner-stone results in convex analysis states that the differentiability dualizes under the Legendre-Fenchel transformation to the strict convexity: the conjugate function f^* is *essentially smooth* if and only if f is strictly convex (Theorem 26.3 in [106]). Let $\mathcal{C} = \text{int}(\text{dom} f^*)$. Here, function f^* is essentially smooth if (1) it is differentiable throughout $\mathcal{C}$, which should not be empty, and (2) $\lim_{i\to\infty} \|\nabla f^*(\boldsymbol{x}_i)\| = +\infty$ whenever $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$, is a sequence in $\mathcal{C}$ converging to a boundary point $\boldsymbol{x}$ of $\mathcal{C}$. Notice that if $\mathcal{C} = \mathbb{R}^n$ a smooth convex function is essentially smooth. When f^* is differentiable, by the equivalence (2.1), we have

$$\nabla f^*(\bar{\boldsymbol{y}}) = \bar{\boldsymbol{x}}$$

for any $\bar{\boldsymbol{y}}$ and $\bar{\boldsymbol{x}}$ satisfying $\bar{\boldsymbol{y}} = \nabla f(\bar{\boldsymbol{x}})$. Thus, $\bar{\boldsymbol{y}}$ maximizes $\bar{\boldsymbol{x}}^T\boldsymbol{y} - f^*(\boldsymbol{y})$, which implies that

$$f^{**}(\bar{\boldsymbol{x}}) = \sup_{\boldsymbol{y}\in\text{dom} f^*} \{\bar{\boldsymbol{x}}^T\boldsymbol{y} - f^*(\boldsymbol{y})\} = \bar{\boldsymbol{x}}^T\bar{\boldsymbol{y}} - f^*(\bar{\boldsymbol{y}}) = f(\bar{\boldsymbol{x}}). \tag{2.2}$$

Conversely, if the last equality in (2.2) holds, then it is true that $\nabla f^*(\bar{\boldsymbol{y}}) = \bar{\boldsymbol{x}}$. Therefore, we have

$$\boldsymbol{y} = \nabla f(\boldsymbol{x}) \Longleftrightarrow \boldsymbol{x} = \nabla f^*(\boldsymbol{y}) \Longleftrightarrow f(\boldsymbol{x}) + f^*(\boldsymbol{y}) = \boldsymbol{x}^T\boldsymbol{y}. \tag{2.3}$$

In addition, if $f(\boldsymbol{x})$ is twice differentiable, then the mapping ∇f is one-to-one if and only if the Hessian matrix $\nabla^2 f$ is nonsingular. Given $\boldsymbol{x}$ and $\boldsymbol{y}$ satisfying $\boldsymbol{y} = \nabla f(\boldsymbol{x})$, let $F(\boldsymbol{x}, \boldsymbol{y}) = \nabla f(\boldsymbol{x}) - \boldsymbol{y}$. If $\nabla^2 f(\boldsymbol{x})$ is nonsingular, which also means

that the Jacobian of $F(\boldsymbol{x}, \boldsymbol{y})$ is not zero, the equation $F(\boldsymbol{x}, \boldsymbol{y}) = 0$ uniquely defines a differentiable function $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{y})$ in the neighborhood of $(\boldsymbol{x}, \boldsymbol{y})$, which implies that $f^*(\boldsymbol{y}) = \boldsymbol{x}^T\boldsymbol{y} - f(\boldsymbol{x})$ is differentiable and we have $\nabla f^*(\boldsymbol{y}) = \boldsymbol{x}(\boldsymbol{y})$. Thus, ∇f^* is the inverse mapping of ∇f and is differentiable, and the conjugate function $f^*(\boldsymbol{y})$ is actually twice differentiable, with $\nabla^2 f^*(\boldsymbol{y}) = \nabla \boldsymbol{x}$. Taking derivative on both sides of $\boldsymbol{y} = \nabla f(\boldsymbol{x})$ with respect to $\boldsymbol{y}$, we have $I = \nabla \boldsymbol{x} \nabla^2 f(\boldsymbol{x})$. Therefore, by replacing $\nabla \boldsymbol{x}$ with $\nabla^2 f^*(\boldsymbol{y})$, it shows that the Hessian matrices of $f(\boldsymbol{x})$ and $f^*(\boldsymbol{x})$ are inverse to each other, i.e.,

$$\nabla^2 f^*(\boldsymbol{y}) = \left(\nabla^2 f(\boldsymbol{x})\right)^{-1}. \tag{2.4}$$

2.2 Optimization problems and optimality conditions

In this section, a general optimization problem is described and then optimality conditions are introduced. A special class of optimization problems, called convex optimization problems, is of particular interest. For the proof of the main results, refer to [12].

2.2.1 Problem statements

A *mathematical optimization problem*, or just *optimization problem*, is generally formulated as:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & \boldsymbol{g}(\boldsymbol{x}) \leq 0 \\ & \boldsymbol{h}(\boldsymbol{x}) = 0 \\ & \boldsymbol{x} \in \mathcal{X}_0, \end{aligned} \tag{2.5}$$

in which $\boldsymbol{x} = \{x_i\}_{i=1}^n \in \mathbb{R}^n$, $\boldsymbol{g}(\boldsymbol{x}) = \{g_i(\boldsymbol{x})\}_{i=1}^m$ and $\boldsymbol{h}(\boldsymbol{x}) = \{h_i(\boldsymbol{x})\}_{i=1}^l$. Here, $f(\boldsymbol{x})$, $g_i(\boldsymbol{x}), i = 1, \ldots, m$ and $h_i(\boldsymbol{x}), i = 1, \ldots, l$ are functions defined on $\mathbb{R}^n$, and $\mathcal{X}_0$ is a subset of $\mathbb{R}^n$.

The function f is called the *objective function* (also called *criterion function* or *cost function*). The restrictions $\boldsymbol{g}(\boldsymbol{x}) \leq 0$ and $\boldsymbol{h}(\boldsymbol{x}) = 0$ are called *constraints*, of which the former consists of *inequality constraints* and the latter consists of *equality constraints*. The set $\mathcal{X}_0$ might typically include lower and upper bounds on the variables, which even if implied by the other constraints can play a useful role in some algorithms. Alternatively, this set might represent some specially structured constraints that are highlighted to be exploited by the optimization routine, or it might represent certain regional containment or other complicating constraints that are to be handled separately via a special mechanism. A vector $\boldsymbol{x} \in \mathcal{X}_0$ satisfying all the constraints is called a *feasible solution* to the problem. The collection of

all such solutions forms the *feasible region*. A feasible solution that minimizes the objective function is called an *optimal solution* or simply a *solution*. If $\bar{\boldsymbol{x}}$ is an optimal solution, then we have $f(\boldsymbol{x}) \geq f(\bar{\boldsymbol{x}})$ for any feasible solution $\boldsymbol{x}$. If there exists an ε-neighborhood $\mathcal{N}_\varepsilon(\bar{\boldsymbol{x}})$ around $\bar{\boldsymbol{x}}$ such that $f(\bar{\boldsymbol{x}}) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{N}_\varepsilon(\bar{\boldsymbol{x}})$, then $\bar{\boldsymbol{x}}$ is called a *local optimal solution* or simply a *local solution*; while if $f(\bar{\boldsymbol{x}}) < f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{N}_\varepsilon(\bar{\boldsymbol{x}})$ and $\boldsymbol{x} \neq \bar{\boldsymbol{x}}$, $\bar{\boldsymbol{x}}$ is called a *strict local solution*. In contrast to local solutions, an optimal solution is also called a *global optimal solution* or simply a *global solution*. Clearly, a global solution is also a local solution. In this context (a minimization problem), a global solution and a local solution can also be, respectively, called a *global minimizer* and a *local minimizer*.

Optimization problems can be categorized into families or classes, by characterizing particular forms of the target and constraint functions. The optimization problem (2.5) is called a *linear program* if the objective function f and constraint functions $g_i, i = 1, \ldots, m$ and $h_i, i = 1, \ldots, l$ are linear and $\mathcal{X}_0$ is a polyhedral set. If the optimization problem is not linear, it is called a *nonlinear program*. From the point of view of convexity, optimization problems can be grouped into *convex optimization problems* and *nonconvex optimization problems*. The problem (2.5) is called a convex optimization problem, if f and $g_i, i = 1, \ldots, m$ are convex, $h_i, i = 1, \ldots, l$ are affine, that is, $\boldsymbol{h}(\boldsymbol{x}) = A\boldsymbol{x} - \boldsymbol{b}$, and $\mathcal{X}_0$ is a convex set; otherwise, it is a nonconvex optimization problem. The standard form of convex optimization problems is

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & \boldsymbol{g}(\boldsymbol{x}) \leq 0 \\ & A\boldsymbol{x} = \boldsymbol{b} \\ & \boldsymbol{x} \in \mathcal{X}_0. \end{aligned} \tag{2.6}$$

2.2.2 Optimality conditions

We assume that $\mathcal{X}_0$ is a nonempty open set, f and $g_i, i = 1, \ldots, m$ are differentiable and $h_i, i = 1, \ldots, l$ are continuously differentiable.

Unconstrained problems

A vector $\boldsymbol{d}$ is called a *descent direction* of $f : \mathbb{R}^n \to \mathbb{R}$ at $\boldsymbol{x}$ if there exists a $\delta > 0$ such that $f(\boldsymbol{x} + \lambda \boldsymbol{d}) < f(\boldsymbol{x})$ for all $\lambda \in (0, \delta)$. It can be proved that if $\nabla f(\boldsymbol{x})^T \boldsymbol{d} < 0$, $\boldsymbol{d}$ is a descent direction of f at $\boldsymbol{x}$; conversely, if $\boldsymbol{d}$ is a descent direction, $\nabla f(\boldsymbol{x})^T \boldsymbol{d} \leq 0$. Thus, for a differentiable function f, if $\bar{\boldsymbol{x}}$ is a local minimizer, we have

$$\nabla f(\bar{\boldsymbol{x}}) = 0.$$

In addition, if f is twice differentiable, the Hessian matrix being positive semidefinite at the point is another necessary condition. Then, for a twice differentiable function f, if $\bar{\boldsymbol{x}}$ is a local minimizer, we have

$$\nabla f(\bar{\boldsymbol{x}}) = 0 \text{ and } \nabla^2 f(\bar{\boldsymbol{x}}) \succeq 0.$$

However, these conditions are only necessary conditions, for if $\nabla^2 f(\bar{\boldsymbol{x}}) = 0$, the point could be neither a local minimizer nor a local maximum (such a point is called a *saddle point*). If $\succeq$ is replaced with $\succ$ in the conditions, they become sufficient conditions for the point being a strict local minimizer. While, if the function f is convex, the condition $\nabla f(\bar{\boldsymbol{x}}) = 0$ alone is sufficient to guarantee that $\bar{\boldsymbol{x}}$ is a local minimizer, and actually, the point $\bar{\boldsymbol{x}}$ is also a global minimizer.

KKT conditions

First we consider the case where there are only inequality constraints. Let $\mathcal{X}_g = \{\boldsymbol{x} \in \mathcal{X}_0 \,|\, \boldsymbol{g}(\boldsymbol{x}) \leq 0\}$ be the feasible region. Let

$$\mathcal{F} = \{\boldsymbol{d} \,|\, \boldsymbol{d} \neq 0, \bar{\boldsymbol{x}} + \lambda \boldsymbol{d} \in \mathcal{X}_g \text{ for all } \lambda \in (0, \delta) \text{ for some } \delta > 0\}$$

be the *cone of feasible directions* of $\mathcal{X}_g$ at a point $\bar{\boldsymbol{x}} \in \mathcal{X}_g$, and

$$\mathcal{D} = \{\boldsymbol{d} \,|\, f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) < f(\bar{\boldsymbol{x}}) \text{ for all } \lambda \in (0, \delta) \text{ for some } \delta > 0\}$$

be the *cone of improving directions* of f at a point $\bar{\boldsymbol{x}} \in \mathcal{X}_g$. Then, if $\bar{\boldsymbol{x}}$ is a local minimizer, it is obvious that $\mathcal{F} \cap \mathcal{D} = \emptyset$, i.e., all the feasible directions will not be improving or descent directions.

Let $I = \{i \,|\, g_i(\bar{\boldsymbol{x}}) = 0\}$ be the index set for the *active* (or *binding* or *tight*) *constraints*. As any vector $\boldsymbol{d}$ satisfying $\nabla g_i(\bar{\boldsymbol{x}})^T \boldsymbol{d} < 0$ is a descent direction of g_i at $\bar{\boldsymbol{x}}$, if let

$$\mathcal{F}_0 = \{\boldsymbol{d} \,|\, \nabla g_i(\bar{\boldsymbol{x}})^T \boldsymbol{d} < 0, i = 1, \ldots, m\},$$

we have $\mathcal{F}_0 \subseteq \mathcal{F}$, and similarly we have $\mathcal{D}_0 \subseteq \mathcal{D}$ where

$$\mathcal{D}_0 = \{\boldsymbol{d} \,|\, \nabla f(\bar{\boldsymbol{x}})^T \boldsymbol{d} < 0\}.$$

Hence, $\mathcal{F}_0 \cap \mathcal{D}_0 = \emptyset$ is a necessary condition for the point $\bar{\boldsymbol{x}}$ being a local minimizer. By the result of separation of two convex sets, $\mathcal{F}_0 \cap \mathcal{D}_0 = \emptyset$ implies that there exist $u_0, u_i, i \in I$ such that

$$u_0 \nabla f(\bar{\boldsymbol{x}}) + \sum_{i \in I} u_i \nabla g_i(\bar{\boldsymbol{x}}) = 0 \tag{2.7}$$

$$(u_0, \boldsymbol{u}_I) \geq 0 \tag{2.8}$$

$$(u_0, \boldsymbol{u}_I) \neq 0, \tag{2.9}$$

where $\boldsymbol{u}_I = \{u_i\}_{i \in I}$.

Conditions (2.7-2.9), together with the feasibility condition, $\bar{\boldsymbol{x}} \in \mathcal{X}_g$, are called FJ (short for Fritz John) conditions, and points satisfying FJ conditions are called FJ points. It is shown above that FJ conditions are necessary conditions of a point being a local minimizer. However, FJ conditions may be trivial: each point with $\nabla g_i(\bar{\boldsymbol{x}}), i \in I$ being linearly dependent is an FJ point, and in some instances, each feasible solution could be an FJ point (see examples in [12]). If $\nabla g_i(\bar{\boldsymbol{x}}), i \in I$ are linearly

independent, it must be true that $u_0 > 0$, which leads to KKT (short for Karush-Kuhn-Tucker) conditions. The KKT conditions are precisely the FJ conditions with the added requirement that $u_0 > 0$, and they encompass FJ points for which there exist values $(u_0, \boldsymbol{u}_I)$ such that $u_0 > 0$ and hence force the gradient of objective function to play a role in the optimality conditions.

When $\nabla g_i(\bar{\boldsymbol{x}}), i \in I$ are linearly independent, without loss of generality, we can let u_0 be equal to one in the conditions (2.7-2.9), which then becomes

$$\begin{aligned} \nabla f(\bar{\boldsymbol{x}}) + \sum_{i \in I} u_i \nabla g_i(\bar{\boldsymbol{x}}) = 0 \\ u_i \geq 0, i \in I \end{aligned}$$

As g_i for $i \notin I$ are also differentiable at $\bar{\boldsymbol{x}}$, the forgoing conditions can also be written as

$$\begin{aligned} \nabla f(\bar{\boldsymbol{x}}) + \sum_{i=1}^{m} u_i \nabla g_i(\bar{\boldsymbol{x}}) = 0 \\ u_i g_i(\bar{\boldsymbol{x}}) = 0, i = 1, \ldots, m \\ u_i \geq 0, i = 1, \ldots, m. \end{aligned}$$

Here, for $i \notin I$, as $g_i(\bar{\boldsymbol{x}}) < 0$, the condition in the second equation will force u_i to be zero, which results in disappearance of $u_i \nabla g_i(\bar{\boldsymbol{x}})$ for $i \notin I$ in the first equation. These conditions, together with the feasibility condition, are called *KKT conditions*:

$$\boldsymbol{g}(\bar{\boldsymbol{x}}) \leq 0 \tag{2.10}$$

$$\bar{\boldsymbol{x}} \in \mathcal{X}_0 \tag{2.11}$$

$$u_i g_i(\bar{\boldsymbol{x}}) = 0, i = 1, \ldots, m \tag{2.12}$$

$$\boldsymbol{u} \geq 0 \tag{2.13}$$

$$\nabla f(\bar{\boldsymbol{x}}) + \nabla \boldsymbol{g}(\bar{\boldsymbol{x}}) \boldsymbol{u} = 0, \tag{2.14}$$

in which $\boldsymbol{u} = \{u_i\}_{i=1}^m$ and $\nabla \boldsymbol{g}(\bar{\boldsymbol{x}}) = (\nabla g_1(\bar{\boldsymbol{x}}), \ldots, \nabla g_m(\bar{\boldsymbol{x}})) \in \mathbb{R}^{n \times m}$. The scalars u_i are called the *Lagrangian* (or *Lagrange*) multipliers. The requirement that $g_i(\bar{\boldsymbol{x}}) \leq 0, i = 1, \ldots, m, \bar{\boldsymbol{x}} \in \mathcal{X}_0$ is called the *primal feasibility condition*, whereas the condition $\nabla f(\bar{\boldsymbol{x}}) + \sum_{i=1}^m u_i \nabla g_i(\bar{\boldsymbol{x}}) = 0, u_i \geq 0, i = 1, \ldots, m$ is referred to as the *dual feasibility condition*. The restriction $u_i g_i(\bar{\boldsymbol{x}}) = 0, i = 1, \ldots, m$ is called the *complementary slackness condition*. Any point $\bar{\boldsymbol{x}}$ meeting the KKT conditions, i.e., there exist Lagrangian multipliers such that conditions (2.10-2.14) hold, is called a *KKT point*.

It should be pointed out that when $\mathcal{X}_0$ is an open set and the requirement is met that $\nabla g_i(\bar{\boldsymbol{x}}), i \in I$ are linearly independent, KKT conditions are necessary local optimality conditions. If the linear independence does not hold true at a local solution, there may not exist multipliers such that conditions (2.12-2.14) are satisfied and hence it is not a KKT point.

Next, consider the general problem with both inequality and equality constraints. KKT conditions can be written as:

$$\boldsymbol{g}(\bar{\boldsymbol{x}}) \leq 0 \tag{2.15}$$
$$\boldsymbol{h}(\bar{\boldsymbol{x}}) = 0 \tag{2.16}$$
$$\bar{\boldsymbol{x}} \in \mathcal{X}_0 \tag{2.17}$$
$$u_i g_i(\bar{\boldsymbol{x}}) = 0, i = 1, \ldots, m \tag{2.18}$$
$$\boldsymbol{u} \geq 0 \tag{2.19}$$
$$\nabla f(\bar{\boldsymbol{x}}) + \nabla \boldsymbol{g}(\bar{\boldsymbol{x}})\boldsymbol{u} + \nabla \boldsymbol{h}(\bar{\boldsymbol{x}})\boldsymbol{v} = 0, \tag{2.20}$$

where $\boldsymbol{v} = \{v_i\}_{i=1}^{l}$ and $\nabla \boldsymbol{h}(\bar{\boldsymbol{x}}) = (\nabla h_1(\bar{\boldsymbol{x}}), \ldots, \nabla h_l(\bar{\boldsymbol{x}})) \in \mathbb{R}^{n \times l}$.

Similarly, for the general problem, KKT conditions may not be necessary for a point being a local solution. There are several requirements under which the KKT conditions become necessary for a point being a local solution. These requirements are referred to as *constraint qualifications*. The requirement that $\nabla g_i(\bar{\boldsymbol{x}}), i \in I$ are linearly independent is called *linear independence constraint qualification* for the problem with only inequality constraints. For the general problem, with both inequality and equality constraints, the linear independence constraint qualification is that

$$\nabla g_i(\bar{\boldsymbol{x}}), i \in I \text{ and } \nabla h_i(\bar{\boldsymbol{x}}), i = 1, \ldots, l \text{ are linearly independent.}$$

For the convex problem (2.6), a widely known constraint qualification is the one called *Slater's constraint qualification*: there exists a feasible solution such that

$$\boldsymbol{g}(\boldsymbol{x}) < 0.$$

Under certain convexity assumptions on f, g_i and h_i, the KKT conditions will be sufficient local optimality conditions. Let $\bar{\boldsymbol{x}}$ be a KKT point. If $h_i, i = 1, \ldots, l$ are affine functions and f and $g_i, i \in I$ are convex in the neighbourhood $\mathcal{N}_\varepsilon(\bar{\boldsymbol{x}})$ for some $\varepsilon > 0$, then $\bar{\boldsymbol{x}}$ is a local minimizer. Immediately, for the convex problem (2.6), any KKT point is a local minimizer and, actually, a global minimizer. Thus, when the Slater's constraint qualification holds true, the KKT conditions become necessary and sufficient for a point being an optimal solution to the convex problem (2.6); that is, $\bar{\boldsymbol{x}}$ is an optimal solution of (2.6) if and only if there exists $(\boldsymbol{u}, \boldsymbol{v})$ such that conditions (2.15-2.20) hold true.

Saddle point optimality conditions

The *Lagrangian function*[1] for the problem (2.5) is defined as

$$L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = f(\boldsymbol{x}) + \boldsymbol{u}^T \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{v}^T \boldsymbol{h}(\boldsymbol{x}).$$

[1]It should be noticed that the original definition of Lagrangian function comes from Lagrange's work on applying the principle of stationary action to classical mechanics [80].

A solution $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ is called a *saddle point* of the Lagrangian function if $\bar{\boldsymbol{x}} \in \mathcal{X}_0$, $\bar{\boldsymbol{u}} \geq 0$ and

$$L(\bar{\boldsymbol{x}}, \boldsymbol{u}, \boldsymbol{v}) \leq L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) \leq L(\boldsymbol{x}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) \tag{2.21}$$

for all $\boldsymbol{x} \in \mathcal{X}_0$ and all $(\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}^{m+l}$ with $\boldsymbol{u} \geq 0$. The equation (2.21) implies that $\bar{\boldsymbol{x}}$ minimises $L(\boldsymbol{x}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ over $\mathcal{X}_0$, i.e.,

$$L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) = \min \left\{ L(\boldsymbol{x}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) \mid \boldsymbol{x} \in \mathcal{X}_0 \right\}, \tag{2.22}$$

and $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ maximises $L(\bar{\boldsymbol{x}}, \boldsymbol{u}, \boldsymbol{v})$ over all $(\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}^{m+l}$ with $\boldsymbol{u} \geq 0$, i.e.,

$$L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) = \max \left\{ L(\bar{\boldsymbol{x}}, \boldsymbol{u}, \boldsymbol{v}) \mid (\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}^{m+l}, \boldsymbol{u} \geq 0 \right\}. \tag{2.23}$$

If $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ is a saddle point of the Lagrangian function, then $\bar{\boldsymbol{x}}$ is a global solution of the problem (2.5). From the primal feasible condition and complementary slackness condition, we have

$$f(\bar{\boldsymbol{x}}) = L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}).$$

Moreover, since $\boldsymbol{h}(\boldsymbol{x}) = 0$ and $\bar{\boldsymbol{u}}^T \boldsymbol{g}(\boldsymbol{x}) \leq 0$ hold true for each feasible solution $\boldsymbol{x}$, we have

$$L(\boldsymbol{x}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) = f(\boldsymbol{x}) + \bar{\boldsymbol{u}}^T \boldsymbol{g}(\boldsymbol{x}) + \bar{\boldsymbol{v}}^T \boldsymbol{h}(\boldsymbol{x}) \leq f(\boldsymbol{x}).$$

By the equation (2.22), it is proved that $f(\bar{\boldsymbol{x}}) \leq f(\boldsymbol{x})$ for all feasible solutions $\boldsymbol{x}$, and thus $\bar{\boldsymbol{x}}$ is a global solution.

A saddle point is always a KKT point. If $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ is a saddle point, it is clear that we must have $\boldsymbol{g}(\bar{\boldsymbol{x}}) \leq 0$ and $\boldsymbol{h}(\bar{\boldsymbol{x}}) = 0$, otherwise the right side of the equation (2.23) will be $+\infty$. From the inequality

$$f(\bar{\boldsymbol{x}}) + \bar{\boldsymbol{u}}^T \boldsymbol{g}(\bar{\boldsymbol{x}}) + \bar{\boldsymbol{v}}^T \boldsymbol{h}(\bar{\boldsymbol{x}}) \geq f(\bar{\boldsymbol{x}}) + \boldsymbol{u}^T \boldsymbol{g}(\bar{\boldsymbol{x}}) + \boldsymbol{v}^T \boldsymbol{h}(\bar{\boldsymbol{x}}),$$

we have $\bar{\boldsymbol{u}}^T \boldsymbol{g}(\bar{\boldsymbol{x}}) \geq \boldsymbol{u}^T \boldsymbol{g}(\bar{\boldsymbol{x}})$. If let $\boldsymbol{u} = 0$, we then have $\bar{\boldsymbol{u}}^T \boldsymbol{g}(\bar{\boldsymbol{x}}) \geq 0$. But the feasibility of $\bar{\boldsymbol{x}}$ shows that $\bar{\boldsymbol{u}}^T \boldsymbol{g}(\bar{\boldsymbol{x}}) \leq 0$. Thus, it must be true that $\bar{u}_i g_i(\bar{\boldsymbol{x}}) = 0, i = 1, \ldots, m$, which is the complementary slackness condition. As $\mathcal{X}_0$ is supposed to be an open set, the minimiser $\bar{\boldsymbol{x}}$ is a stationary point of $L(\boldsymbol{x}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$, that is,

$$\nabla_{\boldsymbol{x}} L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) = \nabla f(\bar{\boldsymbol{x}}) + \nabla \boldsymbol{g}(\bar{\boldsymbol{x}}) \bar{\boldsymbol{u}} + \nabla \boldsymbol{h}(\bar{\boldsymbol{x}}) \bar{\boldsymbol{v}} = 0.$$

Therefore, $\bar{\boldsymbol{x}}$ is a KKT point and $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ are the corresponding Lagrangian multipliers.

However, a KKT point is generally not a saddle point, since it may not even a local minimizer. Suppose $\bar{\boldsymbol{x}}$ is a KKT point and $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ are the corresponding Lagrangian multipliers. Then, if $\bar{\boldsymbol{x}}$ solves the minimization problem in (2.22), $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ is a saddle point. It is obvious that if $L(\boldsymbol{x}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ is a convex function with respect to $\boldsymbol{x}$, $\bar{\boldsymbol{x}}$ is a minimizer of $L(\boldsymbol{x}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ since $\nabla_{\boldsymbol{x}} L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) = 0$. Hence, for the convex problem (2.6), any KKT points are saddle points. Therefore, the KKT conditions and saddle point optimality conditions are equivalent for the convex problem (2.6), which further implies that when the Slater's condition holds true, $\bar{\boldsymbol{x}}$ is a global solution of (2.6) if and only if there exist $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ such that $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ is a saddle point for the Lagrangian function.

2.3 Lagrangian duality and convex optimization problems

Given the problem (2.5), several closely related problems can be derived, which are called *dual problems*. The original problem can be solved indirectly by solving its dual problem and the latter may be easier than the former. Among the various duality formulations, the Lagrangian duality formulation is most well-known. It has been successfully applied to deal with convex and nonconvex problems and also discrete optimization problems.

2.3.1 Lagrangian duality

Let

$$d(\boldsymbol{u}, \boldsymbol{v}) = \inf\{L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) \mid \boldsymbol{x} \in \mathcal{X}_0\}.$$

The *Lagrangian dual problem* for the problem (2.5) is defined as

$$\begin{aligned} &\sup_{\boldsymbol{u},\boldsymbol{v}} d(\boldsymbol{u}, \boldsymbol{v}) \\ &\text{s.t. } \boldsymbol{u} \geq 0. \end{aligned} \tag{2.24}$$

Accordingly, the problem (2.5) is called the *primal problem*. The function $d(\boldsymbol{u}, \boldsymbol{v})$ is called the *Lagrangian dual function*, and the optimization problem that evaluates $d(\boldsymbol{u}, \boldsymbol{v})$ is sometimes referred to as the *Lagrangian dual subproblem*. The Lagrangian multipliers $\boldsymbol{u}$ and $\boldsymbol{v}$ are also called *dual variables*. When the supremum is achievable, we can replace sup with max.

Notice that it is the set $\mathcal{X}_0$ where the infimum is taken for the Lagrangian function. Given an optimization problem, different Lagrangian dual functions can be defined by handling differently the constraints, i.e., which constraints are treated as $\boldsymbol{g}(\boldsymbol{x}) \leq 0$ and $\boldsymbol{h}(\boldsymbol{x}) = 0$ and which constraints are included in the set $\mathcal{X}_0$. This choice can affect both the optimal value of (2.24) and the effort expended in evaluating and updating the dual function during the course of solving the dual problem. Hence, an appropriate selection of the set $\mathcal{X}_0$ must be made, depending on the structure of the problem and the purpose for solving (2.24).

There are significant relations between the primal and dual problems. For any feasible solution $\boldsymbol{x}$ to the primal problem (2.5) and any feasible solution $(\boldsymbol{u}, \boldsymbol{v})$ to the dual problem (2.24), we have

$$d(\boldsymbol{u}, \boldsymbol{v}) \leq f(\boldsymbol{x}) + \boldsymbol{u}^T \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{v}^T \boldsymbol{h}(\boldsymbol{x}) \leq f(\boldsymbol{x}),$$

where the second inequality results from $\boldsymbol{h}(\boldsymbol{x}) = 0$ and $\boldsymbol{u}^T \boldsymbol{g}(\boldsymbol{x}) \leq 0$. It shows that the objective value of any feasible solution to the dual problem yields a lower bound on the objective value of any feasible solution to the primal problem. This result is referred to as the *weak duality*. Immediately, we have

$$\sup\{d(\boldsymbol{u}, \boldsymbol{v}) \mid \boldsymbol{u} \geq 0\} \leq \inf\{f(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{X}_0, \boldsymbol{g}(\boldsymbol{x}) \leq 0, \boldsymbol{h}(\boldsymbol{x}) = 0\}. \tag{2.25}$$

Hence, if there exist $\bar{\boldsymbol{x}}$ and $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ such that $f(\bar{\boldsymbol{x}}) = d(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$, then $\bar{\boldsymbol{x}}$ and $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ are optimal solutions to (2.5) and (2.24), respectively.

The Lagrangian dual function $d(\boldsymbol{u}, \boldsymbol{v})$ may assume the value of $-\infty$ for some vectors $(\boldsymbol{u}, \boldsymbol{v})$. For example, if $\inf\{f(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{X}_0, \boldsymbol{g}(\boldsymbol{x}) \leq 0, \boldsymbol{h}(\boldsymbol{x}) = 0\} = -\infty$, then $d(\boldsymbol{u}, \boldsymbol{v}) = -\infty$ for each $\boldsymbol{u} \geq 0$. Whereas, if $\sup\{d(\boldsymbol{u}, \boldsymbol{v}) \mid \boldsymbol{u} \geq 0\} = \infty$, then the primal problem has no feasible solution, since $d(\boldsymbol{u}, \boldsymbol{v})$ is upper bounded by $f(\boldsymbol{x})$.

If the strict inequality in (2.25) holds true, there is a *duality gap* between the primal and dual problems. Only when there is no duality gap, the primal problem can be solved by solving the dual problem; otherwise, the dual problem is only able to provide a lower bound for the primal problem and it is called the *Lagrangian relaxation problem*. If the equality in (2.25) holds, i.e.,

$$\sup\{d(\boldsymbol{u}, \boldsymbol{v}) \mid \boldsymbol{u} \geq 0\} = \inf\{f(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{X}_0, \boldsymbol{g}(\boldsymbol{x}) \leq 0, \boldsymbol{h}(\boldsymbol{x}) = 0\}, \tag{2.26}$$

then we say that *strong duality* holds.

The strong duality generally does not hold for the problem (2.5). Whereas, for the convex problem (2.6), we usually have strong duality, but not always. Suppose $\mathcal{X}_0$ is a nonempty open set. Then, for the convex problem (2.6), $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ satisfies the KKT conditions if and only if $\bar{\boldsymbol{x}}$ and $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ are optimal solutions to the primal and dual problems and $f(\bar{\boldsymbol{x}}) = d(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$. The proof is direct and simple. If $\bar{\boldsymbol{x}}$ is a KKT points with Lagrangian multipliers $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$, then $\bar{\boldsymbol{x}}$ must be a minimiser of $L(\boldsymbol{x}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ over $\mathcal{X}_0$, because $L(\boldsymbol{x}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ is a convex function with respect to $\boldsymbol{x}$ and the dual feasible conditions implies that $\nabla_{\boldsymbol{x}} L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) = 0$. Thus, we have

$$f(\bar{\boldsymbol{x}}) = f(\bar{\boldsymbol{x}}) + \bar{\boldsymbol{u}}^T \boldsymbol{g}(\bar{\boldsymbol{x}}) + \bar{\boldsymbol{v}}^T \boldsymbol{h}(\bar{\boldsymbol{x}}) = \min\{L(\boldsymbol{x}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) \mid \boldsymbol{x} \in \mathcal{X}_0\} = d(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}),$$

that is, $\bar{\boldsymbol{x}}$ and $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ are optimal solutions to the primal and dual problems with no duality gap. Conversely, from the strong duality $f(\bar{\boldsymbol{x}}) = d(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$, equality holds true through the following

$$\begin{aligned} d(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) &= \inf\{L(\boldsymbol{x}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) \mid \boldsymbol{x} \in \mathcal{X}_0\} \\ &\leq f(\bar{\boldsymbol{x}}) + \bar{\boldsymbol{u}}^T \boldsymbol{g}(\bar{\boldsymbol{x}}) + \bar{\boldsymbol{v}}^T \boldsymbol{h}(\bar{\boldsymbol{x}}) \\ &= f(\bar{\boldsymbol{x}}) + \bar{\boldsymbol{u}}^T \boldsymbol{g}(\bar{\boldsymbol{x}}) \\ &\leq f(\bar{\boldsymbol{x}}). \end{aligned}$$

In particular, we have $\bar{\boldsymbol{u}}^T \boldsymbol{g}(\bar{\boldsymbol{x}}) = 0$ and $\bar{\boldsymbol{x}}$ is a minimizer of $L(\boldsymbol{x}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ in the open set $\mathcal{X}_0$, which implies $\nabla_{\boldsymbol{x}} L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) = 0$. Hence, $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ satisfies the KKT conditions.

As discussed in the previous section, for the convex problem (2.6), the KKT conditions are sufficient conditions for the point being a global solution, but not necessary conditions. When any of the constraint qualifications holds true, a global solution $\bar{\boldsymbol{x}}$ must be also a KKT point, and then there exist Lagrangian multipliers $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ such that $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ solves the dual problem and there is no duality gap. For example, if the Slater's constraint qualification holds true, i.e., there is a point $\boldsymbol{x} \in \mathcal{X}_0$ such that

$$\boldsymbol{g}(\boldsymbol{x}) < 0, \text{ and } A\boldsymbol{x} = \boldsymbol{b},$$

we have the strong duality.

For the convex problem (2.6) with a nonempty open set $\mathcal{X}_0$, the strong duality with the infimum in (2.26) being achievable is also equivalent to the existence of a saddle point: $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ is a saddle point of the Lagrangian function if and only if $\bar{\boldsymbol{x}}$ and $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ are optimal solutions to the primal and dual problems with no duality gap. It is because $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ being a saddle point is equivalent to it satisfying the KKT conditions. The equivalence can also be seen in another way. The following equality holds true

$$\inf\{f(\boldsymbol{x}) \mid \boldsymbol{g}(\boldsymbol{x}) \leq 0, \boldsymbol{h}(\boldsymbol{x}) = 0, \boldsymbol{x} \in \mathcal{X}_0\} = \inf_{\boldsymbol{x} \in \mathcal{X}_0} \sup_{(\boldsymbol{u}, \boldsymbol{v}), \boldsymbol{u} \geq 0} L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}),$$

because the supremum of $L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})$ over $(\boldsymbol{u}, \boldsymbol{v})$ with $\boldsymbol{u} \geq 0$ will be infinity if $\boldsymbol{x}$ does not satisfy $\boldsymbol{g}(\boldsymbol{x}) \leq 0$ and $\boldsymbol{h}(\boldsymbol{x}) = 0$. Then, the strong duality (2.26) can also be expressed as

$$\sup_{(\boldsymbol{u}, \boldsymbol{v}), \boldsymbol{u} \geq 0} \inf_{\boldsymbol{x} \in \mathcal{X}_0} L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = \inf_{\boldsymbol{x} \in \mathcal{X}_0} \sup_{(\boldsymbol{u}, \boldsymbol{v}), \boldsymbol{u} \geq 0} L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}).$$

Thus, if the infimum and supremum are achievable, there must be a saddle point.

2.3.2 Convex optimization problems

Some broadly known convex optimization problems and their Lagrangian dual problems are presented.

Linear program

The *linear program* (LP) is normally written in the following *standard form*:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \boldsymbol{c}^T \boldsymbol{x} \\ \text{s.t.} \quad & A\boldsymbol{x} = \boldsymbol{b} \\ & \boldsymbol{x} \geq 0, \end{aligned} \tag{2.27}$$

where $A \in \mathbb{R}^{l \times n}$, $\boldsymbol{c} \in \mathbb{R}^n$ and $\boldsymbol{b} \in \mathbb{R}^l$. The problem has inequality constraints $\boldsymbol{g}(\boldsymbol{x}) = -\boldsymbol{x} \leq 0$ and equality constraints $\boldsymbol{h}(\boldsymbol{x}) = A\boldsymbol{x} - b$, with $\mathcal{X}_0 = \mathbb{R}^n$.

The Lagrangian function for the problem (2.27) is

$$L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{c}^T \boldsymbol{x} - \boldsymbol{u}^T \boldsymbol{x} + \boldsymbol{v}^T (A\boldsymbol{x} - \boldsymbol{b}),$$

from which we obtain the dual function,

$$d(\boldsymbol{u}, \boldsymbol{v}) = \inf_{\boldsymbol{x} \in \mathbb{R}^n} L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = \begin{cases} -\boldsymbol{v}^T \boldsymbol{b} & A^T \boldsymbol{v} - \boldsymbol{u} + \boldsymbol{c} = 0 \\ -\infty & \text{otherwise} \end{cases}$$

The Lagrangian dual problem of the standard LP is then formulated as

$$\begin{aligned} \max_{\boldsymbol{u},\boldsymbol{v}} \quad & -\boldsymbol{v}^T\boldsymbol{b} \\ \text{s.t.} \quad & A^T\boldsymbol{v} - \boldsymbol{u} + \boldsymbol{c} = 0 \\ & \boldsymbol{u} \geq 0. \end{aligned} \tag{2.28}$$

The dual variable $\boldsymbol{u}$ can be omitted, and the equality constraints will become inequality constraints. The problem (2.28) is then equivalent to

$$\begin{aligned} \max_{\boldsymbol{v}} \quad & -\boldsymbol{v}^T\boldsymbol{b} \\ \text{s.t.} \quad & A^T\boldsymbol{v} + \boldsymbol{c} \geq 0, \end{aligned} \tag{2.29}$$

which is an LP in inequality form. The strong duality between (2.27) and (2.29) always holds true.

Conversely, if the problem (2.29) is treated as the primal problem, then its dual problem is exactly the problem (2.27). It can be verified by rewriting the problem (2.29) into an equivalent minimization problem and then applying the procedure of constructing the Lagrangian dual problem.

Quadratic program

The problem (P_0) is called a *quadratic program* (QP), if the objective function is a quadratic function, and the constraint functions are affine. If the inequality constraints are also quadratic functions, the problem is then a *quadratically constrained quadratic problem* (QCQP):

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \frac{1}{2}\boldsymbol{x}^T Q_0 \boldsymbol{x} - \boldsymbol{x}^T \boldsymbol{f}_0 \\ \text{s.t.} \quad & \frac{1}{2}\boldsymbol{x}^T Q_i \boldsymbol{x} - \boldsymbol{x}^T \boldsymbol{f}_i \leq c_i,\ i = 1, \ldots, m \\ & A\boldsymbol{x} = \boldsymbol{b}, \end{aligned} \tag{2.30}$$

where $Q_i \in \mathbb{S}^n, i = 0, 1, \ldots, m$, $\boldsymbol{f}_i \in \mathbb{R}^n, i = 0, 1, \ldots, m$, $\boldsymbol{b} \in \mathbb{R}^l$, and $c_i \in \mathbb{R}, i = 1, \ldots, m$. Here, for a general QCQP, $Q_i, i = 0, 1, \ldots, m$ are not supposed to be positive semidefinite. While if $Q_i, i = 0, 1, \ldots, m$ are positive semidefinite, the target and inequality constraint functions are convex and the problem becomes a convex QCQP.

The Lagrangian function for the problem (2.30) is

$$L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = \frac{1}{2}\boldsymbol{x}^T G(\boldsymbol{u})\boldsymbol{x} - \boldsymbol{x}^T \boldsymbol{f}(\boldsymbol{u}) - \boldsymbol{u}^T\boldsymbol{c} - \boldsymbol{v}^T\boldsymbol{b}$$

where

$$G(\boldsymbol{u}) = Q_0 + \sum_{i=1}^m u_i Q_i, \text{ and } \boldsymbol{f}(\boldsymbol{u}) = \boldsymbol{f}_0 + \sum_{i=1}^m u_i \boldsymbol{f}_i - A^T\boldsymbol{v}.$$

Notice that given $(\boldsymbol{u}, \boldsymbol{v})$, the positive semidefiniteness of $Q_0 + \sum_{i=1}^m u_i Q_i$ is only necessary for $L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})$ being lower bounded; if $Q_0 + \sum_{i=1}^m u_i Q_i$ has eigenvalues of the value of 0, the sufficiency will be achieved by recruiting the additional condition that $\boldsymbol{f}_0 + \sum_{i=1}^m u_i \boldsymbol{f}_i - A^T \boldsymbol{v}$ is perpendicular to the subspace generated from the corresponding eigenvectors.

If $Q_0 \succ 0$, we have $Q_0 + \sum_{i=1}^m u_i Q_i \succ 0$ for any $\boldsymbol{u} \geq 0$ and thus the Lagrangian function is always lower bounded. The dual function then can be analytically defined and the dual problem can be formulated as

$$\begin{aligned} \max \quad & -\frac{1}{2}\boldsymbol{f}(\boldsymbol{u})^T G(\boldsymbol{u})^{-1} \boldsymbol{f}(\boldsymbol{u}) - \boldsymbol{u}^T \boldsymbol{c} - \boldsymbol{v}^T \boldsymbol{b} \\ \text{s.t.} \quad & \boldsymbol{u} \geq 0. \end{aligned} \tag{2.31}$$

We have the strong duality between (2.30) and (2.31) if the Slater's condition holds true.

Second-order cone program

A quadratic program is called a *second-order cone program* (SOCP) if it is of the form

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \boldsymbol{f}_0^T \boldsymbol{x} \\ \text{s.t.} \quad & \|Q_i \boldsymbol{x} + \boldsymbol{f}_i\| \leq \boldsymbol{c}_i^T \boldsymbol{x} + d_i, i = 1, \ldots, m \\ & A\boldsymbol{x} = \boldsymbol{b}, \end{aligned} \tag{2.32}$$

where $Q_i \in \mathbb{R}^{n_i \times n}$ and $A \in \mathbb{R}^{l \times n}$. Here, the matrices Q_i need not be symmetric. The inequality constraints in (2.32) require that $(Q_i \boldsymbol{x} + \boldsymbol{f}_i, \boldsymbol{c}_i^T \boldsymbol{x} + d_i) \in \mathcal{SOC}^n$, and they are referred to as *second-order cone constraints*. By the fact that each convex quadratic constraint can be written in the form of a second-order cone constraint with $\boldsymbol{c}_i = 0$, a convex QCQP can also be formulated as an SOCP.

Let

$$\boldsymbol{y}_i = Q_i \boldsymbol{x} + \boldsymbol{f}_i \text{ and } t_i = \boldsymbol{c}_i^T \boldsymbol{x} + d_i,$$

and place the constraints $\|\boldsymbol{y}_i\| \leq t_i$ into the set $\mathcal{X}_0$, that is,

$$\mathcal{X}_0 = \{(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{t}) \mid \boldsymbol{x} \in \mathbb{R}^n, \|\boldsymbol{y}_i\| \leq t_i, i = 1, \ldots, m\},$$

where $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)$ and $\boldsymbol{t} = \{t_i\}_{i=1}^m$. The Lagrangian function for the SOCP is

$$\begin{aligned} L(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{t}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = (\boldsymbol{f}_0 - \sum_{i=1}^m Q_i^T \boldsymbol{u}_i - \sum_{i=1}^m v_i \boldsymbol{c}_i + A^T \boldsymbol{w})^T \boldsymbol{x} + \sum_{i=1}^m (\boldsymbol{u}_i^T \boldsymbol{y}_i + v_i t_i) \\ - \sum_{i=1}^m \boldsymbol{u}_i^T \boldsymbol{f}_i - \boldsymbol{v}^T \boldsymbol{d} - \boldsymbol{w}^T \boldsymbol{b} \end{aligned}$$

where $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m)$, $\boldsymbol{v} = \{v_i\}_{i=1}^m$, $\boldsymbol{w} = \{w_i\}_{i=1}^l$ and $\boldsymbol{d} = \{d_i\}_{i=1}^m$. The dual function is lower bounded only when the first item in the Lagrangian function is zero and $(\boldsymbol{u}_i, v_i) \in \mathcal{SOC}^n$ for $i = 1, \ldots, m$. Thus, the dual problem for the problem (2.32) is

$$\begin{aligned} \max_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}} \quad & -\sum_{i=1}^m \boldsymbol{u}_i^T \boldsymbol{f}_i - \boldsymbol{v}^T \boldsymbol{d} - \boldsymbol{w}^T \boldsymbol{b} \\ \text{s.t.} \quad & \boldsymbol{f}_0 - \sum_{i=1}^m Q_i^T \boldsymbol{u}_i - \sum_{i=1}^m v_i \boldsymbol{c}_i + A^T \boldsymbol{w} = 0 \\ & \|\boldsymbol{u}_i\| \le v_i, i = 1, \ldots, m. \end{aligned} \tag{2.33}$$

The Slater's conditions are

$$\|Q_i \boldsymbol{x} + \boldsymbol{f}_i\| < \boldsymbol{c}_i^T \boldsymbol{x} + d_i, i = 1, \ldots, m$$

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

Semidefinite program

A *semidefinite program* (SDP) is an optimization problem in the space $\mathbb{S}^n$ of the form

$$\begin{aligned} \min_X \quad & C \cdot X \\ \text{s.t.} \quad & A_i \cdot X = b_i, i = 1, \ldots, m \\ & X \in \mathbb{S}_+^n, \end{aligned} \tag{2.34}$$

where X is the variable in $\mathbb{S}^n$, and $C, A_i \in \mathbb{S}^n, i = 1, \ldots, m$. As mentioned previously, $C \cdot X = \mathrm{tr}(CX)$ denotes the inner product. There is an analogy between LPs and SDPs: in the SDP, the objective function and equality constraints are also linear, in the space $\mathbb{S}^n$, and the positive semidefiniteness of X is corresponding to the nonnegativity of $\boldsymbol{x}$.

Let $\mathcal{X}_0 = \mathbb{S}_+^n$. The Lagrangian function for the SDP is then

$$\begin{aligned} L(X, \boldsymbol{u}) &= C \cdot X + \sum_{i=1}^m u_i (A_i \cdot X - b_i) \\ &= (C + \sum_{i=1}^m u_i A_i) \cdot X - \boldsymbol{u}^T \boldsymbol{b}, \end{aligned}$$

in which $\boldsymbol{u} = \{u_i\}_{i=1}^m$ and $\boldsymbol{b} = \{b_i\}_{i=1}^m$. Since the cone $\mathbb{S}_+^n$ is self-dual, if $C + \sum_{i=1}^m u_i A_i \notin \mathbb{S}_+^n$, the dual function will be equal to $-\infty$. Thus, the dual problem for (2.34) is

$$\begin{aligned} \max_{\boldsymbol{u}} \quad & -\boldsymbol{u}^T \boldsymbol{b} \\ \text{s.t.} \quad & C + \sum_{i=1}^m u_i A_i \succeq 0. \end{aligned} \tag{2.35}$$

For the SDP, the Slater's condition is that there exists a feasible solution $X \in \mathbb{S}^n$ such that $X \succ 0$.

Copositive program

If $\mathbb{S}^n_+$ expands into $\mathcal{COP}^n$, the SDP problem becomes a *copositive program* (COP),

$$\begin{aligned} \min_X \quad & C \cdot X \\ \text{s.t.} \quad & A_i \cdot X = b_i, i = 1, \ldots, m \\ & X \in \mathcal{COP}^n. \end{aligned} \tag{2.36}$$

Since $\mathbb{S}^n_+ \subseteq \mathcal{COP}^n$, every SDP is a COP. Comparing to SDPs, the Lagrangian dual problem can be similarly constructed for the COP. As the dual cone of $\mathcal{COP}^n$ is $\mathcal{CP}^n$, the dual function is equal to $-\boldsymbol{u}^T\boldsymbol{b}$ when $C + \sum_{i=1}^m u_i A_i \in \mathcal{CP}^n$; otherwise, the dual function is unbounded below. So the dual problem is

$$\begin{aligned} \max_{\boldsymbol{u}} \quad & -\boldsymbol{u}^T\boldsymbol{b} \\ \text{s.t.} \quad & C + \sum_{i=1}^m u_i A_i \in \mathcal{CP}^n. \end{aligned} \tag{2.37}$$

The Slater's condition for the COP is that there exists a feasible solution X of (2.36) such that $X \in \text{int}\mathcal{COP}^n$.

Geometric program

An optimization problem of the following form is called a *geometric program* (GP),

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f_0(\boldsymbol{x}) \\ \text{s.t.} \quad & f_i(\boldsymbol{x}) \leq 1, i = 1, \ldots, m \\ & h_i(\boldsymbol{x}) = 1, i = 1, \ldots, l, \end{aligned} \tag{2.38}$$

where f_i are *posynomials*,

$$f_i(\boldsymbol{x}) = \sum_{k=1}^{K_i} \alpha_{ik} x_1^{a_{ik}^{(1)}} x_2^{a_{ik}^{(2)}} \cdots x_n^{a_{ik}^{(n)}}, \; i = 0, 1, \ldots, m,$$

and h_i are *monomials*,

$$h_i(\boldsymbol{x}) = \beta_i x_1^{c_i^{(1)}} x_2^{c_i^{(2)}} \cdots x_n^{c_i^{(n)}}, \; i = 1, \ldots, l.$$

Here, $\alpha_{ik} > 0, i = 0, 1, \ldots, m, k = 1, \ldots, K_i$ and $\beta_i > 0, i = 1, \ldots, l$. The problem (2.38) is the *standard form* of a GP. The GP in the standard form is not a convex optimization problem, since posynomials are not convex functions. However, it can be transformed to a convex problem by a change of variables.

Let $y_i = \log x_i$, $b_{ik} = \log \alpha_{ik}$ and $d_i = \log \beta_i$. The problem (2.38) can be equivalently turned into the following problem:

$$\begin{aligned} \min_{\boldsymbol{y}} \quad & \sum_{k=1}^{K_0} \exp(\boldsymbol{a}_{0k}^T \boldsymbol{y} + b_{0k}) \\ \text{s.t.} \quad & \sum_{k=1}^{K_i} \exp(\boldsymbol{a}_{ik}^T \boldsymbol{y} + b_{ik}) \leq 1, i = 1, \ldots, m \\ & \exp(\boldsymbol{c}_i^T \boldsymbol{y} + d_i) = 1, i = 1, \ldots, l, \end{aligned} \tag{2.39}$$

where $\boldsymbol{a}_{ik} = \{a_{ik}^{(j)}\}_{j=1}^n$ and $\boldsymbol{c}_i = \{c_i^{(j)}\}_{j=1}^n$, which is further equivalent to

$$\begin{aligned} \min_{\boldsymbol{y}} \quad & \tilde{f}_0(\boldsymbol{y}) = \log \sum_{k=1}^{K_0} \exp(\boldsymbol{a}_{0k}^T \boldsymbol{y} + b_{0k}) \\ \text{s.t.} \quad & \tilde{f}_i(\boldsymbol{y}) = \log \sum_{k=1}^{K_i} \exp(\boldsymbol{a}_{ik}^T \boldsymbol{y} + b_{ik}) \leq 0, i = 1, \ldots, m \\ & \tilde{h}_i(\boldsymbol{y}) = \boldsymbol{c}_i^T \boldsymbol{y} + d_i = 0, i = 1, \ldots, l. \end{aligned} \tag{2.40}$$

By the fact that the *log-sum-exp function*

$$g(\boldsymbol{x}) = \log \sum_{i=1}^{n} \exp(x_i)$$

is convex, the problem (2.40) is a convex problem. It is referred to as the *convex form* for the GP.

In order to obtain the dual problem, first let

$$t_{ik} = \boldsymbol{a}_{ik}^T \boldsymbol{y} + b_{ik},\ i = 0, 1, \ldots, m,\ k = 1, \ldots, K_i,$$

and replace $\boldsymbol{a}_{ik}^T \boldsymbol{y} + b_{ik}$ in the problem (2.40) with t_{ik}. The Lagrangian function is then formulated as

$$\begin{aligned} L(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = & \log \sum_{k=1}^{K_0} \exp(t_{0k}) + \sum_{i=1}^{m} u_i \log \sum_{k=1}^{K_i} \exp(t_{ik}) \\ & + \sum_{i=0}^{m} \sum_{k=1}^{K_i} v_{ik}(\boldsymbol{a}_{ik}^T \boldsymbol{y} + b_{ik} - t_{ik}) + \sum_{i=1}^{l} w_i(\boldsymbol{c}_i^T \boldsymbol{y} + d_i), \end{aligned}$$

in which $\boldsymbol{t} = \{t_{ik}\}$, $\boldsymbol{u} = \{u_i\}$, $\boldsymbol{v} = \{v_{ik}\}$ and $\boldsymbol{w} = \{w_i\}$. Given $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$, the Lagrangian function L is convex with respect to $\boldsymbol{t}$ and linear with respect to $\boldsymbol{y}$. Take derivatives of L with respect to $(\boldsymbol{y}, \boldsymbol{t})$, we can analytically define the dual function as

$$d(\boldsymbol{v}, \boldsymbol{w}) = -\sum_{k=1}^{K_0} v_{0k} \log v_{0k} - \sum_{i=1}^{m} \sum_{k=1}^{K_i} v_{ik} \log \frac{v_{ik}}{\sum_{k=1}^{K_i} v_{ik}} + \sum_{i=0}^{m} \sum_{k=1}^{K_i} v_{ik} b_{ik} + \sum_{i=1}^{l} w_i d_i.$$

Notice that the multipliers u_i do not appear in the dual function, which results from the relation $\sum_{k=1}^{K_i} v_{ik} = u_i$, induced by letting the derivatives of L be equal to zero. Therefore, the dual problem is

$$\begin{aligned} \max_{\boldsymbol{v},\boldsymbol{w}} \quad & d(\boldsymbol{v},\boldsymbol{w}) \\ \text{s.t.} \quad & \sum_{i=0}^{m}\sum_{k=1}^{K_i} v_{ik}\boldsymbol{a}_{ik} + \sum_{i=1}^{l} w_i = 0, \\ & \sum_{k=1}^{K_0} v_{0k} = 1, \\ & v_{ik} \geq 0,\ i = 0,\ldots,m,\ k = 1,\ldots,K_i. \end{aligned} \tag{2.41}$$

The Slater's condition for a GP requires that there exists a point $\boldsymbol{y}$ such that the inequalities constraints hold strictly,

$$\log \sum_{k=1}^{K_i} \exp(\boldsymbol{a}_{ik}^T \boldsymbol{y} + b_{ik}) < 0,\ i = 1,\ldots,m.$$

Chapter 3

Canonical Duality Theory

3.1 Problem statements

Consider the following global optimization problem:

$$(\mathcal{P}) \qquad \begin{array}{ll} \min\limits_{\boldsymbol{x}} & \Pi(\boldsymbol{x}) = g_0(\boldsymbol{x}) \\ \text{s.t.} & g_i(\boldsymbol{x}) \leq 0,\ i = 1, \ldots, m, \\ & \boldsymbol{x} \in \mathbb{R}^n. \end{array} \tag{3.1}$$

It is assumed that $g_i(\boldsymbol{x}), i = 0, 1, \ldots, m$ are functions (not necessary convex) that can be written as

$$g_i(\boldsymbol{x}) = V_i(\boldsymbol{\Lambda}_i(\boldsymbol{x})) + \Lambda_{i0}(\boldsymbol{x})$$

with $\boldsymbol{\Lambda}_i(\boldsymbol{x}) : \mathbb{R}^n \to \mathcal{E}_i \subseteq \mathbb{R}^p$ and $\Lambda_{i0}(\boldsymbol{x}) : \mathbb{R}^n \to \mathcal{E}_{i0} \subseteq \mathbb{R}$ being twice continuously differentiable, where $\boldsymbol{\Lambda}_i(\boldsymbol{x}) = \{\Lambda_{ik}(\boldsymbol{x})\}_{k=1}^p$, and $V_i(\boldsymbol{\xi}_i) : \mathbb{R}^p \to \mathbb{R}$ being strictly convex and differentiable. We denote as

$$\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^n \mid g_i(\boldsymbol{x}) \leq 0, i = 1, \ldots, m\}$$

the feasible region of the problem $(\mathcal{P})$. Of course, It is assumed that the problem $(\mathcal{P})$ has at least one optimal solution in $\mathcal{X}$.

A broad range of optimization problems can be formulated as the problem $(\mathcal{P})$, which is illustrated by the applications in the following chapters. In [83], a related problem is discussed. Based on the problem $(\mathcal{P})$, the three parts of the canonical duality theory are detailedly presented. The complementary-dual principle is further developed, and it is proved that there exists a one-to-one correspondence between KKT points of the primal problem and the canonical dual problem and each pair of corresponding KKT points share the same function value. Then the triality theory is presented, and it is proved that solving globally the primal problem becomes a convex optimization problems as long as there is a KKT point in the positive semidefinite region in the dual space. It is shown that the canonical duality covers the classical Lagrangian duality for convex optimization problems.

The rest of the chapter is organised as follows: the canonical duality theory for the problem $(\mathcal{P})$ is discussed in Section 3.2. Then, in Section 3.3, the case where operators are quadratic is discussed, for which the triality theory is proposed and proved. All the results are explained by examples in the end.

3.2 Canonical duality theory

3.2.1 Canonical dual problem

For each of $V_i(\boldsymbol{\xi}_i), i = 0, 1, \ldots, m$, a duality mapping is introduced

$$\boldsymbol{\varsigma}_i = \{\varsigma_{ik}\}_{k=1}^{p} = \nabla V_i(\boldsymbol{\xi}_i) : \ \mathcal{E}_i \to \mathcal{E}_i^* \subseteq \mathbb{R}^p. \tag{3.2}$$

If $V_i, i = 0, 1, \ldots, m$ are *canonical functions* [38], that is, the inverse mapping $(\nabla V_i)^{-1}(\boldsymbol{\varsigma}_i)$ is single-valued, we have

$$\boldsymbol{\xi}_i = \nabla V_i^*(\boldsymbol{\varsigma}_i)$$

where $V_i^*(\boldsymbol{\varsigma}_i)$ is the conjugate of $V_i(\boldsymbol{\xi}_i)$ and it is defined by

$$V_i^*(\boldsymbol{\varsigma}_i) = \sup_{\boldsymbol{\xi}_i \in \mathbb{R}^n} \{\boldsymbol{\varsigma}_i^T \boldsymbol{\xi}_i - V_i(\boldsymbol{\xi}_i)\}.$$

Since $V_i, i = 0, 1, \ldots, m$ are assumed to be strictly convex in $\mathbb{R}^p$, by Theorem 26.3 [106] the conjugate functions $V_i^*(\boldsymbol{\varsigma}_i)$ are differentiable. Then the following equivalence holds true:

$$\boldsymbol{\varsigma}_i = \nabla V_i(\boldsymbol{\xi}_i) \ \Leftrightarrow \ \boldsymbol{\xi}_i = \nabla V_i^*(\boldsymbol{\varsigma}_i) \ \Leftrightarrow \ \boldsymbol{\xi}_i^T \boldsymbol{\varsigma}_i = V_i(\boldsymbol{\xi}_i) + V_i^*(\boldsymbol{\varsigma}_i). \tag{3.3}$$

From the equation (3.3), we then have

$$V_i(\boldsymbol{\Lambda}_i(\boldsymbol{x})) = \boldsymbol{\varsigma}_i^T \boldsymbol{\Lambda}_i(\boldsymbol{x}) - V_i^*(\boldsymbol{\varsigma}_i) \tag{3.4}$$

for each $\boldsymbol{x}$ satisfying $\boldsymbol{\varsigma}_i = \nabla V_i(\boldsymbol{\Lambda}_i(\boldsymbol{x}))$.

Let $\boldsymbol{\xi} = (\boldsymbol{\xi}_0, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m)$, $\boldsymbol{\varsigma} = (\boldsymbol{\varsigma}_0, \boldsymbol{\varsigma}_1, \ldots, \boldsymbol{\varsigma}_m)$, $\mathcal{E}_a = \mathcal{E}_0 \times \mathcal{E}_1 \times \cdots \times \mathcal{E}_m$ and $\mathcal{E}_a^* = \mathcal{E}_0^* \times \mathcal{E}_1^* \times \cdots \times \mathcal{E}_m^*$. The pair $(\boldsymbol{\xi}, \boldsymbol{\varsigma})$ is called *canonical duality pair* on $\mathcal{E}_a \times \mathcal{E}_a^*$. As the Lagrangian is introduced in Lagrangian duality, here the *total complementary function* [43, 45, 51, 47, 56, 116] (or the *extended Lagrangian* [38, 54]) is defined from $\mathbb{R}^n \times \mathbb{R}_+^m \times \mathcal{E}_a^*$ to $\mathbb{R}$:

$$\Xi(\boldsymbol{x}, \boldsymbol{\sigma}, \boldsymbol{\varsigma}) = \Lambda_{00}(\boldsymbol{x}) + \boldsymbol{\varsigma}_0^T \boldsymbol{\Lambda}_0(\boldsymbol{x}) - V_0^*(\boldsymbol{\varsigma}_0) + \sum_{i=1}^{m} \sigma_i(\Lambda_{i0}(\boldsymbol{x}) + \boldsymbol{\varsigma}_i^T \boldsymbol{\Lambda}_i(\boldsymbol{x}) - V_i^*(\boldsymbol{\varsigma}_i)),$$

where $\boldsymbol{\sigma} = \{\sigma_i\}_{i=1}^m \in \mathbb{R}_+^m$ are the Lagrangian multipliers associated with inequality constraints $g_i(\boldsymbol{x}) \leq 0, i = 1, \ldots, m$. From the total complementary function, the

canonical dual function is then defined by

$$\begin{aligned}\Pi^d(\boldsymbol{\sigma},\boldsymbol{\varsigma}) &= \mathrm{ext}_{\boldsymbol{x}\in\mathbb{R}^n}\{\Xi(\boldsymbol{x},\boldsymbol{\sigma},\boldsymbol{\varsigma})\} \\ &= U(\boldsymbol{\sigma},\boldsymbol{\varsigma}) - V_0^*(\boldsymbol{\varsigma}_0) - \sum_{i=1}^m \sigma_i V_i^*(\boldsymbol{\varsigma}_i),\end{aligned} \tag{3.5}$$

where

$$U(\boldsymbol{\sigma},\boldsymbol{\varsigma}) = \mathrm{ext}_{\boldsymbol{x}\in\mathbb{R}^n}\{\Lambda_{00}(\boldsymbol{x}) + \boldsymbol{\varsigma}_0^T\boldsymbol{\Lambda}_0(\boldsymbol{x}) + \sum_{i=1}^m \sigma_i(\Lambda_{i0}(\boldsymbol{x}) + \boldsymbol{\varsigma}_i^T\boldsymbol{\Lambda}_i(\boldsymbol{x}))\}$$

and $\mathrm{ext}\{\cdot\}$ denotes computing extrema for the function in braces.

Since Ξ is differentiable with respect to $\boldsymbol{x}$, all extrema happen at critical points. Let $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{\sigma},\boldsymbol{\varsigma}) = \{F_i(\boldsymbol{x},\boldsymbol{\sigma},\boldsymbol{\varsigma})\}_{i=1}^n$ denote the partial derivative of the function in $U(\boldsymbol{\sigma},\boldsymbol{\varsigma})$ with respect to $\boldsymbol{x}$,

$$\boldsymbol{F}(\boldsymbol{x},\boldsymbol{\sigma},\boldsymbol{\varsigma}) = \nabla\Lambda_{00}(\boldsymbol{x}) + \nabla\boldsymbol{\Lambda}_0(\boldsymbol{x})\boldsymbol{\varsigma}_0 + \sum_{i=1}^m \sigma_i(\nabla\Lambda_{i0}(\boldsymbol{x}) + \nabla\boldsymbol{\Lambda}_i(\boldsymbol{x})\boldsymbol{\varsigma}_i), \tag{3.6}$$

which is also the partial derivative of $\Xi(\boldsymbol{x},\boldsymbol{\sigma},\boldsymbol{\varsigma})$ with respect to $\boldsymbol{x}$. As $\boldsymbol{\Lambda}_i(\boldsymbol{x}), i = 0,1,\ldots,m$ are assumed twice continuously differentiable, $F_i(\boldsymbol{x},\boldsymbol{\sigma},\boldsymbol{\varsigma}), i = 1,\ldots,n$ then have continuous partial derivatives with respect to $\boldsymbol{x}$. On the other hand, with respect to $(\boldsymbol{\sigma},\boldsymbol{\varsigma})$, $F_i(\boldsymbol{x},\boldsymbol{\sigma},\boldsymbol{\varsigma}), i = 1,\ldots,n$ also have continuous partial derivatives. Given a pair $(\boldsymbol{\sigma},\boldsymbol{\varsigma})$, any stationary point $\boldsymbol{x}$ in (3.5) is a solution of the following system of equations

$$\boldsymbol{F}(\boldsymbol{x},\boldsymbol{\sigma},\boldsymbol{\varsigma}) = 0. \tag{3.7}$$

By the implicit function theorem, in an appropriate neighborhood of each $(\boldsymbol{x},\boldsymbol{\sigma},\boldsymbol{\varsigma})$ that satisfies (3.7) and has a nonzero Jacobin

$$J_F = \det(\frac{\partial\boldsymbol{F}(\boldsymbol{x},\boldsymbol{\sigma},\boldsymbol{\varsigma})}{\partial\boldsymbol{x}}) \neq 0, \tag{3.8}$$

a unique set of continuous functions

$$\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{\sigma},\boldsymbol{\varsigma}) \tag{3.9}$$

is determined by the system (3.7). Then, in the neighborhood, the function $U(\boldsymbol{\sigma},\boldsymbol{\varsigma})$ and thus the dual function $\Pi^d(\boldsymbol{\sigma},\boldsymbol{\varsigma})$ are well-defined. Substituting $\boldsymbol{x}$ with $\boldsymbol{x}(\boldsymbol{\sigma},\boldsymbol{\varsigma})$ in $U(\boldsymbol{\sigma},\boldsymbol{\varsigma})$, the dual function $\Pi^d(\boldsymbol{\sigma},\boldsymbol{\varsigma})$ can then be rewritten as

$$\begin{aligned}\Pi^d(\boldsymbol{\sigma},\boldsymbol{\varsigma}) =& \Lambda_{00}(\boldsymbol{x}(\boldsymbol{\sigma},\boldsymbol{\varsigma})) + \boldsymbol{\varsigma}_0^T\boldsymbol{\Lambda}_0(\boldsymbol{x}(\boldsymbol{\sigma},\boldsymbol{\varsigma})) - V_0^*(\boldsymbol{\varsigma}_0) + \\ & \sum_{i=1}^m \sigma_i(\Lambda_{i0}(\boldsymbol{x}(\boldsymbol{\sigma},\boldsymbol{\varsigma})) + \boldsymbol{\varsigma}_i^T\boldsymbol{\Lambda}_i(\boldsymbol{x}(\boldsymbol{\sigma},\boldsymbol{\varsigma})) - V_i^*(\boldsymbol{\varsigma}_i)).\end{aligned} \tag{3.10}$$

Because the implicit function theorem also guarantees that the function $\boldsymbol{x}(\boldsymbol{\sigma}, \boldsymbol{\varsigma})$ is continuously differentiable, the dual function is differentiable. Taking a partial derivative with respect to σ_i, we have

$$\frac{\partial \Pi^d(\boldsymbol{\sigma}, \boldsymbol{\varsigma})}{\partial \sigma_i} = \frac{\partial \boldsymbol{x}(\boldsymbol{\sigma}, \boldsymbol{\varsigma})}{\partial \sigma_i} \boldsymbol{F}(\boldsymbol{x}, \boldsymbol{\sigma}, \boldsymbol{\varsigma}) + \Lambda_{i0}(\boldsymbol{x}) + \boldsymbol{\varsigma}_i^T \boldsymbol{\Lambda}_i(\boldsymbol{x}) - V_i^*(\boldsymbol{\varsigma}_i).$$

Since $\boldsymbol{F}(\boldsymbol{x}(\boldsymbol{\sigma}, \boldsymbol{\varsigma}), \boldsymbol{\sigma}, \boldsymbol{\varsigma}) = 0$, the first item in the partial derivative vanishes. Hence, we have

$$\frac{\partial \Pi^d(\boldsymbol{\sigma}, \boldsymbol{\varsigma})}{\partial \boldsymbol{\sigma}} = \left\{ \Lambda_{i0}(\boldsymbol{x}) + \boldsymbol{\varsigma}_i^T \boldsymbol{\Lambda}_i(\boldsymbol{x}) - V_i^*(\boldsymbol{\varsigma}_i) \right\}_{i=1}^m . \tag{3.11}$$

Similarly, the partial derivatives of $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\varsigma})$ with respect to $\boldsymbol{\varsigma}$ are

$$\frac{\partial \Pi^d(\boldsymbol{\sigma}, \boldsymbol{\varsigma})}{\partial \boldsymbol{\varsigma}_0} = \boldsymbol{\Lambda}_0(\boldsymbol{x}) - \nabla V_0^*(\boldsymbol{\varsigma}_0), \tag{3.12}$$

$$\frac{\partial \Pi^d(\boldsymbol{\sigma}, \boldsymbol{\varsigma})}{\partial \boldsymbol{\varsigma}_i} = \sigma_i \left(\boldsymbol{\Lambda}_i(\boldsymbol{x}) - \nabla V_i^*(\boldsymbol{\varsigma}_i) \right) . \; i = 1, \ldots, m \tag{3.13}$$

Let $\mathcal{S}_a \subseteq \mathbb{R}^m \times \mathcal{E}_a^*$ denote a dual feasible region on which the canonical dual function $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\varsigma})$ is well-defined, that is, for each $(\boldsymbol{\sigma}, \boldsymbol{\varsigma}) \in \mathcal{S}_a$ there exists a vector $\boldsymbol{x}$ such that (3.7) and (3.8) hold. The *canonical dual problem* is defined as

$$(\mathcal{P}^d) \qquad \text{ext}\{\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\varsigma}) \mid \boldsymbol{\sigma} \geq 0, (\boldsymbol{\sigma}, \boldsymbol{\varsigma}) \in \mathcal{S}_a\}. \tag{3.14}$$

The complementary-dual principle discussed below states that the canonical dual problem $(\mathcal{P}^d)$ is perfectly dual to the primal problem $(\mathcal{P})$, that is, there is no duality gap between $(\mathcal{P})$ and $(\mathcal{P}^d)$. If the global optimality condition holds true, solving the problem $(\mathcal{P})$, either convex or nonconvex, becomes solving a convex subproblem in the dual space.

3.2.2 Complementary-dual principle

The duality relation between $(\mathcal{P})$ and $(\mathcal{P}^d)$ is illustrated by the following result.

Theorem 1 *Assume $\bar{\boldsymbol{x}}$ is a KKT point of the primal problem $(\mathcal{P})$ with a Lagrangian multiplier $\bar{\boldsymbol{\sigma}}$. Let $I = \{i \mid \bar{\sigma}_i > 0\}$ and $\bar{\boldsymbol{\varsigma}}_i, i = 0, 1, \ldots, m$ be any vectors that satisfy $\bar{\boldsymbol{\varsigma}}_i = \nabla V_i(\boldsymbol{\Lambda}_i(\bar{\boldsymbol{x}}))$ for $i \in \{0\} \cup I$. If $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}}) \in \mathcal{S}_a$, then $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}})$ is a KKT point of the dual problem $(\mathcal{P}^d)$.*

Let $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}})$ be a KKT point of the dual problem $(\mathcal{P}^d)$ and $\bar{\boldsymbol{x}}$ be a vector defined by (3.9), i.e., $\bar{\boldsymbol{x}} = \boldsymbol{x}(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}})$. If $\bar{\boldsymbol{x}}$ satisfies $g_i(\bar{\boldsymbol{x}}) \leq 0$ for $i = 1, \ldots, m$, then $\bar{\boldsymbol{x}}$ is a KKT point of the primal problem $(\mathcal{P})$ and $\bar{\boldsymbol{\sigma}}$ is the corresponding Lagrangian multiplier.

Moreover, for both statements, we have

$$\Pi(\bar{\boldsymbol{x}}) = \Xi(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}}) = \Pi^d(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}}). \tag{3.15}$$

Proof: From the assumption that $\bar{\boldsymbol{x}}$ is a KKT point of the primal problem $(\mathcal{P})$ and $\bar{\boldsymbol{\sigma}}$ is the corresponding the Lagrangian multiplier, the following KKT conditions hold

$$\nabla g_0(\bar{\boldsymbol{x}}) + \sum_{i=1}^{m} \bar{\sigma}_i \nabla g_i(\bar{\boldsymbol{x}}) = 0 \tag{3.16}$$

$$g_i(\bar{\boldsymbol{x}}) \leq 0, i = 1, \ldots, m \tag{3.17}$$

$$\bar{\boldsymbol{\sigma}} \geq 0 \tag{3.18}$$

$$\bar{\sigma}_i g_i(\bar{\boldsymbol{x}}) = 0, i = 1, \ldots, m \tag{3.19}$$

The condition (3.16) can also be written as

$$\nabla \Lambda_{00}(\bar{\boldsymbol{x}}) + \nabla \boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}}) \nabla V_0(\bar{\boldsymbol{\xi}}_0) + \sum_{i=1}^{m} \bar{\sigma}_i \left(\nabla \Lambda_{i0}(\bar{\boldsymbol{x}}) + \nabla \boldsymbol{\Lambda}_i(\bar{\boldsymbol{x}}) \nabla V_i(\bar{\boldsymbol{\xi}}_i) \right) = 0, \tag{3.20}$$

where $\bar{\boldsymbol{\xi}}_i = \boldsymbol{\Lambda}_i(\bar{\boldsymbol{x}}), i = 0, 1, \ldots, m$. By the definition, $\bar{\boldsymbol{\varsigma}}_i = \nabla V_i(\bar{\boldsymbol{\xi}}_i)$ for $i \in \{0\} \cup I$, the equation (3.20) means that we have

$$\boldsymbol{F}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}}) = 0.$$

Hence, the partial derivatives of Π^d at $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}})$ will follow the expressions in (3.11-3.13). By the equivalence in (3.3) and the condition (3.17), we have $0 \geq g_i(\bar{\boldsymbol{x}}) = \Lambda_{i0}(\bar{\boldsymbol{x}}) + \bar{\boldsymbol{\varsigma}}_i^T \boldsymbol{\Lambda}_i(\bar{\boldsymbol{x}}) - V_i^*(\bar{\boldsymbol{\varsigma}}_i), i \in I$. On the other hand, the Fenchel-Young inequality shows that we always have $\Lambda_{i0}(\boldsymbol{x}) + \bar{\boldsymbol{\varsigma}}_i^T \boldsymbol{\Lambda}_i(\bar{\boldsymbol{x}}) - V_i^*(\bar{\boldsymbol{\varsigma}}_i) \leq g_i(\bar{\boldsymbol{x}}) \leq 0$ for $i \in \{1, 2, \ldots, m\} \backslash I$. Thus, there exits a vector $\bar{\boldsymbol{\eta}} \geq 0$ such that $\nabla_{\boldsymbol{\sigma}} \Pi^d(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}}) + \bar{\boldsymbol{\eta}} = 0$, where, by the condition (3.19), we have $\bar{\eta}_i = 0$ for $i \in I$. Therefore, $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}})$ and $\bar{\boldsymbol{\eta}}$ satisfy the following conditions

$$\nabla_{\boldsymbol{\sigma}} \Pi^d(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}}) + \bar{\boldsymbol{\eta}} = 0 \tag{3.21}$$

$$\nabla_{\boldsymbol{\varsigma}_0} \Pi^d(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}}) = 0 \tag{3.22}$$

$$\nabla_{\boldsymbol{\varsigma}_i} \Pi^d(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}}) = 0 \tag{3.23}$$

$$\bar{\boldsymbol{\sigma}} \geq 0 \tag{3.24}$$

$$\bar{\boldsymbol{\eta}} \geq 0 \tag{3.25}$$

$$\bar{\sigma}_i \bar{\eta}_i = 0, i = 1, \ldots, m \tag{3.26}$$

It is proved that $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}})$ is a KKT point of the dual problem with the multiplier $\bar{\boldsymbol{\eta}}$.

Conversely, let $\bar{\boldsymbol{\eta}}$ be the multiplier that, together with $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}})$, satisfies the conditions (3.21–3.26). Let $I = \{i : \bar{\sigma}_i > 0\}$. The conditions (3.22) and (3.23) implies that we have $\bar{\boldsymbol{\varsigma}}_i = \nabla V_i(\boldsymbol{\Lambda}_i(\bar{\boldsymbol{x}}))$ and $g_0(\bar{\boldsymbol{x}}) = \Lambda_{i0}(\bar{\boldsymbol{x}}) + V_i(\boldsymbol{\Lambda}_i(\bar{\boldsymbol{x}})) = \Lambda_{i0}(\bar{\boldsymbol{x}}) + \bar{\boldsymbol{\varsigma}}_i^T \boldsymbol{\Lambda}_i(\bar{\boldsymbol{x}}) - V_i^*(\bar{\boldsymbol{\varsigma}}_i)$ for $i \in \{0\} \cup I$. Then we have conditions (3.17) and (3.19) hold because of (3.21), (3.25) and (3.26). The condition (3.16) is proved by the fact that

$$\nabla g_0(\bar{\boldsymbol{x}}) + \sum_{i=1}^{m} \bar{\sigma}_i \nabla g_i(\bar{\boldsymbol{x}}) = \boldsymbol{F}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}}) = 0.$$

So $\bar{\boldsymbol{x}}$ is proved to be a KKT point of the primal problem with the multiplier $\bar{\boldsymbol{\sigma}}$.

Since $\bar{\boldsymbol{x}}$ and $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}})$ satisfy the relation in (3.9), it is obvious that

$$\Pi^d(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}}) = \Xi(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}}).$$

Because of the condition (3.19) and $\Pi(\bar{\boldsymbol{x}}) = g_0(\bar{\boldsymbol{x}}) = \Lambda_{00}(\bar{\boldsymbol{x}}) + \boldsymbol{\varsigma}_0^T \boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}}) - V_0^*(\boldsymbol{\varsigma}_0)$, we have

$$\Xi(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}}) = \Pi(\bar{\boldsymbol{x}}).$$

Thus, the equation (3.15) is proved. □

Given a KKT point of the dual problem, if $\bar{\boldsymbol{\sigma}} > 0$, then $\bar{\boldsymbol{x}}$ must be a feasible solution of the primal problem. Actually, we have $g_i(\bar{\boldsymbol{x}}) = 0, i = 1, \ldots, m$. It is because, by the condition (3.26), $\bar{\boldsymbol{\sigma}} > 0$ implies that $\bar{\boldsymbol{\eta}} = 0$, and, by the condition (3.23), we have $\boldsymbol{\Lambda}_i(\boldsymbol{x}) = \nabla V_i^*(\boldsymbol{\varsigma}_i)$ for $i = 1, \ldots, m$. The conditions (3.21–3.23) with $\bar{\boldsymbol{\eta}} = 0$ also means that $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}})$ is a critical point of the dual function. Thus, we have the following result.

Corollary 2 *If $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}}) \in \mathcal{S}_a$ is a critical point of $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\varsigma})$ with $\bar{\boldsymbol{\sigma}} > 0$, then the vector $\bar{\boldsymbol{x}} = \boldsymbol{x}(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}})$ defined by (3.9) is a KKT point of the primal problem $(\mathcal{P})$.*

3.2.3 Global optimality condition

Let $G(\boldsymbol{x}, \boldsymbol{\sigma}, \boldsymbol{\varsigma})$ denote the Jacobian matrix of $\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{\sigma}, \boldsymbol{\varsigma})$, that is,

$$G(\boldsymbol{x}, \boldsymbol{\sigma}, \boldsymbol{\varsigma}) = \frac{\partial \boldsymbol{F}(\boldsymbol{x}, \boldsymbol{\sigma}, \boldsymbol{\varsigma})}{\partial \boldsymbol{x}}.$$

The matrix G is also the second partial derivative of the total complementary function $\Xi(\boldsymbol{x}, \boldsymbol{\sigma}, \boldsymbol{\varsigma})$ with respect to $\boldsymbol{x}$. The function Ξ is certainly twice differentiable on $\boldsymbol{x}$, as $\boldsymbol{\Lambda}_i(\boldsymbol{x}), i = 0, 1, \ldots, m$ are assumed to be twice differentiable. Let

$$\mathcal{S}_c^+ = \{(\boldsymbol{\sigma}, \boldsymbol{\varsigma}) \in \mathcal{S}_a \mid \boldsymbol{\sigma} \geq 0, G(\boldsymbol{x}, \boldsymbol{\sigma}, \boldsymbol{\varsigma}) \succeq 0, \forall \boldsymbol{x} \in \mathbb{R}^n\}.$$

We have the following results about the global optimality.

Theorem 3 *Let $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}})$ be a KKT point of the dual problem $(\mathcal{P}^d)$ and $\bar{\boldsymbol{x}} = \boldsymbol{x}(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}})$ be defined by (3.9). If $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}}) \in \mathcal{S}_c^+$, then $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}})$ is a global maximizer of $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\varsigma})$ on $\mathcal{S}_c^+$; if, in addition, $g_i(\bar{\boldsymbol{x}}) \geq 0, i = 1, \ldots, m$, then $\bar{\boldsymbol{x}}$ is a global solution for the primal problem $(\mathcal{P})$, with*

$$\Pi(\bar{\boldsymbol{x}}) = \min_{\boldsymbol{x} \in \mathcal{X}} \Pi(\boldsymbol{x}) = \max_{(\boldsymbol{\sigma}, \boldsymbol{\varsigma}) \in \mathcal{S}_c^+} \Pi^d(\boldsymbol{\sigma}, \boldsymbol{\varsigma}) = \Pi^d(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}}). \tag{3.27}$$

Proof: If $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}}) \in \mathcal{S}_c^+$, the definition of $\mathcal{S}_c^+$ implies that the funtcion $\Xi(\boldsymbol{x}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}})$ is convex with respect to $\boldsymbol{x}$. Thus, we have

$$\begin{aligned}
\Xi(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}}) &= \min_{\boldsymbol{x}\in\mathbb{R}^n} \Xi(\boldsymbol{x}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}}) \\
&\leq \min_{\boldsymbol{x}\in\mathbb{R}^n} \Pi(\boldsymbol{x}) + \sum_{i=1}^{m} \bar{\sigma}_i g_i(\boldsymbol{x}) \\
&\leq \min_{\boldsymbol{x}\in\mathbb{R}^n, g_i(\boldsymbol{x})\leq 0} \Pi(\boldsymbol{x}) + \sum_{i=1}^{m} \bar{\sigma}_i g_i(\boldsymbol{x}) \\
&\leq \min_{\boldsymbol{x}\in\mathbb{R}^n, g_i(\boldsymbol{x})\leq 0} \Pi(\boldsymbol{x}) \qquad (3.28)
\end{aligned}$$

where the first inequality results from the facts that $\bar{\boldsymbol{\sigma}} \geq 0$, $\Pi(\boldsymbol{x}) \geq \bar{\boldsymbol{\varsigma}}_0^T \boldsymbol{\Lambda}_0(\boldsymbol{x}) - V_0^*(\bar{\boldsymbol{\varsigma}}_0)$ and $g_i(\boldsymbol{x}) \geq \bar{\boldsymbol{\varsigma}}_i^T \boldsymbol{\Lambda}_i(\boldsymbol{x}) - V_i^*(\bar{\boldsymbol{\varsigma}}_i)$. While, because $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}})$ is a KKT point and $\bar{\boldsymbol{x}}$ is defined by (3.9), as shown in Theorem 1, if $g_i(\bar{\boldsymbol{x}}) \geq 0, i = 1, \ldots, m$, the vector $\bar{\boldsymbol{x}}$ is a KKT point of the primal problem and the equation (3.15) holds. Therefore, $\bar{\boldsymbol{x}}$ must be a global solution of the primal problem and the first equality in (3.27) is proved.

On the other hand, given $\boldsymbol{x}$, the function $\Xi(\boldsymbol{x}, \boldsymbol{\sigma}, \boldsymbol{\varsigma})$ is concave with $\boldsymbol{\varsigma}$, because of the convexity of $V_i^*(\boldsymbol{\varsigma}_i), i = 0, 1, \ldots, m$. Thus,

$$\begin{aligned}
\Xi(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}}) &= \max_{(\boldsymbol{\sigma},\boldsymbol{\varsigma})\in\mathcal{S}_c^+} \Xi(\bar{\boldsymbol{x}}, \boldsymbol{\sigma}, \boldsymbol{\varsigma}) \\
&\geq \max_{(\boldsymbol{\sigma},\boldsymbol{\varsigma})\in\mathcal{S}_c^+} \min_{\boldsymbol{x}\in\mathbb{R}^n} \Xi(\boldsymbol{x}, \boldsymbol{\sigma}, \boldsymbol{\varsigma}) \\
&= \max_{(\boldsymbol{\sigma},\boldsymbol{\varsigma})\in\mathcal{S}_c^+} \Pi^d(\boldsymbol{\sigma}, \boldsymbol{\varsigma}) \qquad (3.29)
\end{aligned}$$

where the inequality results from the convexity of $\Xi(\boldsymbol{x}, \boldsymbol{\sigma}, \boldsymbol{\varsigma})$ with respect to $\boldsymbol{x}$ when $(\boldsymbol{\sigma}, \boldsymbol{\varsigma}) \in \mathcal{S}_c^+$. Combining with the equation (3.15), we have proved that $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}})$ must be a global solution of the dual function over $\mathcal{S}_c^+$ and thus the last equality in (3.27) is true. □

By the Corollary 2, if $\bar{\boldsymbol{\sigma}} > 0$, the feasibility of $\bar{\boldsymbol{x}}$ will become redundant.

Corollary 4 *Let $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}})$ be a critical point of $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\varsigma})$. If $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}}) \in \mathcal{S}_c^+$ and $\bar{\boldsymbol{\sigma}} > 0$, then the vector $\bar{\boldsymbol{x}} = \boldsymbol{x}(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}})$ defined by (3.9) is a global solution of the primal problem $(\mathcal{P})$.*

Combining with the complementary-dual principle, the theorem also indicates that $\Pi^d(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}})$ has the smallest value among all the KKT points of the dual problem with the corresponding $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{\sigma}, \boldsymbol{\varsigma})$ being feasible.

The dual function Π^d may not have a maximizer in $\mathcal{S}_c^+$, because the function Π^d may not be defined on the boundary of $\mathcal{S}_c^+$. However, the weak duality always holds true.

Corollary 5 *It is always true that*

$$\sup_{(\boldsymbol{\sigma},\boldsymbol{\varsigma})\in\mathcal{S}_c^+} \Pi^d(\boldsymbol{\sigma}, \boldsymbol{\varsigma}) \;\leq\; \min_{\boldsymbol{x}\in\mathcal{X}} \Pi(\boldsymbol{x}). \qquad (3.30)$$

3.3 Quadratic operators and triality theory

In this section, we discuss the case where the geometrical operator $\boldsymbol{\Lambda}(\boldsymbol{x})$ is quadratic, which commonly arises in applications.

3.3.1 Canonical duality with quadratic operators

For $i = 0, 1, \ldots, m$, let

$$\Lambda_{i0}(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T A_0^i \boldsymbol{x} - \boldsymbol{x}^T \boldsymbol{b}_0^i, \text{ and } \boldsymbol{\Lambda}_i(\boldsymbol{x}) = \left\{ \frac{1}{2}\boldsymbol{x}^T A_k^i \boldsymbol{x} - \boldsymbol{x}^T \boldsymbol{b}_k^i \right\}_{k=1}^{p}, \tag{3.31}$$

in which $A_k^i \in \mathbb{S}^n$ and $\boldsymbol{b}_k^i \in \mathbb{R}^n$.

The total complementary function $\Xi(\boldsymbol{x}, \boldsymbol{\sigma}, \boldsymbol{\varsigma})$ in this case can be written as

$$\Xi(\boldsymbol{x}, \boldsymbol{\sigma}, \boldsymbol{\varsigma}) = \frac{1}{2}\boldsymbol{x}^T G(\boldsymbol{\sigma}, \boldsymbol{\varsigma})\boldsymbol{x} - \boldsymbol{x}^T \boldsymbol{f}(\boldsymbol{\sigma}, \boldsymbol{\varsigma}) - V_0^*(\boldsymbol{\varsigma}_0) - \sum_{i=1}^{m} \sigma_i V_i^*(\boldsymbol{\varsigma}_i) \tag{3.32}$$

with

$$\begin{aligned} G(\boldsymbol{\sigma}, \boldsymbol{\varsigma}) &= A_0^0 + \sum_{k=1}^{p} \varsigma_{0k} A_k^0 + \sum_{i=1}^{m} \sigma_i A_0^i + \sum_{i=1}^{m}\sum_{k=1}^{p} \sigma_i \varsigma_{ik} A_k^i, \text{ and} \\ \boldsymbol{f}(\boldsymbol{\sigma}, \boldsymbol{\varsigma}) &= \boldsymbol{b}_0^0 + \sum_{k=1}^{p} \varsigma_{0k} \boldsymbol{b}_k^0 + \sum_{i=1}^{m} \sigma_i \boldsymbol{b}_0^i + \sum_{i=1}^{m}\sum_{k=1}^{p} \sigma_i \varsigma_{ik} \boldsymbol{b}_k^i. \end{aligned}$$

Then, the equation (3.7) becomes

$$\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{\sigma}, \boldsymbol{\varsigma}) = G(\boldsymbol{\sigma}, \boldsymbol{\varsigma})\boldsymbol{x} - \boldsymbol{f}(\boldsymbol{\sigma}, \boldsymbol{\varsigma}) = 0. \tag{3.33}$$

So the Jacobin matrix is $G(\boldsymbol{\sigma}, \boldsymbol{\varsigma})$, and the condition (3.8) becomes

$$\det(G(\boldsymbol{\sigma}, \boldsymbol{\varsigma})) \neq 0.$$

The region $\mathcal{S}_a$ where the dual function is well-defined now can be written as

$$\mathcal{S}_a = \{(\boldsymbol{\sigma}, \boldsymbol{\varsigma}) \in \mathbb{R}^m \times \mathcal{E}_a^* \mid \det(G(\boldsymbol{\sigma}, \boldsymbol{\varsigma})) \neq 0\}, \tag{3.34}$$

and the dual function becomes

$$\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\varsigma}) = -\frac{1}{2}\boldsymbol{f}(\boldsymbol{\sigma}, \boldsymbol{\varsigma})^T G(\boldsymbol{\sigma}, \boldsymbol{\varsigma})^{-1} \boldsymbol{f}(\boldsymbol{\sigma}, \boldsymbol{\varsigma}) - V_0^*(\boldsymbol{\varsigma}_0) - \sum_{i=1}^{m} \sigma_i V_i^*(\boldsymbol{\varsigma}_i). \tag{3.35}$$

Notice that as the matrix $G(\boldsymbol{\sigma}, \boldsymbol{\varsigma})$ appears in the dual function $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\varsigma})$ in the form of the inverse. $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\varsigma})$ may converge to infinity as points approach to any point $(\boldsymbol{\sigma}, \boldsymbol{\varsigma})$ where $\det(G(\boldsymbol{\sigma}, \boldsymbol{\varsigma})) = 0$. So it may not possess Lipschitz continuity around points with zero determinant of the matrix G. The dual function $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\varsigma})$ is differentiable, and its partial derivatives (3.11), (3.12) and (3.13) can be obtained by applying the derivative of the inverse (see Appendix).

3.3.2 Triality theory

We then consider a case of the problem $(\mathcal{P})$ where there are no inequality constraints:

$$(\mathcal{P}_0) \qquad \min_{\boldsymbol{x}\in\mathbb{R}^n} \Pi(\boldsymbol{x}). \tag{3.36}$$

A more special problem of $(\mathcal{P}_0)$ has been addressed in [46, 56]. It is a problem of minimizing a *forth-order (or quartic) polynomial*,

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\sum_{k=1}^{m} \beta_k(\frac{1}{2}\boldsymbol{x}^T A^k \boldsymbol{x} - c^k)^2 + \frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x} - \boldsymbol{x}^T \boldsymbol{f}$$

in which $Q, A^k \in \mathbb{S}^n$, $\boldsymbol{f} \in \mathbb{R}^n$, $\beta_k \in \mathbb{R}_{++}$ and $c^k \in \mathbb{R}$. The objective function is a discretized form of the so-called double-well potential, first proposed by van der Waals in thermodynamics in 1895 [108], which is the mathematical model for natural phenomena of bifurcation and phase transitions in areas such as cosmology, continuum mechanics, material science, and quantum field theory [57, 71, 75].

As there are no inequality constraints in the problem $(\mathcal{P}_0)$, the variables $\boldsymbol{\sigma}$ and $\boldsymbol{\varsigma}_i$ for $i = 1, \ldots, m$ will disappear from the dual problem and hence $\boldsymbol{\varsigma} = \boldsymbol{\varsigma}_0 \in \mathbb{R}^p$. The equation (3.33) then becomes

$$G(\boldsymbol{\varsigma})\boldsymbol{x} - \boldsymbol{f}(\boldsymbol{\varsigma}) = 0, \tag{3.37}$$

with

$$G(\boldsymbol{\varsigma}) = A_0^0 + \sum_{k=1}^{m} \varsigma_k A_k^0 \text{ and } \boldsymbol{f}(\boldsymbol{\varsigma}) = \boldsymbol{b}_0^0 + \sum_{k=1}^{m} \varsigma_k \boldsymbol{b}_k^0.$$

The dual function then becomes

$$\Pi^d(\boldsymbol{\varsigma}) = -\frac{1}{2}\boldsymbol{f}(\boldsymbol{\varsigma})^T G(\boldsymbol{\varsigma})^{-1}\boldsymbol{f}(\boldsymbol{\varsigma}) - V_0^*(\boldsymbol{\varsigma}) \tag{3.38}$$

with the feasible region

$$\mathcal{S}_a = \{\boldsymbol{\varsigma} \in \mathcal{E}_a^* \mid \det(G(\boldsymbol{\varsigma})) \neq 0\}.$$

In (3.38), if $G(\boldsymbol{\varsigma}) \succ 0$, the first item of the dual function Π^d will be concave in the neighborhood of $\boldsymbol{\varsigma}$. Combining with the convexity of $V_0^*(\boldsymbol{\varsigma})$, the dual function Π^d will be concave in the neighborhood. Thus, we have the following result about the convexity of Π^d.

Lemma 6 *If*

$$\mathcal{S}_c^+ = \{\boldsymbol{\varsigma} \in \mathcal{S}_a \mid G(\boldsymbol{\varsigma}) \succeq 0\}$$

is convex, the dual function Π^d is a concave function on $\mathcal{S}_c^+$.

Before we present the following result, another subregion in $\mathcal{S}_a$ is introduced

$$\mathcal{S}_c^- = \{\boldsymbol{\varsigma} \in \mathcal{S}_a \mid G(\boldsymbol{\varsigma}) \preceq 0\}.$$

Theorem 7 (Triality Theorem) *Suppose that $\bar{\boldsymbol{\varsigma}}$ is a critical point of the dual function $\Pi^d(\boldsymbol{\varsigma})$ in (3.38) and $\bar{\boldsymbol{x}} = G(\bar{\boldsymbol{\varsigma}})^{-1}\boldsymbol{f}(\bar{\boldsymbol{\varsigma}})$.*

1. *(min-max duality) If $\bar{\boldsymbol{\varsigma}} \in \mathcal{S}_c^+$, then $\bar{\boldsymbol{\varsigma}}$ is a global maximizer of $\Pi^d(\boldsymbol{\varsigma})$ over $\mathcal{S}_c^+$ and $\bar{\boldsymbol{x}}$ is a global minimizer of $\Pi(\boldsymbol{x})$; moreover, the following equalities hold*

$$\Pi(\bar{\boldsymbol{x}}) = \min_{\boldsymbol{x}\in\mathbb{R}^n} \Pi(\boldsymbol{x}) = \max_{\boldsymbol{\varsigma}\in\mathcal{S}_c^+} \Pi^d(\boldsymbol{\varsigma}) = \Pi^d(\bar{\boldsymbol{\varsigma}}). \tag{3.39}$$

2. *(double-max duality) If $\bar{\boldsymbol{\varsigma}} \in \mathcal{S}_c^-$, then $\bar{\boldsymbol{\varsigma}}$ is a local maximizer of $\Pi^d(\boldsymbol{\varsigma})$ if and only if $\bar{\boldsymbol{x}}$ is a local maximizer of $\Pi(\boldsymbol{x})$; there exists a neighborhood of $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\varsigma}})$, $\mathcal{X}_o \times \mathcal{S}_o \subset \mathbb{R}^n \times \mathcal{S}_c^-$, such that*

$$\Pi(\bar{\boldsymbol{x}}) = \max_{\boldsymbol{x}\in\mathcal{X}_o} \Pi(\boldsymbol{x}) = \max_{\boldsymbol{\varsigma}\in\mathcal{S}_o} \Pi^d(\boldsymbol{\varsigma}) = \Pi^d(\bar{\boldsymbol{\varsigma}}). \tag{3.40}$$

3. *(double-min duality) If $\bar{\boldsymbol{\varsigma}} \in \mathcal{S}_c^-$, we have the following cases:*

   (a) *If $p = n$, then $\bar{\boldsymbol{\varsigma}}$ is a local minimizer of $\Pi^d(\boldsymbol{\varsigma})$ if and only if $\bar{\boldsymbol{x}}$ is a local minimizer of $\Pi(\boldsymbol{x})$; there exists a neighborhood of $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\varsigma}})$, $\mathcal{X}_o \times \mathcal{S}_o \subset \mathbb{R}^n \times \mathcal{S}_c^-$, such that*

$$\Pi(\bar{\boldsymbol{x}}) = \min_{\boldsymbol{x}\in\mathcal{X}_o} \Pi(\boldsymbol{x}) = \min_{\boldsymbol{\varsigma}\in\mathcal{S}_o} \Pi^d(\boldsymbol{\varsigma}) = \Pi^d(\bar{\boldsymbol{\varsigma}}). \tag{3.41}$$

   (b) *If $p < n$, then $\bar{\boldsymbol{\varsigma}}$ is a local minimizer or a saddle point of $\Pi^d(\boldsymbol{\varsigma})$ if and only if $\bar{\boldsymbol{x}}$ is a saddle point of $\Pi(\boldsymbol{x})$;*

   (c) *If $p > n$, then $\bar{\boldsymbol{\varsigma}}$ is a saddle point of $\Pi^d(\boldsymbol{\varsigma})$ if and only if $\bar{\boldsymbol{x}}$ is a local minimizer or a saddle point of $\Pi(\boldsymbol{x})$.*

Proof: If $\bar{\boldsymbol{\varsigma}} \in \mathcal{S}_c^+$, from the global optimality in Theorem 3, it is true that $\bar{\boldsymbol{\varsigma}}$ is a global maximizer of $\Pi^d(\boldsymbol{\varsigma})$ over $\mathcal{S}_c^+$, $\bar{\boldsymbol{x}} = G(\bar{\boldsymbol{\varsigma}})^{-1}\boldsymbol{f}(\bar{\boldsymbol{\varsigma}})$ is a global minimizer of $(\mathcal{P}_0)$ and equalities in (3.39) hold.

In the rest of this proof, we assume that $\bar{\boldsymbol{\varsigma}}$ is a critical point of $\Pi^d(\boldsymbol{\varsigma})$ in $\mathcal{S}_c^-$. Since $\bar{\boldsymbol{\varsigma}}$ is a critical point, we have

$$0 = \frac{\partial \Pi^d(\bar{\boldsymbol{\varsigma}})}{\partial \varsigma_k} = \frac{1}{2}\boldsymbol{f}(\bar{\boldsymbol{\varsigma}})^T G(\bar{\boldsymbol{\varsigma}})^{-1} A_k^0 G(\bar{\boldsymbol{\varsigma}})^{-1}\boldsymbol{f}(\bar{\boldsymbol{\varsigma}}) - \boldsymbol{b}_k^{0T} G(\bar{\boldsymbol{\varsigma}})^{-1}\boldsymbol{f}(\bar{\boldsymbol{\varsigma}}) - c_k^0 - \frac{\partial V_0^*(\bar{\boldsymbol{\varsigma}})}{\partial \varsigma_k},$$

for $k = 1, \ldots, p$. By substituting $G(\bar{\boldsymbol{\varsigma}})^{-1}\boldsymbol{f}(\bar{\boldsymbol{\varsigma}})$ with $\bar{\boldsymbol{x}}$, we then have

$$\nabla V_0^*(\bar{\boldsymbol{\varsigma}}) = \left\{\frac{1}{2}\bar{\boldsymbol{x}}^T A_k^0 \bar{\boldsymbol{x}} - \bar{\boldsymbol{x}}^T \boldsymbol{b}_k^0\right\}_{k=1}^p = \boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}}), \tag{3.42}$$

which, by (3.3), is equivalent to

$$\bar{\boldsymbol{\varsigma}} = \nabla V_0(\bar{\boldsymbol{\xi}}) \tag{3.43}$$

where $\bar{\boldsymbol{\xi}} = \boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}})$. Hessian matrices of $\Pi(\boldsymbol{x})$ and $\Pi^d(\boldsymbol{\varsigma})$ at $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{\varsigma}}$ can then be written as

$$\nabla^2\Pi(\bar{\boldsymbol{x}}) = G(\bar{\boldsymbol{\varsigma}}) + \nabla\boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}})\nabla^2 V_0(\bar{\boldsymbol{\xi}})\nabla\boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}})^T, \tag{3.44}$$

$$\nabla^2\Pi^d(\bar{\boldsymbol{\varsigma}}) = -\nabla^2 V_0^*(\bar{\boldsymbol{\varsigma}}) - \nabla\boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}})^T G(\bar{\boldsymbol{\varsigma}})^{-1}\nabla\boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}}). \tag{3.45}$$

Then Lemma 44 and the fact of $\nabla^2 V_0^*(\bar{\boldsymbol{\varsigma}}) = (\nabla^2 V_0(\bar{\boldsymbol{\xi}}))^{-1}$ show that, by letting $P = G(\bar{\boldsymbol{\varsigma}})$, $U = \nabla^2 V_0(\bar{\boldsymbol{\xi}})$ and $D = \nabla\boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}})$, we have the following equivalence

$$\nabla^2\Pi(\bar{\boldsymbol{x}}) \preceq 0 \Leftrightarrow \nabla^2\Pi^d(\bar{\boldsymbol{\varsigma}}) \preceq 0. \tag{3.46}$$

Thus, it is proved that $\bar{\boldsymbol{\varsigma}}$ being a local maximizer of $(\mathcal{P}_0^d)$ is equivalent to $\bar{\boldsymbol{x}}$ being a local maximizer of $(\mathcal{P}_0)$.

Then we prove the double-min duality. From the expressions in (3.44) and (3.45), we know that only when the second items in $\nabla^2\Pi(\bar{\boldsymbol{x}})$ and $\nabla^2\Pi^d(\bar{\boldsymbol{\varsigma}})$ are positive definite, the conditions $\nabla^2\Pi(\bar{\boldsymbol{x}}) \succeq 0$ and $\nabla^2\Pi^d(\bar{\boldsymbol{\varsigma}}) \succeq 0$ may hold true. Thus, when the condition $\nabla^2\Pi(\bar{\boldsymbol{x}}) \succeq 0$ is true, it can be proved that $\nabla^2 V_0(\bar{\boldsymbol{\xi}})$ is invertible and $\operatorname{rank}(\nabla\boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}})) = p$. When $p = n$, the matrix $\nabla\boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}})$ is also invertible. By letting $P = -\nabla\boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}})\nabla^2 V_0(\bar{\boldsymbol{\xi}})\nabla\boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}})^T$, $U = -G(\bar{\boldsymbol{\varsigma}})$ and $D = I$ in Lemma 44, the following equivalent relation can be derived

$$\begin{aligned} &\nabla\boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}})\nabla^2 V_0(\bar{\boldsymbol{\xi}})\nabla\boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}})^T + G(\bar{\boldsymbol{\varsigma}}) \succeq 0 \\ &\Longleftrightarrow -G(\bar{\boldsymbol{\varsigma}})^{-1} - \left(\nabla\boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}})\nabla^2 V_0(\bar{\boldsymbol{\xi}})\nabla\boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}})^T\right)^{-1} \succeq 0. \end{aligned} \tag{3.47}$$

The positive semidefiniteness on the right side is further equivalent to

$$-\nabla\boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}})^T G(\bar{\boldsymbol{\varsigma}})^{-1}\nabla\boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}}) - \nabla^2 V_0(\bar{\boldsymbol{\xi}})^{-1} \succeq 0,$$

of which the left side is $\nabla^2\Pi^d(\bar{\boldsymbol{\varsigma}})$, as $\nabla^2 V_0(\bar{\boldsymbol{\xi}})^{-1} = \nabla^2 V_0^*(\bar{\boldsymbol{\varsigma}})$. Thus, it is proved that when $p = n$, $\bar{\boldsymbol{\varsigma}}$ being a local minimizer of $(\mathcal{P}_p^d)$ is equivalent to $\bar{\boldsymbol{x}}$ being a local minimizer of $(\mathcal{P}_p)$

If the second items in $\nabla^2\Pi(\bar{\boldsymbol{x}})$ and $\nabla^2\Pi^d(\bar{\boldsymbol{\varsigma}})$ are positive definite, we have the following inequalities about the rank,

$$n = \operatorname{rank}(\nabla\boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}})\nabla^2 V_0(\bar{\boldsymbol{\xi}})\nabla\boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}})^T) \leq \operatorname{rank}(\nabla\boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}})) \leq \min\{n, p\} \tag{3.48}$$

$$p = \operatorname{rank}(\nabla\boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}})^T G(\bar{\boldsymbol{\varsigma}})^{-1}\nabla\boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}})) \leq \operatorname{rank}(\nabla\boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}})) \leq \min\{n, p\}. \tag{3.49}$$

Thus, when $p < n$, the Hessian $\nabla^2\Pi(\bar{\boldsymbol{x}})$ can not be positive semidefinite, otherwise, as mentioned above, it must be true that $\operatorname{rank}(\nabla\boldsymbol{\Lambda}_0(\bar{\boldsymbol{x}})) = p$, which will cause contradiction in (3.48). Hence, $\bar{\boldsymbol{x}}$ can not be a local minimizer when $p < n$. Combining with the double-max duality, we have proved that $\bar{\boldsymbol{\varsigma}}$ is a local minimizer or a saddle point of $\Pi^d(\boldsymbol{\varsigma})$ if and only if $\bar{\boldsymbol{x}}$ is a saddle point of $\Pi(\boldsymbol{x})$. Similarly, when $p > n$, we can prove that $\bar{\boldsymbol{\varsigma}}$ is a saddle point of $\Pi^d(\boldsymbol{\varsigma})$ if and only if $\bar{\boldsymbol{x}}$ is a local minimizer or a saddle point of $\Pi(\boldsymbol{x})$.

Combining with Theorem 1, equations (3.39), (3.41) and (3.41) are also proved. □

3.3.3 Hidden convexity

Three special cases are discussed, which, appearing as nonconvex problems, are actually equivalent to convex problems. The hidden convex nature of these problems can be easily verified by examining their canonical dual problems.

Case 1

In the first case, all the matrices in (3.31) are diagonal matrices,

$$A_k^i = \mathrm{diag}(\boldsymbol{a}_k^i), k = 0, 1, \ldots, p, i = 0, 1, \ldots, m,$$

and, except $\boldsymbol{b}_0^0$, all coefficients of linear items in the quadratic operators all vanish, that is,

$$\boldsymbol{b}_k^0 = 0, k = 1, \ldots, p, \text{ and } \boldsymbol{b}_k^i = 0, k = 0, 1, \ldots, p, i = 1, \ldots, m.$$

For simplicity, we assume that $\mathrm{dom}V_i^* = \mathbb{R}^n, i = 0, 1, \ldots, m$.

Then, G and $\boldsymbol{f}$ becomes

$$\begin{aligned} G(\boldsymbol{\sigma}, \boldsymbol{\varsigma}) &= \mathrm{diag}\left(\boldsymbol{a}_0^0 + \sum_{k=1}^{p} \varsigma_{0k}\boldsymbol{a}_k^0 + \sum_{i=1}^{m} \sigma_i \boldsymbol{a}_0^i + \sum_{i=1}^{m}\sum_{k=1}^{p} \sigma_i\varsigma_{ik}\boldsymbol{a}_k^i\right), \\ \boldsymbol{f}(\boldsymbol{\sigma}, \boldsymbol{\varsigma}) &= \boldsymbol{b}_0^0, \end{aligned}$$

and the canonical dual function in (3.35) can be written as

$$\begin{aligned} \Pi^d(\boldsymbol{\sigma}, \boldsymbol{\varsigma}) = &-\frac{1}{2}\sum_{j=1}^{n}(b_{0j}^0)^2/(a_{0j}^0 + \sum_{k=1}^{p}\varsigma_{0k}a_{kj}^0 + \sum_{i=1}^{m}\sigma_i a_{0j}^i + \sum_{i=1}^{m}\sum_{k=1}^{p}\sigma_i\varsigma_{ik}a_{kj}^i) \\ &- V_0^*(\boldsymbol{\varsigma}_0) - \sum_{i=1}^{m}\sigma_i V_i^*(\boldsymbol{\varsigma}_i), \end{aligned} \tag{3.50}$$

where b_{0j}^0 is the jth entry of $\boldsymbol{b}_0^0$ and a_{kj}^i is the jth entry of $\boldsymbol{a}_k^i$ for $k = 0, 1, \ldots, p, i = 0, 1, \ldots, m$.

It is obvious that, as long as $b_{0j}^0 \neq 0$, there is a 'wall', the hyperplane defined by letting the denominator be equal to zero, to which when the point $(\boldsymbol{\sigma}, \boldsymbol{\varsigma})$ approaches from one side of the wall, the function value converges to $-\infty$ or $+\infty$. These hyperplanes form the boundary of the positive semidefinite region $\mathcal{S}_c^+ = \{(\boldsymbol{\sigma}, \boldsymbol{\varsigma}) \mid G \succ 0\}$, which in this case is a convex polytope, consisting of all points that can make the denominators be positive. Immediately, we have the following result. The proof of the result is direct and hence omitted here.

Theorem 8 *If, for any j such that the hyperplane*

$$a_{0j}^0 + \sum_{k=1}^{p}\varsigma_{0k}a_{kj}^0 + \sum_{i=1}^{m}\sigma_i a_{0j}^i + \sum_{i=1}^{m}\sum_{k=1}^{p}\sigma_i\varsigma_{ik}a_{kj}^i = 0$$

contains a part of the boundary of $\mathcal{S}_c^+$, we have $b_{0j}^0 \neq 0$, then there must exist a critical point of $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\varsigma})$ in $\mathcal{S}_c^+$, and the corresponding vector $\boldsymbol{x} = G^{-1}\boldsymbol{f}$ is a global solution of the primal problem.

Case 2

The second case is of the form of (3.36), where the operator $\mathbf{\Lambda}_0(\boldsymbol{x})$ is a scalar function, hence denoted by $\Lambda_0(\boldsymbol{x})$. Accordingly, V_0 and the canonical dual function will become univariant functions. We assume that $A_1^0 \succ 0$ and $\mathrm{dom}V_0^* = \mathbb{R}$.

By applying the eigendecomposition, the matrices A_1^0 and A_0^0 can be simultaneously diagonalized. Thus, without loss of generality, we can just consider the following problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \Pi(\boldsymbol{x}) = V_0(\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x}) + \frac{1}{2}\boldsymbol{x}^T\mathrm{diag}(\boldsymbol{q})\boldsymbol{x} - \boldsymbol{p}^T\boldsymbol{x}, \tag{3.51}$$

where $\boldsymbol{q} = \{q_i\}_{i=1}^n$ with

$$q_1 = \ldots = q_k < q_{k+1} \leq \ldots \leq q_n.$$

Its canonical dual function is

$$\Pi^d(\varsigma) = -\frac{1}{2}\sum_{i=1}^{n} \frac{(p_i + \varsigma b_i)^2}{q_i + \varsigma} - V_0^*(\varsigma). \tag{3.52}$$

The region $\mathcal{S}_c^+$, under the assumption of $\mathrm{dom}V_0^* = \mathbb{R}$, becomes

$$\mathcal{S}_c^+ = \{\varsigma \,|\, \varsigma > -q_1\}.$$

Notice that, if $\sum_{i=1}^k (p_i - q_1 b_i)^2 \neq 0$, the line of $\varsigma = -q_1$ is a pole of the dual function $\Pi^d(\varsigma)$, i.e., as ς approaches to $-q_1$ from the right side the function value of $\Pi^d(\varsigma)$ converges to $-\infty$. When $\sum_{i=1}^k (p_i - q_1 b_i)^2 = 0$, the first derivative

$$\nabla\Pi^d(\varsigma) = \frac{1}{2}\sum_{i=1}^{n}\left(\frac{(p_i - q_i b_i)^2}{(q_i + \varsigma)^2} - b_i^2\right) - \nabla V_0^*(\varsigma)$$

will be well-defined at $\varsigma = -q_1$. If $\nabla\Pi^d(-q_1) \leq 0$, because of the strict concavity of Π^d, there will not exist any critical points in $\mathcal{S}_c^+$. While for both situations, $\sum_{i=1}^k (p_i - q_1 b_i)^2 \neq 0$ or $\nabla\Pi^d(-q_1) > 0$, if $\lim_{\varsigma\to+\infty} \nabla\Pi^d(\varsigma) < 0$, there must be a critical point in $\mathcal{S}_c^+$. Hence, we have the following result, which provides necessary and sufficient conditions for the existence of critical points in $\mathcal{S}_c^+$.

Theorem 9 *Assume*

$$\lim_{\varsigma\to+\infty} \nabla V_0^*(\varsigma) > -\frac{1}{2}\boldsymbol{b}^T\boldsymbol{b}.$$

The function $\Pi^d(\varsigma)$ in (3.52) has a critical point in $\mathcal{S}_c^+$ if and only if

$$\sum_{i=1}^{k}(p_i - q_1 b_i)^2 \neq 0 \tag{3.53}$$

or

$$\frac{1}{2}\sum_{i=k+1}^{n}\frac{(p_i - q_i b_i)^2}{(q_i - q_1)^2} - \frac{1}{2}\boldsymbol{b}^T\boldsymbol{b} - \nabla V_0^*(-q_1) > 0. \tag{3.54}$$

If we define

$$\bar{\Pi}^d(\varsigma) = \begin{cases} \Pi^d(\varsigma) & \varsigma \in \mathcal{S}_c^+ \\ -\frac{1}{2}\sum_{i=k+1}^n \frac{(p_i - q_1 b_i)^2}{q_i - q_1} - V_0^*(-q_1) & \varsigma = -q_1 \end{cases}$$

the new dual function $\bar{\Pi}^d(\varsigma)$ is well-defined on

$$\bar{\mathcal{S}}_a^+ = \{\varsigma \,|\, \varsigma \geq -q_1\}.$$

Moreover, it is differentiable, and the derivative $\nabla\bar{\Pi}^d(\varsigma)$ is equal to $\nabla\Pi^d(\varsigma)$. The following result shows that the problem (3.51) actually possesses hidden convexity.

Theorem 10 *Assume*

$$\lim_{\varsigma \to +\infty} \nabla V_0^*(\varsigma) > -\frac{1}{2}\boldsymbol{b}^T\boldsymbol{b}$$

and let $\bar{\varsigma}$ be a maximizer of the concave maximization problem

$$\max\{\bar{\Pi}^d(\varsigma) \,|\, \varsigma \in \bar{\mathcal{S}}_a^+\}. \tag{3.55}$$

If $\bar{\varsigma} > -q_1$, let

$$\bar{x}_i = \frac{p_i + \bar{\varsigma} b_i}{q_i + \bar{\varsigma}}, i = 1, \ldots, n;$$

if $\bar{\varsigma} = -q_1$, let $\bar{\boldsymbol{x}}$ be any vector satisfying

$$\frac{1}{2}\bar{\boldsymbol{x}}^T\bar{\boldsymbol{x}} - \boldsymbol{b}^T\bar{\boldsymbol{x}} = \nabla V_0^*(-q_1), \text{ and } \bar{x}_i = \frac{p_i - q_1 b_i}{q_i - q_1}, i = k+1, \ldots, n.$$

Then, $\bar{\boldsymbol{x}}$ is a global solution of the primal problem (3.51), and

$$\Pi(\bar{\boldsymbol{x}}) = \bar{\Pi}^d(\bar{\varsigma}).$$

Proof: If $\bar{\varsigma} > -q_1$, an interior point of $\mathcal{S}_c^+$, it is a critical point, and hence the point $\bar{\boldsymbol{x}} = G(\bar{\varsigma})^{-1}\boldsymbol{f}(\bar{\varsigma})$ is a global solution of the problem (3.51). Now we discuss the situation when $\bar{\varsigma} = -q_1$. By Theorem 9, this situation can only happen when

$$\sum_{i=1}^k (p_i - q_1 b_i)^2 = 0 \text{ and } \frac{1}{2}\sum_{i=k+1}^n \frac{(p_i - q_i b_i)^2}{(q_i - q_1)^2} - \frac{1}{2}\boldsymbol{b}^T\boldsymbol{b} - \nabla V_0^*(-q_1) \leq 0.$$

On the other hand, any solution $\bar{\boldsymbol{x}}$ with

$$\bar{x}_i = \frac{p_i - q_1 b_i}{q_i - q_1}, i = k+1, \ldots, n, \text{ and } \bar{x}_i \in \mathbb{R}, i = 1, \ldots, k$$

is a solution of the equation $G(\bar{\varsigma})\boldsymbol{x} = \boldsymbol{f}(\bar{\varsigma})$. Then we have

$$\frac{1}{2}\bar{\boldsymbol{x}}^T\bar{\boldsymbol{x}} - \boldsymbol{b}^T\bar{\boldsymbol{x}} = \sum_{i=1}^k (\frac{1}{2}\bar{x}_i^2 - b_i\bar{x}_i) + \frac{1}{2}\sum_{i=k+1}^n \frac{(p_i - q_i b_i)^2}{(q_i - q_1)^2} - \frac{1}{2}\boldsymbol{b}^T\boldsymbol{b},$$

and we are always able to choose $\bar{x}_i, i = 1, \ldots, k$ such that $\bar{\xi} = \frac{1}{2}\bar{\boldsymbol{x}}^T\bar{\boldsymbol{x}} - \boldsymbol{b}^T\bar{\boldsymbol{x}} = \nabla V_0^*(-q_1)$. By the equivalence in (3.3), it holds true that $V_0(\bar{\xi}) = \bar{\xi}\bar{\varsigma} - V_0^*(\bar{\varsigma})$, from which we have $\Pi(\bar{\boldsymbol{x}}) = \bar{\Pi}^d(\bar{\varsigma})$. Thus $\bar{\boldsymbol{x}}$ is a global solution of the primal problem (3.51). The theorem is proved. □

This result shows that, by solving the problem (3.62), which is a convex optimization problem, we can find a global solution for the primal problem (3.51). It also shows that if the maximizer is on the boundary of $\bar{\mathcal{S}}_a^+$, the primal problem has infinitely many global solutions.

Case 3

The third case is a constrained problem, with only one constraint:

$$\begin{aligned} &\min_{\boldsymbol{x}} \ \Lambda_{00}(\boldsymbol{x}) \\ &\text{s.t. } V_1(\Lambda_{11}(\boldsymbol{x})) \leq 0, \end{aligned} \tag{3.56}$$

where we assume that the matrix A_1^1 in $\Lambda_{11}(\boldsymbol{x})$ is positive definite and $\text{dom}V_1^* = \mathbb{R}$.

Similar to the situation in the case 2 above, A_0^0 and A_1^1 can be simultaneously diagonalized, by rotating and scaling the vector $\boldsymbol{x}$, and the problem (3.56) can be equivalently transformed into the following problem:

$$\begin{aligned} &\min_{\boldsymbol{x}} \ \frac{1}{2}\boldsymbol{x}^T\text{diag}(\boldsymbol{q})\boldsymbol{x} - \boldsymbol{p}^T\boldsymbol{x} \\ &\text{s.t. } V_1(\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x}) \leq 0, \end{aligned} \tag{3.57}$$

where the entries of $\boldsymbol{q} = \{q_i\}_{i=1}^n$ are in nondecreasing order, i.e.,

$$q_1 = \ldots = q_k < q_{k+1} \leq \ldots \leq q_n.$$

Its canonical dual function is then

$$\Pi^d(\sigma, \varsigma) = -\frac{1}{2}\sum_{i=1}^{n} \frac{(p_i + \sigma\varsigma b_i)^2}{q_i + \sigma\varsigma} - \sigma V_1^*(\varsigma). \tag{3.58}$$

The set $\mathcal{S}_c^+$ then is

$$\mathcal{S}_c^+ = \{(\sigma, \varsigma) \,|\, \sigma \geq 0, \sigma\varsigma > -q_1\}.$$

Unfortunately, the set $\mathcal{S}_c^+$ is not always convex: if $q_1 \leq 0$, $\mathcal{S}_c^+$ is convex; if $q_1 > 0$, $\mathcal{S}_c^+$ is not convex. If let $\tau = \sigma\varsigma$, we can transform the set $\mathcal{S}_c^+$ into a convex one and cancel the variable ς in $\Pi^d(\sigma, \varsigma)$ by replacing ς with τ/σ. However, the replacing is not legal when $\sigma = 0$.

Let

$$\hat{\Pi}^d(\sigma, \tau) = -\frac{1}{2}\sum_{i=1}^{n} \frac{(p_i + \tau b_i)^2}{q_i + \tau} - \sigma V_1^*(\frac{\tau}{\sigma}), \tag{3.59}$$

and

$$\hat{\mathcal{S}}_a^+ = \{(\sigma, \tau) \mid \sigma > 0, \tau > -q_1\}.$$

It can be verified that $\hat{\Pi}^d(\sigma, \tau)$ is a convex function over $\hat{\mathcal{S}}_a^+$. Inspired by Theorem 9, we have the following result.

Lemma 11 *Assume*

$$\lim_{\tau \to +\infty} \nabla V_1^*(\frac{\tau}{\sigma}) > -\frac{1}{2}\boldsymbol{b}^T\boldsymbol{b}$$

for any $\sigma > 0$. If $(\bar{\sigma}, \bar{\tau}) \in \hat{\mathcal{S}}_a^+$ is a critical point of $\hat{\Pi}^d(\sigma, \tau)$, then we have

$$\sum_{i=1}^{k}(p_i - q_1 b_i)^2 \neq 0 \tag{3.60}$$

or

$$\frac{1}{2}\sum_{i=k+1}^{n}\frac{(p_i - q_i b_i)^2}{(q_i - q_1)^2} - \frac{1}{2}\boldsymbol{b}^T\boldsymbol{b} - \nabla V_1^*(\frac{-q_1}{\bar{\sigma}}) > 0. \tag{3.61}$$

However, here only necessary conditions are identified, because it is difficult to predict the behavior of $\hat{\Pi}^d(\sigma, \tau)$ when σ approaches to the boundary of $\hat{\mathcal{S}}_a^+$. If, in addition, we know the value of $-\sigma V_1^*(\frac{\tau}{\sigma})$ in $\hat{\Pi}^d(\sigma, \tau)$ will converge to $-\infty$ when σ approaches to the boundary, the conditions (3.60) and (3.61) will become sufficient for the existence of a critical point.

Let

$$\bar{\mathcal{S}}_a^+ = \{(\sigma, \tau) \mid \sigma > 0, \tau \geq -q_1\}$$

and

$$\bar{\Pi}^d(\sigma, \tau) = \begin{cases} \hat{\Pi}^d(\sigma, \tau) & (\sigma, \tau) \in \hat{\mathcal{S}}_a^+ \\ -\frac{1}{2}\sum_{i=k+1}^{n}\frac{(p_i - q_1 b_i)^2}{q_i - q_1} - \sigma V_1^*(\frac{-q_1}{\sigma}) & \tau = -q_1, \sigma > 0 \end{cases}$$

Then we have the following result, which is similar to Theorem 10.

Theorem 12 *Assume*

$$\lim_{\tau \to +\infty} \nabla V_1^*(\frac{\tau}{\sigma}) > -\frac{1}{2}\boldsymbol{b}^T\boldsymbol{b}$$

for any $\sigma > 0$, and let $(\bar{\sigma}, \bar{\tau})$ be a maximizer of the concave maximization problem

$$\max\{\bar{\Pi}^d(\sigma, \tau) \mid (\sigma, \tau) \in \bar{\mathcal{S}}_a^+\}. \tag{3.62}$$

If $\bar{\tau} > -q_1$, let

$$\bar{x}_i = \frac{p_i + \bar{\tau} b_i}{q_i + \bar{\tau}}, i = 1, \ldots, n;$$

if $\bar{\tau} = -q_1$, let $\bar{\boldsymbol{x}}$ be any vector satisfying

$$\frac{1}{2}\bar{\boldsymbol{x}}^T\bar{\boldsymbol{x}} - \boldsymbol{b}^T\bar{\boldsymbol{x}} = \nabla V_1^*(\frac{-q_1}{\bar{\sigma}}), \text{ and } \bar{x}_i = \frac{p_i - q_1 b_i}{q_i - q_1}, i = k+1, \ldots, n.$$

Then, $\bar{\boldsymbol{x}}$ is a global solution of the primal problem (3.51), and

$$\Pi(\bar{\boldsymbol{x}}) = \bar{\Pi}^d(\bar{\sigma}, \bar{\tau}).$$

3.3.4 Convex optimization for global solutions

From the expression of Hessian matrix of $\Pi^d(\boldsymbol{\varsigma})$ in (3.45), it is obvious that if $G(\boldsymbol{\varsigma})$ is positive definite $\nabla^2\Pi^d(\boldsymbol{\varsigma})$ will be negative definite. Thus, the dual function $\Pi^d(\boldsymbol{\varsigma})$ is concave over $\mathcal{S}_c^+$, and calculating the maximizer over $\mathcal{S}_c^+$, i.e.

$$\max_{\boldsymbol{\varsigma}\in\mathcal{S}_c^+} \Pi^d(\boldsymbol{\varsigma}) \tag{3.63}$$

becomes a convex optimization problem. As soon as we solve this convex problem, the maximizer can be checked: if it is a critical point, the Triality Theorem (Theorem 7) guarantees that the corresponding primal solution is a global solution.

We then discuss the general quartic polynomial problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \Pi(\boldsymbol{x}) = \frac{1}{2}\sum_{i=1}^{m}(\frac{1}{2}\boldsymbol{x}^T A_i\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{b}_i)^2 + \frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{p}. \tag{3.64}$$

Its canonical dual function is

$$\Pi^d(\boldsymbol{\varsigma}) = -\frac{1}{2}\boldsymbol{f}(\boldsymbol{\varsigma})^T G(\boldsymbol{\varsigma})^{-1}\boldsymbol{f}(\boldsymbol{\varsigma}) - \frac{1}{2}\boldsymbol{\varsigma}^T\boldsymbol{\varsigma}, \tag{3.65}$$

where

$$G(\boldsymbol{\varsigma}) = Q + \sum_{i=1}^{m}\varsigma_i A_i \text{ and } \boldsymbol{f}(\boldsymbol{\varsigma}) = \boldsymbol{p} + \sum_{i=1}^{m}\varsigma_i\boldsymbol{b}_i.$$

By introducing an extra variable t, the maximization problem (3.63) then equivalently becomes

$$\begin{aligned} \max_{\boldsymbol{\varsigma}} \ & t \\ \text{s.t.} \ & \boldsymbol{\varsigma} \in \mathcal{S}_c^+ \\ & \Pi^d(\boldsymbol{\varsigma}) - t \geq 0 \end{aligned} \tag{3.66}$$

Let

$$X = \begin{bmatrix} 2G(\boldsymbol{\varsigma}) & 0 & \boldsymbol{f}(\boldsymbol{\varsigma}) \\ 0 & 2I & \boldsymbol{\varsigma} \\ \boldsymbol{f}(\boldsymbol{\varsigma})^T & \boldsymbol{\varsigma}^T & -t \end{bmatrix} \text{ and } B = \begin{bmatrix} 2G(\boldsymbol{\varsigma}) & 0 \\ 0 & 2I \end{bmatrix}.$$

Then $\Pi^d(\boldsymbol{\varsigma}) - t$ is the Schur complement of B in X. From the Schur complement condition for positive definiteness, the last constraint in (3.66) is equivalent to

$$X \succeq 0.$$

If relax the semidefinite region $\mathcal{S}_c^+$ into $\{\boldsymbol{\varsigma} \in \mathbb{R}^m \mid G(\boldsymbol{\varsigma}) \succeq 0\}$, we get an SDP problem

$$\begin{aligned} \max_{\boldsymbol{\varsigma}} \ & t \\ \text{s.t.} \ & X \succeq 0 \end{aligned} \tag{3.67}$$

Corollary 13 *If $\bar{\boldsymbol{\varsigma}}$ is an optimal solution of the SDP problem (3.67) with $G(\bar{\boldsymbol{\varsigma}}) \succ 0$, then it is a critical point of the canonical dual function $\Pi^d(\boldsymbol{\varsigma})$ and the corresponding solution $\bar{\boldsymbol{x}} = G(\bar{\boldsymbol{\varsigma}})^{-1}\boldsymbol{f}(\bar{\boldsymbol{\varsigma}})$ is an optimal solution of the quartic polynomial problem (3.64).*

3.3.5 Examples

In the end, we use examples to illustrate results discussed above. In order to use graphs to clearly show the duality relations, all examples are of dimensions not larger than 2.

A strictly convex quadratic function is a natural choice for the function $V_i(\boldsymbol{\xi}_i)$, and it arises in many applications. If all V_i are quadratic functions, as mentioned above, the function $\Pi(\boldsymbol{x})$ becomes a quartic polynomial. This case has been discussed in [46, 56], where examples can be found for understanding the triality theory.

Here, we discuss a more general case where V_i could be a quadratic function or a fourth-order polynomial,

$$V_i(\xi_i) = \alpha \xi_i^4,$$

which is strictly convex when $\alpha > 0$. Consider the following minimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \left\{ \Pi(\boldsymbol{x}) = W(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{x}^T Q \boldsymbol{x} - \boldsymbol{x}^T \boldsymbol{p} \right\},$$

where

$$W(\boldsymbol{x}) = \frac{1}{4}\left(\frac{1}{2}\boldsymbol{x}^T A_1 \boldsymbol{x} - \boldsymbol{x}^T \boldsymbol{b}_1 - c_1\right)^4 + \frac{1}{2}\left(\frac{1}{2}\boldsymbol{x}^T A_2 \boldsymbol{x} - \boldsymbol{x}^T \boldsymbol{b}_2 - c_2\right)^2.$$

Here, $A_1, A_2, Q \in \mathbb{S}^2$, $\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{p} \in \mathbb{R}^2$ and $c_1, c_2 \in \mathbb{R}$. Its canonical dual function, which is of dimension 2, is then formulated as

$$\Pi^d(\boldsymbol{\varsigma}) = -\frac{1}{2}\boldsymbol{f}(\boldsymbol{\varsigma}) G(\boldsymbol{\varsigma})^{-1} \boldsymbol{f}(\boldsymbol{\varsigma}) - c_1\varsigma_1 - c_2\varsigma_2 - \frac{3}{4}\varsigma_1^{4/3} - \frac{1}{2}\varsigma_2^2,$$

in which

$$G(\boldsymbol{\varsigma}) = Q + \varsigma_1 A_1 + \varsigma_2 A_2, \text{ and } \boldsymbol{f}(\boldsymbol{\varsigma}) = \boldsymbol{p} + \varsigma_1 \boldsymbol{b}_1 + \varsigma_2 \boldsymbol{b}_2.$$

Three instances are given below to show the duality relations described in the Triality Theorem for the three cases, $n = p$, $n > p$ and $n < p$.

Instance 1

In the function $\Pi(\boldsymbol{x})$, we let the matrices be

$$A_1 = \begin{pmatrix} 3 & 0 \\ 0 & 8 \end{pmatrix}, A_2 = \begin{pmatrix} 10 & 6 \\ 6 & -3 \end{pmatrix}, \text{ and } Q = \begin{pmatrix} 2 & 4 \\ 4 & 0 \end{pmatrix},$$

the vectors be

$$\boldsymbol{b}_1 = \begin{pmatrix} -6 \\ -7 \end{pmatrix}, \boldsymbol{b}_2 = \begin{pmatrix} -6 \\ -2 \end{pmatrix}, \text{ and } \boldsymbol{p} = \begin{pmatrix} -6 \\ -5 \end{pmatrix},$$

and the scalars be $c_1 = 6$ and $c_2 = -5$. Hence, for this instance, we have $n = p = 2$. The graphs and contours of $\Pi(\boldsymbol{x})$ and $\Pi^d(\boldsymbol{\varsigma})$ are shown in Figure 3.1.

$\boldsymbol{x}$	$\Pi(\boldsymbol{x})$	optimality	$\boldsymbol{\varsigma}$	$\Pi^d(\boldsymbol{\varsigma})$	$\mathcal{S}_a$	optimality
$\boldsymbol{x}_1$	-22.921	global minimizer	$\boldsymbol{\varsigma}_1$	-22.921	$\mathcal{S}_c^+$	local maximizer
$\boldsymbol{x}_2$	8.405	local minimizer	$\boldsymbol{\varsigma}_2$	8.405		saddle point
$\boldsymbol{x}_3$	16.99	local minimizer	$\boldsymbol{\varsigma}_3$	16.99	$\mathcal{S}_c^-$	local minimizer
$\boldsymbol{x}_4$	37.70	saddle point	$\boldsymbol{\varsigma}_4$	37.70	$\mathcal{S}_c^-$	saddle point
$\boldsymbol{x}_5$	66.17	saddle point	$\boldsymbol{\varsigma}_5$	66.17		local maximizer
$\boldsymbol{x}_6$	5037.1	saddle point	$\boldsymbol{\varsigma}_6$	5037.1	$\mathcal{S}_c^-$	saddle point
$\boldsymbol{x}_7$	13092.9	local maximizer	$\boldsymbol{\varsigma}_7$	13092.9	$\mathcal{S}_c^-$	local maximizer

Table 3.1: Dualities for Instance 1.

The primal function $\Pi(\boldsymbol{x})$ has 7 critical points,

$$\begin{aligned}
&\boldsymbol{x}_1 = (-2.124, 1.136), \quad && \boldsymbol{x}_2 = (0.927, -1.510), \quad && \boldsymbol{x}_3 = (-0.734, -2.588),\\
&\boldsymbol{x}_4 = (0.082, -2.169), && \boldsymbol{x}_5 = (0.732, -0.285), && \boldsymbol{x}_6 = (-4.031, -1.218),\\
&\boldsymbol{x}_7 = (-2.041, -0.881),
\end{aligned}$$

which are corresponding to the 7 critical points of the dual function $\Pi^d(\varsigma_1, \varsigma_2)$,

$$\begin{aligned}
&\boldsymbol{\varsigma}_1 = (1.440, 0.683), \quad && \boldsymbol{\varsigma}_2 = (-0.217, 0.016), \quad && \boldsymbol{\varsigma}_3 = (-0.783, -0.521),\\
&\boldsymbol{\varsigma}_4 = (-6.479, -6.935), && \boldsymbol{\varsigma}_5 = (-15.122, 10.130), && \boldsymbol{\varsigma}_6 = (-594.32, 86.835),\\
&\boldsymbol{\varsigma}_7 = (-3415.5, 21.457).
\end{aligned}$$

The Table 3.1 shows clearly the min-max, double-max and double-min dualities of Triality Theorem.

Instance 2

In this instance, we let the second item in $W(\boldsymbol{x})$ vanish, i.e., $A_2 = 0$, $\boldsymbol{b}_2 = 0$ and $c_2 = 0$, and all other coefficients be the same with Instance 1. We still have $n = 2$, but $p = 1$ and Π^d becomes a univariate function. The contour of $\Pi(\boldsymbol{x})$ and graph of $\Pi^d(\varsigma)$ are shown in Figure 3.2.

The primal function $\Pi(\boldsymbol{x})$ has 5 critical points,

$$\begin{aligned}
&\boldsymbol{x}_1 = (-3.109, 1.014), \quad && \boldsymbol{x}_2 = (0.973, -1.415), \quad && \boldsymbol{x}_3 = (1.068, -1.098),\\
&\boldsymbol{x}_4 = (-3.987, -2.281), && \boldsymbol{x}_5 = (-2.000, -0.875),
\end{aligned}$$

and, correspondingly, the dual function $\Pi^d(\varsigma)$ has 5 critical points,

$$\begin{aligned}
&\varsigma_1 = 1.154, \quad && \varsigma_2 = -0.256, \quad && \varsigma_3 = -0.407,\\
&\varsigma_4 = -1.862, && \varsigma_5 = -3417.4.
\end{aligned}$$

The dualities are shown in Table 3.2.

$\boldsymbol{x}$	$\Pi(\boldsymbol{x})$	optimality	ς	$\Pi^d(\varsigma)$	$\mathcal{S}_a$	optimality
$\boldsymbol{x}_1$	-26.358	global minimizer	ς_1	-26.358	$\mathcal{S}_c^+$	local maximizer
$\boldsymbol{x}_2$	8.394	local minimizer	ς_2	8.394		local minimizer
$\boldsymbol{x}_3$	8.425	saddle point	ς_3	8.425		local maximizer
$\boldsymbol{x}_4$	40.334	saddle point	ς_4	40.334	$\mathcal{S}_c^-$	local minimizer
$\boldsymbol{x}_5$	12871.9	local maximizer	ς_5	12871.9	$\mathcal{S}_c^-$	local maximizer

Table 3.2: Dualities for Instance 2.

x	$\Pi(x)$	optimality	$\boldsymbol{\varsigma}$	$\Pi^d(\boldsymbol{\varsigma})$	$\mathcal{S}_a$	optimality
x_1	-20.887	global minimizer	$\boldsymbol{\varsigma}_1$	-20.887	$\mathcal{S}_c^+$	local maximizer
x_2	6.973	local minimizer	$\boldsymbol{\varsigma}_2$	6.973	$\mathcal{S}_c^-$	saddle point
x_3	113.22	local maximizer	$\boldsymbol{\varsigma}_3$	113.22	$\mathcal{S}_c^-$	local maximizer

Table 3.3: Dualities for Instance 3.

Instance 3

At last, we have a look at an instance with the primal problem being of one dimension. The coefficients A_1, A_2, Q, $\boldsymbol{b}_1$, $\boldsymbol{b}_2$ and $\boldsymbol{f}$ are all scalars now. Let

$$A_1 = -8, A_2 = 8, Q = -8, b_1 = -5, b_2 = 2, f = 8, c_1 = -3, c_2 = 5.$$

The primal function $\Pi(x)$ is univariate and the dual function $\Pi^d(\boldsymbol{\varsigma})$ is of dimension 2.

The function $\Pi(x)$ has 3 critical points,

$$x_1 = 1.61, \qquad x_2 = -0.567, \qquad x_3 = 0.591,$$

and, correspondingly, the function $\Pi^d(\boldsymbol{\varsigma})$ has 3 critical points,

$$\boldsymbol{\varsigma}_1 = (0.317, 2.148), \qquad \boldsymbol{\varsigma}_2 = (-1.405, -2.581), \qquad \boldsymbol{\varsigma}_3 = (94.69, -4.785).$$

The dualities are shown in Table 3.3.

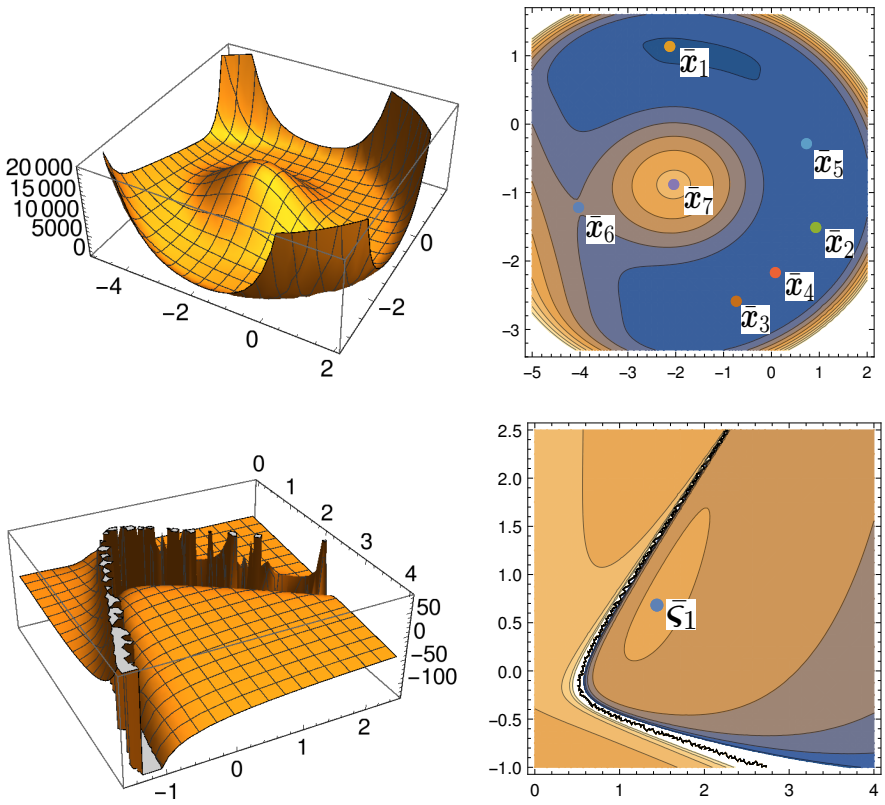

Figure 3.1: Instance 1: The two above are the graph and contour of $\Pi(\boldsymbol{x})$; the two below are the graph and contour of $\Pi^d(\varsigma_1, \varsigma_2)$ on $\mathcal{S}_c^+$.

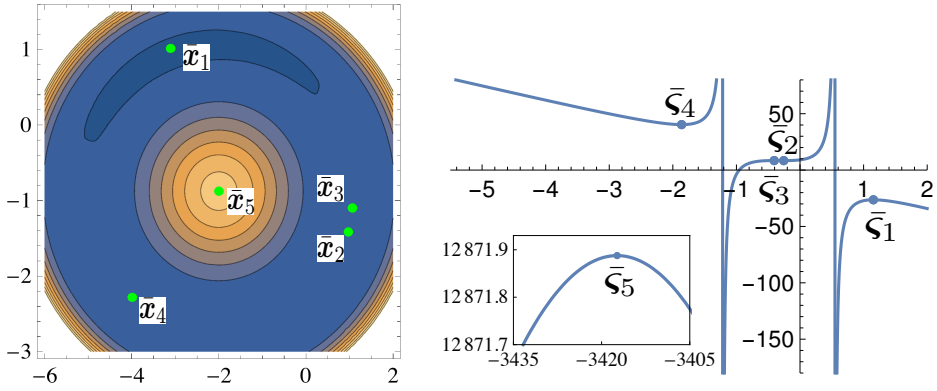

Figure 3.2: Instance 2: The left one is the contour of $\Pi(\boldsymbol{x})$; the right one is the graph of $\Pi^d(\varsigma)$.

Chapter 4

Spherically constrained quadratic minimization

4.1 Introduction

In this chapter, we consider the spherically constrained quadratic minimization problem,

$$(\mathcal{P}_{sqp}) \qquad \min_{\boldsymbol{x}} \ \Pi(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} - 2\boldsymbol{f}^T\boldsymbol{x} \qquad \text{s.t.} \ \ \boldsymbol{x} \in \mathcal{X} \tag{4.1}$$

where $\boldsymbol{Q} \in \mathbb{S}^n$ and $\boldsymbol{f} \in \mathbb{R}^n$. The feasible region is defined as

$$\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^n \ | \ \|\boldsymbol{x}\| \leq r\},$$

with r being a positive real number and $\|\boldsymbol{x}\| = \|\boldsymbol{x}\|_2$ representing ℓ_2 norm in $\mathbb{R}^n$.

Problem $(\mathcal{P}_{sqp})$ arises naturally in computational mathematical physics with extensive applications in engineering sciences. From the point of view of the systems theory, if the vector $\boldsymbol{f} \in \mathbb{R}^n$ is considered as an input (or source), then the solution $\boldsymbol{x} \in \mathbb{R}^n$ is referred to as the output (or state) of the system. By the fact that the capacity of any given system is limited, the spherical constraint in $\mathcal{X}$ is naturally required for virtually every real-world system. For example, in engineering structural analysis, if the applied force $\boldsymbol{f} \in \mathbb{R}^\infty$ is big enough, the stress distribution in the structure will reach its elastic limit and the structure will collapse. For elasto-perfectly plastic materials, the well-known von Mises yield condition is a nonlinear inequality constraint $\|\boldsymbol{x}\|_2 \leq r$ imposed on each material point[1] (see Chapter 7 of [38]). By the finite element method, the variational problem in structural limit analysis can be formulated as a large-size nonlinear optimization problem with m

[1]The well-known Tresca yield condition $\|\boldsymbol{x}\|_\infty \leq r$ is equivalent to a box constraint at each material point. It was shown in the well-known experiment by Taylor and Quinney in 1931 that the von Mises yield condition is better than the Tesca yield condition for metal structures (see [38] p. 404.)

quadratic inequality constraints (m depends on the number of total finite elements). Such problems have been studied extensively in computational mechanics for more than fifty years and the so-called penalty-duality finite element programming [35, 34] is one of the well-developed efficient methods for solving this type of problem in engineering sciences.

In mathematical programming, the problem $(\mathcal{P}_{sqp})$ is known as a trust region subproblem, which arises in trust region methods [29, 102]. In papers, two similar problems are also discussed: in [72, 131, 21], the convexity of the quadratic constraint is removed; while in [122, 15], the constraint is replaced by a two-sided (lower and upper bounded) quadratic constraint. Although the function $\Pi(\boldsymbol{x})$ may be nonconvex, it is proved that the problem $(\mathcal{P}_{sqp})$ possesses the *hidden convexity*, which means that $(\mathcal{P}_{sqp})$ is actually equivalent to a convex optimization problem [15, 130]. For each optimal solution $\bar{\boldsymbol{x}}$, there exists a Lagrange multiplier $\bar{\mu}$ such that the following conditions hold [120]:

$$(\boldsymbol{Q} + \bar{\mu}\boldsymbol{I})\bar{\boldsymbol{x}} = \boldsymbol{f}, \tag{4.2}$$

$$\boldsymbol{Q} + \bar{\mu}\boldsymbol{I} \succeq 0, \tag{4.3}$$

$$\|\bar{\boldsymbol{x}}\| \leq r, \quad \bar{\mu} \geq 0, \quad \bar{\mu}(\|\bar{\boldsymbol{x}}\| - r) = 0. \tag{4.4}$$

Let λ_1 be the smallest eigenvalue of the matrix $\boldsymbol{Q}$. From conditions (4.3) and (4.4), we have

$$\bar{\mu} \geq \max\{0, -\lambda_1\}.$$

If the problem $(\mathcal{P}_{sqp})$ has no solutions on the boundary of $\mathcal{X}$, then $\boldsymbol{Q}$ must be positive definite and $\|\boldsymbol{Q}^{-1}\boldsymbol{f}\| < r$, which leads to $\bar{\mu} = 0$. Now suppose the solution $\bar{\boldsymbol{x}}$ is on the boundary of $\mathcal{X}$. If $(\boldsymbol{Q} + \bar{\mu}\boldsymbol{I}) \succ 0$, we have $\|(\boldsymbol{Q} + \bar{\mu}\boldsymbol{I})^{-1}\boldsymbol{f}\| = r$ and the multiplier $\bar{\mu}$ can be easily found. While if $\det(\boldsymbol{Q} + \bar{\mu}\boldsymbol{I}) = 0$, it becomes very challenging to solve the problem [121, 104, 73, 107, 33] and the situation is referred to as "hard case" (see [90]). Mathematically speaking, when the problem is in the hard case, there are multiple solutions for the equation $(\boldsymbol{Q} + \bar{\mu}\boldsymbol{I})\boldsymbol{x} = \boldsymbol{f}$ and they are in the form $\boldsymbol{x} = (\boldsymbol{Q} + \bar{\mu}\boldsymbol{I})^{\dagger}\boldsymbol{f} + \tau\tilde{\boldsymbol{x}}$ with $(\boldsymbol{Q} + \bar{\mu}\boldsymbol{I})\tilde{\boldsymbol{x}} = 0$. As pointed out in [107, 121, 33, 65], the hard case always implies that $\boldsymbol{f}$ is perpendicular to the subspace generated by all the eigenvectors corresponding to λ_1. We show by Theorem 16 and Example 2 that this condition is only a necessary condition for the problem being in the hard case. Many methods have been proposed for handling the problem $(\mathcal{P}_{sqp})$, especially focusing on the hard case: Newton type methods [59, 90], methods recasting the problem in terms of a parameterized eigenvalue problem [121, 107], methods sequential searching Krylov subspaces [61, 65], semidefinite programming methods [104, 33], and the D.C. (difference of convex functions) method [124].

Here, we discuss global solutions for the problem $(\mathcal{P}_{sqp})$ via the canonical duality theory, especially when it is in the hard case. We first show in the next section that by the canonical dual transformation, this constrained nonconvex problem can be reformulated as a one-dimensional optimization problem. The complementary-dual principle shows that the one-dimensional problem is canonically dual to $(\mathcal{P}_{sqp})$

in the sense that there is a one-to-one correspondence between KKT points of the two problems and each pair of corresponding KKT points share the same function value. The canonical min-max duality in Triality Theorem provides a sufficient and necessary condition for identifying global optimal solutions. In order to solve the hard case, a perturbation method is proposed in Section 4.3 and then a canonical primal-dual algorithm is developed in Section 4.4. Numerical results are presented in Section 4.5. The chapter finishes with some conclusion remarks.

4.2 Canonical duality and optimality

4.2.1 Canonical dual problem

By the fact that the condition $\|\boldsymbol{x}\| \leq r$ is a physical constraint (required by mathematical model), it must be written in canonical form. Therefore, instead of the ℓ_2 norm, the canonical dual transformation is to introduce a quadratic measure $\xi = \Lambda(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{x} : \mathbb{R}^n \to \mathcal{E}_a = \{\xi \in \mathbb{R} \,|\, \xi \geq 0\}$

Then the total complementary function can be obtained:

$$\Xi(\boldsymbol{x}, \sigma) = \boldsymbol{x}^T\boldsymbol{G}(\sigma)\boldsymbol{x} - 2\boldsymbol{f}^T\boldsymbol{x},$$

where $\boldsymbol{G}(\sigma) = \boldsymbol{Q} + \sigma\boldsymbol{I}$. The canonical dual feasible region is

$$\mathcal{S}_a = \{\sigma \in \mathbb{R} \,|\, \det \boldsymbol{G}(\sigma) \neq 0 \}.$$

For any given $\sigma \in \mathcal{S}_a$, the canonical dual function $\Pi^d : \mathcal{S}_a \to \mathbb{R}$ is well defined and can be formulated as

$$\Pi^d(\sigma) = \text{ext}\,\{\Xi(\boldsymbol{x}, \sigma) \,|\, \boldsymbol{x} \in \mathbb{R}^n\} = -\boldsymbol{f}^T\boldsymbol{G}(\sigma)^{-1}\boldsymbol{f} - r^2\sigma.$$

The canonical dual problem is to find extreme points $\bar{\sigma}$ of $\Pi^d(\sigma)$ such that

$$\Pi^d(\bar{\sigma}) = \text{ext}\left\{\Pi^d(\sigma) \,|\, \sigma \geq 0,\, \sigma \in \mathcal{S}_a\right\}. \tag{4.5}$$

We need to emphasize that $\Pi^d(\sigma)$ is a function of a scalar variable $\sigma \in \mathcal{S}_a \subset \mathbb{R}$, regardless of the dimension of the primal problem, and the inequality $\det \boldsymbol{G}(\sigma) \neq 0$ is actually not a constraint (the Lagrange multiplier for this inequality is zero). Therefore, the KKT points for this canonical dual problem are much easier to be obtained than that for the primal problem. By the canonical duality theory, we have the following result.

Theorem 14 (**Analytical Solution and Complementary-Dual Principle** [42]) *Suppose that the symmetrical matrix $\boldsymbol{Q}$ has m ($\leq n$) distinct eigenvalues $\lambda_i, i = 1, \ldots, m$ and $i_d \leq m$ of them are strictly negative such that $\lambda_1 < \lambda_2 < \cdots < \lambda_{i_d} < 0 \leq \lambda_{i_d+1} < \cdots < \lambda_m$. Then for a given vector $\boldsymbol{f} \in \mathbb{R}^n$ and a sufficiently large $r > 0$, the canonical dual problem (4.5) has at most $2i_d + 1$ KKT points $\bar{\sigma}_i$ satisfying*

$$\bar{\sigma}_1 > -\lambda_1 > \bar{\sigma}_2 \geq \bar{\sigma}_3 > -\lambda_2 > \cdots > -\lambda_{i_d} > \bar{\sigma}_{2i_d} \geq \bar{\sigma}_{2i_d+1} > 0.$$

For each $\bar{\sigma}_i$, $i = 1, \ldots, 2i_d + 1$, *the vector*

$$\bar{\boldsymbol{x}}_i = \boldsymbol{G}(\bar{\sigma}_i)^{-1}\boldsymbol{f} \tag{4.6}$$

is a KKT point of the primal problem $(\mathcal{P}_{sqp})$*, and we have*

$$\Pi(\bar{\boldsymbol{x}}_j) \geq \Pi(\bar{\boldsymbol{x}}_i) = \Xi(\bar{\boldsymbol{x}}_i, \bar{\sigma}_i) = \Pi^d(\bar{\sigma}_i) \leq \Pi^d(\bar{\sigma}_j) \;\; \forall i, j = 1, \ldots, 2i_d + 1, \;\; i \leq j.$$

Theorem 14 shows that the nonconvex function $\Pi(\boldsymbol{x})$ is canonically dual (without duality gaps) to $\Pi^d(\sigma)$ at each KKT point $(\bar{\boldsymbol{x}}_i, \bar{\sigma}_i)$, and the function values of $\Pi^d(\sigma_i)$ are in an opposite order with its critical points $\sigma_1 > \sigma_2 \geq \ldots$ (see Figure 4.1). Clearly, the KKT solution $\bar{\boldsymbol{x}}_1$ is a global minimizer of the primal problem $(\mathcal{P}_{sqp})$.

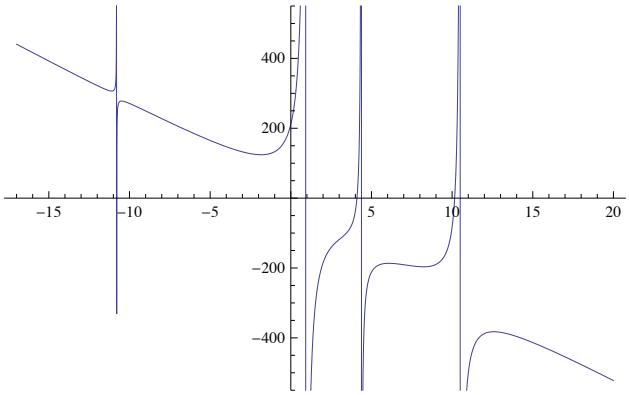

Figure 4.1: The graph of canonical dual function $\Pi^d(\sigma)$ for $n = 4$ (see Example 3 for details).

4.2.2 Global optimality condition

In order to identify global optimal solutions among all the critical points of $\Pi^d(\sigma)$, a subset of $\mathcal{S}_a$ is needed:

$$\mathcal{S}_c^+ = \{\sigma \in \mathcal{S}_a \,|\, \sigma \geq 0, \, \boldsymbol{G}(\sigma) \succ \boldsymbol{0}\}\,.$$

The problem canonically dual to $(\mathcal{P}_{sqp})$ can be proposed as the following

$$(\mathcal{P}_{sqp}^d) \qquad \max\left\{\Pi^d(\sigma) \,|\, \sigma \in \mathcal{S}_c^+\right\}. \tag{4.7}$$

Theorem 15 (**Global Optimality Condition** [38, 51]) *Suppose that* $\bar{\sigma}$ *is a critical point of* $\Pi^d(\sigma)$*. If* $\bar{\sigma} \in \mathcal{S}_c^+$*, then* $\bar{\sigma}$ *is a global maximal solution of the problem* $(\mathcal{P}_{sqp}^d)$ *and* $\bar{\boldsymbol{x}} = \boldsymbol{G}(\bar{\sigma})^{-1}\boldsymbol{f}$ *is a global minimal solution of the primal problem* $(\mathcal{P}_{sqp})$*, that is,*

$$\Pi(\bar{\boldsymbol{x}}) = \min_{\boldsymbol{x} \in \mathcal{X}} \Pi(\boldsymbol{x}) = \max_{\sigma \in \mathcal{S}_c^+} \Pi^d(\sigma) = \Pi^d(\bar{\sigma}). \tag{4.8}$$

According to Triality Theorem, the global optimality condition (4.8) is called canonical min-max duality. By the fact that $\Pi^d(\sigma)$ is strictly concave on the convex set $\mathcal{S}_c^+$, this theorem guarantees that if there is a critical point in $\mathcal{S}_c^+$, it must be unique and the nonconvex minimization problem $(\mathcal{P}_{sqp})$ is equivalent to a concave maximization problem $(\mathcal{P}_{sqp}^d)$. Similar result is also discussed by Corollary 5.3 in [122] and Theorem 1 in [104]. Moreover, for the case when $n = 1$, the double-min duality statement in the weak-triality theory proven recently (see [56, 88, 89]) shows that the problem $(\mathcal{P}_{sqp})$ has at most one local minimizer, which is corresponding to a critical point $\bar{\sigma} \in \mathcal{S}_c^- = \{\sigma \in \mathcal{S}_a \mid \boldsymbol{G}(\sigma) \prec 0\}$. All these previous results show that the canonical duality-triality theory provides detailed information on a complete set of solutions to the nonconvex problem $(\mathcal{P}_{sqp})$.

Duality theory for quadratic minimization problems with ℓ_2-norm constraints was discussed extensively in plastic mechanics fifty years ago. It was shown by Gao in [34] that for the quadratic ℓ_2^2 constraint, the canonical dual can be easily formulated and a primal-dual finite element programming algorithm was first developed for solving minimal potential variational problems in infinite dimensional space [35]. By the fact that the geometrical measure $\xi(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{x}$ is quadratic, the first term in $\Xi(\boldsymbol{x}, \sigma)$ is the so-called (generalized) *complementary gap function* [41, 53] denoted by

$$G_{ap}(\boldsymbol{x}, \sigma) = \xi(\boldsymbol{x})\sigma + \boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{G}(\sigma)\boldsymbol{x}.$$

Clearly, $G_{ap}(\boldsymbol{x}, \sigma) \geq 0 \quad \forall \boldsymbol{x} \in \mathbb{R}^n$ if and only if $\sigma \in \mathcal{S}_c^+$. Therefore, $\Xi(\boldsymbol{x}, \sigma)$ is a saddle function on $\mathbb{R}^n \times \mathbb{R}$ if $G_{ap}(\boldsymbol{x}, \sigma) \geq 0 \quad \forall \boldsymbol{x} \in \mathbb{R}^n$. This result was first discovered by Gao and Strang in nonconvex mechanics [58], where they proved that this gap function recovers a broken symmetry in geometrically nonlinear systems and provides a global optimality condition for general nonconvex variational problems in mathematical physics. Particularly, the total complementary function $\Xi(\boldsymbol{x}, \sigma)$ on $\mathbb{R}^n \times \mathbb{R}_+ = \{\sigma \in \mathbb{R} \mid \sigma \geq 0\}$ has a simple form

$$\Xi(\boldsymbol{x}, \sigma) = \boldsymbol{x}^T\boldsymbol{G}(\sigma)\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{f} - r^2\sigma = \Pi(\boldsymbol{x}) + \sigma(\boldsymbol{x}^T\boldsymbol{x} - r^2),$$

which can be viewed as a Lagrangian of $(\mathcal{P}_{sqp})$ for the ℓ_2^2-norm constraint $\boldsymbol{x}^T\boldsymbol{x} \leq r^2$. Indeed, the total complementary function $\Xi(\boldsymbol{x}, \sigma)$ was also called nonlinear Lagrangian in [38] or extended Lagrangian in [37]. However, for nonconvex objective function $\Pi(\boldsymbol{x})$, the classical Lagrangian duality theory will produce a well-known duality gap unless the global optimality condition $G_{ap}(\boldsymbol{x}, \sigma) \geq 0 \quad \forall \boldsymbol{x} \in \mathbb{R}^n$ is satisfied. Therefore, the Lagrangian duality theory is only a special case of the canonical duality theory for certain problems. Also, by the fact that a large class of nonconvex/discrete global optimization problems can be equivalently reformulated as a unified canonical dual form (4.7) (see [41, 46, 53]), which is equivalent to a convex minimization problem over a convex feasible set, the so-called "hidden-convexity" is indeed a special case of the canonical min-max duality theory.

For the hard case, the matrix $\boldsymbol{G}(\sigma)$ is singular at the KKT point $\bar{\sigma}$, the canonical dual $\Pi^d(\sigma)$ should be replaced by (see [47])

$$\Pi^d(\sigma) = -\boldsymbol{f}^T\boldsymbol{G}(\sigma)^\dagger\boldsymbol{f} - r^2\sigma,$$

where $\boldsymbol{G}(\sigma)^\dagger$ stands for the pseudo-inverse of $\boldsymbol{G}(\sigma)$. In [104, 122], the dual function is also presented in discussions of the strong duality. Since this function is not strictly concave on the closure of $\mathcal{S}_c^+$, it may have multiple critical points located on the boundary of $\mathcal{S}_c^+$. In the following sections, we will first study the existence conditions of these critical points, and then study an associated algorithm for computing these solutions.

4.2.3 Existence conditions

As $\boldsymbol{Q}$ is symmetrical, there exist a diagonal matrix Λ and an orthogonal matrix $\boldsymbol{U}$ such that $\boldsymbol{Q} = \boldsymbol{U}\Lambda\boldsymbol{U}^T$. The diagonal entities of Λ are the eigenvalues of $\boldsymbol{Q}$ and are arranged in a nondecreasing order,

$$\lambda_1 = \cdots = \lambda_k < \lambda_{k+1} \leq \cdots \leq \lambda_n.$$

The columns of $\boldsymbol{U}$ are corresponding eigenvectors.

Let $\hat{\boldsymbol{f}} = \boldsymbol{U}^T\boldsymbol{f}$. Because $(\boldsymbol{Q} + \sigma\boldsymbol{I})^{-1} = \boldsymbol{U}(\Lambda + \sigma\boldsymbol{I})^{-1}\boldsymbol{U}^T$, we can rewrite the canonical dual function $\Pi^d(\sigma)$ as

$$\Pi^d(\sigma) = -\frac{\sum_{i=1}^k \hat{f}_i^2}{\lambda_1 + \sigma} - \sum_{i=k+1}^n \frac{\hat{f}_i^2}{\lambda_i + \sigma} - r^2\sigma, \tag{4.9}$$

where $\hat{f}_i, i = 1, \ldots, n$ are elements of $\hat{\boldsymbol{f}}$. It is now easy to see that as long as $\boldsymbol{f} \neq 0$, $\Pi^d(\sigma)$ has stationary points in $\mathcal{S}$ and thus the canonical dual problem (4.5) is well defined. Whereas, for the case when $\boldsymbol{f} = 0$, a perturbation should be introduced, which will be discussed in the next section.

Theorem 16 (**Existence Conditions**) *Suppose that for any given $\boldsymbol{Q} \in \mathbb{S}^n$ and $\boldsymbol{f} \in \mathbb{R}^n$, λ_i and $\hat{f}_i$ are defined as above.*

The canonical dual function $\Pi^d(\sigma)$ has a critical point $\bar{\sigma}$ in $(-\lambda_1, +\infty)$ if and only if either $\sum_{i=1}^k \hat{f}_i^2 \neq 0$ or $\sum_{i=k+1}^n \frac{\hat{f}_i^2}{(\lambda_i - \lambda_1)^2} > r^2$ holds true. Furthermore, if $\lambda_1 \leq 0$, then $\bar{\boldsymbol{x}} = \boldsymbol{G}(\bar{\sigma})^{-1}\boldsymbol{f}$ is the unique solution of the primal problem ($\mathcal{P}_{sqp}$).

If $\Pi^d(\sigma)$ has no critical points in $(-\lambda_1, +\infty)$, the primal problem ($\mathcal{P}_{sqp}$) has exactly two global solutions when the multiplicity of λ_1 is $k = 1$ and has infinite number of solutions when $k > 1$.

Proof: First, we prove that the existence of a critical point of $\Pi^d(\sigma)$ in $(-\lambda_1, +\infty)$ implies that either $\sum_{i=1}^k \hat{f}_i^2 \neq 0$ or $\sum_{i=k+1}^n \frac{\hat{f}_i^2}{(\lambda_i - \lambda_1)^2} > r^2$ holds true. It is equivalent to prove that if $\sum_{i=1}^k \hat{f}_i^2 = 0$ and $\sum_{i=k+1}^n \frac{\hat{f}_i^2}{(\lambda_i - \lambda_1)^2} \leq r^2$ the dual function $\Pi^d(\sigma)$ will have no critical points in $(-\lambda_1, +\infty)$. The first item in the expression (4.9) vanishes when $\sum_{i=1}^k \hat{f}_i^2 = 0$. Then because $\sum_{i=k+1}^n \frac{\hat{f}_i^2}{(\lambda_i - \lambda_1)^2} \leq r^2$, the first-order derivative of the dual function

$$(\Pi^d(\sigma))' = \sum_{i=k+1}^n \frac{\hat{f}_i^2}{(\lambda_i + \sigma)^2} - r^2$$

is always negative in $(-\lambda_1, +\infty)$. Therefore, the dual function $\Pi^d(\sigma)$ will have no critical points in $(-\lambda_1, +\infty)$.

Next we will give the proof of the sufficiency, which is divided into two parts:

1) If $\sum_{i=1}^{k} \hat{f}_i^2 \neq 0$, then $\sigma = -\lambda_1$ is a pole of $\Pi^d(\sigma)$, i.e., as σ approaches $-\lambda_1$ from the right side, $\Pi^d(\sigma)$ approaches $-\infty$. The value of $\Pi^d(\sigma)$ also approaches $-\infty$, when σ approaches $+\infty$. Thus, $-\Pi^d(\sigma)$ is coercive on $(-\lambda_1, +\infty)$. Since, for any $\sigma \in (-\lambda_1, +\infty)$, $\boldsymbol{G}(\sigma)$ is positive definite, $\Pi^d(\sigma)$ is strictly concave on $(-\lambda_1, +\infty)$. Thus there exists a unique critical point in $(-\lambda_1, +\infty)$.

2) If $\sum_{i=1}^{k} \hat{f}_i^2 = 0$ and $\sum_{i=k+1}^{n} \frac{\hat{f}_i^2}{(\lambda_i - \lambda_1)^2} > r^2$, $(\Pi^d(\sigma))'$ is positive at $\sigma = -\lambda_1$. Moreover, $(\Pi^d(\sigma))'$ approaches $-r^2$ as σ approaches ∞. Therefore, there exists at least one root for the equation $(\Pi^d(\sigma))' = 0$ in $(-\lambda_1, +\infty)$, which means $\Pi^d(\sigma)$ has at least one critical point in $(-\lambda_1, +\infty)$. Similarly, because of the strict concavity of $\Pi^d(\sigma)$ over $(-\lambda_1, +\infty)$, the critical point is unique.

Suppose $\lambda_1 \leq 0$. The uniqueness of global solution $\bar{\boldsymbol{x}}$ will be proved, if it can be proved that $(\bar{\boldsymbol{x}}, \bar{\sigma})$ is the only pair that satisfies the KKT conditions (4.2-4.4). As mentioned above, the dual function $\Pi^d(\sigma)$ is strictly concave on $(-\lambda_1, +\infty)$, which, plus the criticality of $\bar{\sigma}$, implies that $(\Pi^d(\sigma))' = \|\boldsymbol{x}\|^2 - r^2 > 0$ for $\sigma \in (-\lambda_1, \bar{\sigma})$ and < 0 for $\sigma \in (\bar{\sigma}, +\infty)$, where $\boldsymbol{x} = \boldsymbol{G}(\sigma)^{-1}\boldsymbol{f}$. Thus, for any $\sigma \neq \bar{\sigma}$ in $(-\lambda_1, +\infty)$, there is no $\boldsymbol{x}$ such that $(\boldsymbol{x}, \sigma)$ satisfies the KKT conditions (4.2-4.4). Except for the interval $(-\lambda_1, +\infty)$, $\sigma = -\lambda_1$ is the last candidate. However, if $\sum_{i=1}^{k} \hat{f}_i^2 \neq 0$, the equation $\boldsymbol{G}(-\lambda_1)\boldsymbol{x} = \boldsymbol{f}$ has no solutions, and if $\sum_{i=1}^{k} \hat{f}_i^2 = 0$ and $\sum_{i=k+1}^{n} \frac{\hat{f}_i^2}{(\lambda_i - \lambda_1)^2} > r^2$, the feasibility of any solution of $\boldsymbol{G}(-\lambda_1)\boldsymbol{x} = \boldsymbol{f}$ is violated by the fact that $\|\boldsymbol{x}\|^2 - r^2 = \sum_{i=k+1}^{n} \frac{\hat{f}_i^2}{(\lambda_i - \lambda_1)^2} - r^2 > 0$. Then, $\sigma = -\lambda_1$ can not make the KKT conditions hold true. Therefore, $(\bar{\boldsymbol{x}}, \bar{\sigma})$ is the unique pair that satisfies the KKT conditions (4.2-4.4).

Finally, suppose that there are no critical points in $(-\lambda_1, +\infty)$, which, from the above proof, is equivalent to $\sum_{i=1}^{k} \hat{f}_i^2 = 0$ and $\sum_{i=k+1}^{n} \frac{\hat{f}_i^2}{(\lambda_i - \lambda_1)^2} \leq r^2$. Then, for any global solution, we have $\bar{\sigma} = -\lambda_1$. Let $\bar{\boldsymbol{x}}$ be a global solution and $\bar{\boldsymbol{y}} = \boldsymbol{U}^T\bar{\boldsymbol{x}}$. Then the canonical equilibrium equation $\boldsymbol{G}(\bar{\sigma})\bar{\boldsymbol{x}} = \boldsymbol{f}$ can be equivalently transformed into $\mathrm{diag}(\{\lambda_i + \bar{\sigma}\})\bar{\boldsymbol{y}} = \hat{\boldsymbol{f}}$. If $k = 1$, i.e., the multiplicity of λ_1 is one, the equation uniquely determines $\bar{y}_i, i = 2, \ldots, n$, but not $\bar{y}_1$. By the fact that $\bar{\boldsymbol{y}}^T\bar{\boldsymbol{y}} = r^2$, $\bar{y}_1$ has exactly two values, corresponding to the two global solutions of $(\mathcal{P}_{sqp})$. While, if $k > 1$, i.e., the matrix $\boldsymbol{Q}$ has at least two repeated eigenvalues $\lambda_1 = \lambda_2 = \cdots = \lambda_k \leq 0$, the equations $\mathrm{diag}(\{\lambda_i + \bar{\sigma}\})\bar{\boldsymbol{y}} = \hat{\boldsymbol{f}}$ and $\bar{\boldsymbol{y}}^T\bar{\boldsymbol{y}} = r^2$ have infinite number of solutions. □

The complementarity relations between the primal problem $(\mathcal{P}_{sqp})$ and its canonical dual problem $(\mathcal{P}_{sqp}^d)$ are significant. When $\lambda_1 > 0$, i.e., $\boldsymbol{Q}$ is positive definite, if $(\mathcal{P}_{sqp})$ has a global solution in the interior of $\mathcal{X}$, which must be the stationary point of $\Pi(\boldsymbol{x})$ and can be easily calculated, its canonical dual $(\mathcal{P}_{sqp}^d)$ has no critical point in $\mathcal{S}_c^+ = [0, +\infty)$ due to $(\Pi^d(0))' = \|\bar{\boldsymbol{x}}\|^2 - r^2 < 0$, where $\bar{\boldsymbol{x}} = \boldsymbol{G}(0)^{-1}\boldsymbol{f}$ is the stationary point of $\Pi(\boldsymbol{x})$. Dually, when $\lambda_1 \leq 0$, the primal function $\Pi(\boldsymbol{x})$ is nonconvex and the global minimizer of $(\mathcal{P}_{sqp})$ must be on the boundary of $\mathcal{X}$. In this case, if the canonical dual $(\mathcal{P}_{sqp}^d)$ has a critical point in $\mathcal{S}_c^+ = (-\lambda_1, +\infty)$, the primal problem $(\mathcal{P}_{sqp})$

is then not in the hard case and has a unique solution, which can be easily obtained by solving the canonical dual problem. Whereas if $(\mathcal{P}^d_{sqp})$ has no critical points in $\mathcal{S}_c^+$, i.e., $\Pi^d(-\lambda_1) = \sup\{\Pi^d(\sigma) \mid \sigma \in \mathcal{S}_c^+\}$, the primal problem $(\mathcal{P}_{sqp})$ is in the hard case, because, for any $\sigma \in \mathcal{S}_c^+$ and $\boldsymbol{x} = \boldsymbol{G}(\sigma)^{-1}\boldsymbol{f}$, we have $(\Pi^d(\sigma))' = \|\boldsymbol{x}\|^2 - r^2 < 0$, which destroys the complementary condition in (4.4), and only $\sigma = -\lambda_1$ can make the KKT conditions (4.2-4.4) hold.

Therefore, combining with Theorem 16, we have the following result.

Corollary 17 *If $\lambda_1 \leq 0$, the nonconvex problem $(\mathcal{P}_{sqp})$ is in the hard case if and only if both conditions (i) $\sum_{i=1}^{k} \hat{f}_i^2 = 0$ and (ii) $\sum_{i=k+1}^{n} \frac{\hat{f}_i^2}{(\lambda_i - \lambda_1)^2} \leq r^2$ hold true.*

The condition (i) is well-known: the trust region subproblem could be in the hard case only if the coefficient $\boldsymbol{f}$ is perpendicular to the subspace generated by eigenvectors of the smallest eigenvalue. The condition (ii) is new, which shows that the hard case of $(\mathcal{P}_{sqp})$ depends not only on the direction of $\boldsymbol{f}$, but also on its norm.

Theorem 16 and Corollary 17 show an important fact that the given vector $\boldsymbol{f}$ plays an important role to the solutions of the problem $(\mathcal{P}_{sqp})$. From the point of view of solid mechanics, if $\boldsymbol{f}$ is considered as an applied force, then the decision variable $\boldsymbol{x}$ is the displacement and the spherical constraint $\|\boldsymbol{x}\| \leq r$ is corresponding to the von Mises yield condition, which represents the capacity of the system. If the norm of $\boldsymbol{f}$ is big enough, the deformation $\boldsymbol{x}$ should reach the limit $\|\boldsymbol{x}\| = r$ and the problem $(\mathcal{P}_{sqp})$ has a solution on the boundary of $\mathcal{X}$. By the canonical duality, the problem $(\mathcal{P}^d_{sqp})$ must have a critical point in $\mathcal{S}_c^+$. If the norm of $\boldsymbol{f}$ is too small, the primal problem $(\mathcal{P}_{sqp})$ could have multiple solutions. In this case, $(\mathcal{P}^d_{sqp})$ has no critical point in $\mathcal{S}_c^+$ and $(\mathcal{P}_{sqp})$ could be in the hard case.

To illustrate Theorem 16, let us consider a 3-dimensional problem with coefficients

$$\boldsymbol{Q} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{f} = \begin{pmatrix} 0 \\ 0 \\ -1.8 \end{pmatrix}, \text{ and } r = 2.$$

In this case, the eigenvalues of $\boldsymbol{Q}$ are $\lambda_1 = \lambda_2 = -1$, and $\lambda_3 = 1$. So we have $k = 2$ and the target function

$$\Pi(\boldsymbol{x}) = -\frac{1}{2}(x_1^2 + x_2^2) + \frac{1}{2}x_3^2 + 1.8x_3$$

is nonconvex, whose minimizers are on the boundary of the feasible region. Replacing $x_1^2 + x_2^2$ with $r^2 - x_3^2$, the objective function $\Pi(\boldsymbol{x})$ can be reformulated as a univariate function of x_3,

$$g(x_3) = x_3^2 + 1.8x_3 - 2,$$

which achieves the minimum at $x_3 = -0.9$. Then we obtain the following equation

$$x_1^2 + x_2^2 = r^2 - x_3^2 = 2^2 - (-0.9)^2 = 3.19.$$

So all $\bar{\boldsymbol{x}} \in \mathbb{R}^3$ satisfying $\bar{x}_1^2 + \bar{x}_2^2 = 3.19$ and $\bar{x}_3 = -0.9$ are global minimizers of the problem.

By the fact that $\sum_{i=1}^{2} \hat{f}_i^2 = 0$ and $\sum_{i=2+1}^{3} \frac{\hat{f}_i^2}{(\lambda_i - \lambda_1)^2} = (-1.8)^2/(1+1)^2 \leq r^2 = 4$, Theorem 16 shows that $\Pi^d(\sigma)$ has no critical point in $\mathcal{S}_c^+$, and $(\mathcal{P}_{sqp})$ is indeed in the hard case and has infinite number of global solutions. If we choose either a smaller r or a vector $\boldsymbol{f}$ with a larger magnitude such that $\sum_{i=2+1}^{3} \frac{\hat{f}_i^2}{(\lambda_i - \lambda_1)^2} > r^2$, the global solution will be unique. For example, let $r = 0.5$. Then $x_3 = -0.9$ is no longer the minimizer of $g(x_3)$ and the problem $\min\{g(x_3) \mid x_3^2 \leq 0.5^2\}$ leads to $x_3 = -0.5$. From $x_1^2 + x_2^2 = r^2 - x_3^2 = 0.5^2 - (-0.5)^2 = 0$, we know the unique global solution of $(\mathcal{P}_{sqp})$ is $\bar{\boldsymbol{x}} = (0, 0, -0.5)^T$.

In [87], Martinez investigated the "local-nonglobal minimizers" of the problem $(\mathcal{P}_{sqp})$, of which the main result (Theorem 3.1 in [87]) can be restated in the following theorem.

Theorem 18 *(i) If $\bar{\boldsymbol{x}}$ is a local-nonglobal minimizer of $(\mathcal{P}_{sqp})$, then there is a $\bar{\sigma} \in (\max\{0, -\lambda_2\}, -\lambda_1)$ such that $\boldsymbol{G}(\bar{\sigma})\bar{\boldsymbol{x}} = \boldsymbol{f}$ and $(\Pi^d(\bar{\sigma}))'' \geq 0$. (ii) There exists at most one local-nonglobal minimizer of $(\mathcal{P}_{sqp})$. (iii) If $\|\bar{\boldsymbol{x}}\| = r$, $\boldsymbol{G}(\bar{\sigma})\bar{\boldsymbol{x}} = \boldsymbol{f}$ for some $\bar{\sigma} \in (-\lambda_2, -\lambda_1)$, $\bar{\sigma} > 0$ and $(\Pi^d(\bar{\sigma}))'' > 0$, then $\bar{\boldsymbol{x}}$ is a strict local minimizer of $(\mathcal{P}_{sqp})$.*

From the point of view of the canonical duality theory, the $\bar{\sigma}$ in this theorem is actually a critical point of $\Pi^d(\sigma)$. The case of $(\mathcal{P}_{sqp})$ having no local-nonglobal minimizers implies that all the local minimizers are global solutions. The situations that leads to this case include i) the multiplicity of λ_1 being larger than one; ii) no critical point in $(\max\{0, -\lambda_2\}, -\lambda_1)$, and iii) $\boldsymbol{f}$ being perpendicular to the eigenvector of λ_1. The first situation results in $(-\lambda_2, -\lambda_1) = \emptyset$. The last situation violates the necessary condition $(\Pi^d(\sigma))'' \geq 0$, which can be observed from the expression of $(\Pi^d(\sigma))''$,

$$(\Pi^d(\sigma))'' = -2\sum_{i=1}^{n} \frac{\hat{f}_i^2}{(\lambda_i + \sigma)^3}.$$

For any $\sigma \in (-\lambda_2, -\lambda_1)$, the only nonnegative item in $(\Pi^d(\sigma))''$ is the first term $-2\hat{f}_1^2/(\lambda_1 + \sigma)^3$. Thus $(\Pi^d(\sigma))''$ will be negative if $\hat{f}_1^2 = 0$. As shown in Figure 4.1, there is a critical point $\bar{\sigma}_2 \in (-\lambda_2, -\lambda_1) = (4.37, 10.51)$ and the corresponding solution $\bar{\boldsymbol{x}}_2$ obtained from the equation (4.6) is a local minimizer.

4.2.4 A quartic polynomial minimization

A closely related problem is the following quartic polynomial minimization:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \Pi_1(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T Q \boldsymbol{x} - \boldsymbol{p}^T \boldsymbol{x} + \frac{1}{2}\left(\frac{1}{2}\boldsymbol{x}^T \boldsymbol{x} - \boldsymbol{x}^T \boldsymbol{b} - c\right)^2 \tag{4.10}$$

where $Q \in \mathbb{S}^n$, $\boldsymbol{p}, \boldsymbol{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Here, in the fourth-order item, the coefficient matrix could be any positive definite matrix, which can be transformed into an identity matrix without changing the problem. The canonical dual function is

$$\Pi_1^d(\sigma) = -\frac{1}{2}\boldsymbol{f}^T G^{-1}\boldsymbol{f} - c\sigma - \frac{1}{2}\sigma^2,$$

in which

$$G = G(\sigma) = Q + \sigma I, \text{ and } \boldsymbol{f} = \boldsymbol{f}(\sigma) = \boldsymbol{p} + \sigma \boldsymbol{b}.$$

Similarly, assume that Q has the eigendecomposition of $Q = U\Lambda U^T$, with the diagonal entities of Λ being in nondecreasing order,

$$\lambda_1 = \cdots = \lambda_k < \lambda_{k+1} \leq \cdots \leq \lambda_n.$$

Let $\hat{\boldsymbol{p}} = U^T\boldsymbol{p}$ and $\hat{\boldsymbol{b}} = U^T\boldsymbol{b}$. The dual function can then be rewritten as

$$\Pi_1^d(\sigma) = -\frac{1}{2}\sum_{i=1}^{n} \frac{(\hat{p}_i + \sigma\hat{b}_i)^2}{\lambda_i + \sigma} - \frac{1}{2}\sigma^2 - c\sigma. \tag{4.11}$$

The first- and second-order derivatives of the dual function $\Pi_1^d(\sigma)$ are

$$\nabla\Pi_1^d(\sigma) = \frac{1}{2}\sum_{i=1}^{n} \frac{(\hat{p}_i - \lambda_i\hat{b}_i)^2}{(\lambda_i + \sigma)^2} - \sigma - \frac{1}{2}\sum_{i=1}^{n}\hat{b}_i^2 - c,$$

$$\nabla^2\Pi_1^d(\sigma) = -\sum_{i=1}^{n} \frac{(\hat{p}_i - \lambda_i\hat{b}_i)^2}{(\lambda_i + \sigma)^3} - 1.$$

Then we have the following result, which is similar to Theorem 16.

Theorem 19 *Suppose that λ_i are defined as above. Then there exists a critical point of $\Pi_1^d(\sigma)$ in $\mathcal{S}_c^+$ if and only if*

$$\sum_{i=1}^{k}(\hat{p}_i - \lambda_1\hat{b}_i)^2 \neq 0 \text{ or } \frac{1}{2}\sum_{i=k+1}^{n} \frac{(\hat{p}_i - \lambda_i\hat{b}_i)^2}{(\lambda_i + \sigma)^2} - \sigma - \frac{1}{2}\sum_{i=1}^{n}\hat{b}_i^2 - c > 0. \tag{4.12}$$

If $\Pi_1^d(\sigma)$ has a critical point in $\mathcal{S}_c^+$, the critical point is unique. Let $\bar{\sigma}$ denote the critical point. Then $\bar{\boldsymbol{x}} = G(\bar{\sigma})^{-1}\boldsymbol{f}(\bar{\sigma})$ is a global solution of the problem (4.10).

Notice that if $\lambda_1 - \frac{1}{2}\sum_{i=1}^{n}\hat{b}_i^2 - c > 0$, the second condition in (4.12) holds and thus there must be a critical point in $\mathcal{S}_c^+$.

4.3 A perturbation method

This section is devoted to compute solutions for the problem when the canonical dual problem $(\mathcal{P}^d_{sqp})$ has no critical point in $(-\lambda_1, +\infty)$. Since a necessary condition for the hard case is $\sum_{i=1}^k \hat{f}_i^2 = 0$, a perturbation can be introduced such that this condition does not hold true any more. Impressively, once we obtain the critical point in $\mathcal{S}_c^+$, all the global solutions can be determined. Our approach has been applied successfully in the canonical duality theory for solving nonlinear algebraic equations [119], chaotic dynamical systems [116], as well as a class of NP-hard problems in the global optimization [47, 117, 128].

In order to establish the existence conditions, a perturbation $\sum_{i=1}^k \alpha_i \boldsymbol{U}_i$ with parameters

$$\boldsymbol{\alpha} = \{\alpha_i\}_{i=1}^k \neq 0 \tag{4.13}$$

is introduced to $\boldsymbol{f}$. Let

$$\boldsymbol{p} = \boldsymbol{f} + \sum_{i=1}^k \alpha_i \boldsymbol{U}_i, \quad \hat{\boldsymbol{p}} = \boldsymbol{U}^T \boldsymbol{p}, \text{ and } \Pi_\alpha(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} - 2\boldsymbol{p}^T \boldsymbol{x}.$$

It is true that the existence conditions hold for the perturbed problem

$$(\mathcal{P}_\alpha) \qquad \min\{\Pi_\alpha(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{X}_a\}, \tag{4.14}$$

for $\sum_{i=1}^k \hat{p}_i^2 \neq 0$ is guaranteed by (4.13).

The following theorem states that if the parameter $\boldsymbol{\alpha}$ is chosen appropriately, the optimal solution of the perturbed problem approximates that of the primal problem $(\mathcal{P}_{sqp})$.

Theorem 20 *Suppose that $\lambda_1 \leq 0$, there is no critical point of $\Pi^d(\sigma)$ in $\mathcal{S}_c^+$, and $\bar{\boldsymbol{x}}^*$ is the optimal solution of the problem $(\mathcal{P}_\alpha)$. Then, there is a global solution of the problem $(\mathcal{P}_{sqp})$, denoted as $\bar{\boldsymbol{x}}$, which is on the boundary of $\mathcal{X}$ and, for any $\varepsilon > 0$, if the parameter $\boldsymbol{\alpha}$ satisfies*

$$\|\boldsymbol{\alpha}\|^2 \leq (\lambda_2 - \lambda_1)^2 \left(r^2 - \sum_{i=k+1}^n \frac{\hat{f}_i^2}{(\lambda_i - \lambda_1)^2} \right) (1/\sqrt{2(1 - \cos(\varepsilon/r))} - 1)^{-2}, \tag{4.15}$$

we have $\|\bar{\boldsymbol{x}}^* - \bar{\boldsymbol{x}}\| \leq \varepsilon$.

Proof: For simplicity, the coordinate system is rotated and let $\boldsymbol{y} = \boldsymbol{U}^T \boldsymbol{x}$, $\boldsymbol{y}_k = \{y_i\}_{i=1}^k$ and $\boldsymbol{y}_\ell = \{y_i\}_{i=k+1}^n$. Since $\hat{f}_i = 0$ for $i = 1, \ldots, k$, variables y_i for $i = 1, \ldots, k$ appear in the objective function only in the form of squares. On the boundary of $\mathcal{X}$, the problem $(\mathcal{P}_{sqp})$ is then equivalent to the following problem in $\mathbb{R}^{n-k}$:

$$\min_{\|\boldsymbol{y}_\ell\| \leq r} \Pi^\ell(\boldsymbol{y}_\ell) = \sum_{i=k+1}^n (\lambda_i - \lambda_1) y_i^2 - \sum_{i=k+1}^n 2\hat{f}_i y_i + \lambda_1 r^2. \tag{4.16}$$

Since $\Pi^\ell(\boldsymbol{y}_\ell)$ is a strictly convex function, it has a unique stationary point,

$$\bar{\boldsymbol{y}}_\ell = \left\{ \frac{\hat{f}_i}{\lambda_i - \lambda_1} \right\}_{i=k+1}^{n}.$$

Combining with the assumption of no critical point in $\mathcal{S}_c^+$, we know that this stationary point is the global optimal solution of the problem (4.16). Then, all $\bar{\boldsymbol{y}}$ that satisfies $\bar{\boldsymbol{y}}_k^T \bar{\boldsymbol{y}}_k = r^2 - \bar{\boldsymbol{y}}_\ell^T \bar{\boldsymbol{y}}_\ell$ are solutions of the problem $(\mathcal{P}_{sqp})$. Here we choose one particular solution with

$$\bar{\boldsymbol{y}}_k = h\bar{\boldsymbol{y}}_k^*, \quad h = \frac{1}{\|\bar{\boldsymbol{y}}_k^*\|}\sqrt{r^2 - \bar{\boldsymbol{y}}_\ell^T \bar{\boldsymbol{y}}_\ell},$$

where $\bar{\boldsymbol{y}}^* = U\bar{\boldsymbol{x}}^*$, and let $\bar{\boldsymbol{x}} = \boldsymbol{U}\bar{\boldsymbol{y}}$.

By canceling variables $y_i, i = 1, \ldots, k$, the perturbed problem (4.14) with the equality constraint is equivalent to

$$\min_{\|\boldsymbol{y}_\ell\| \le r} \Pi_\alpha^\ell(\boldsymbol{y}_\ell) = \sum_{i=k+1}^{n} (\lambda_i - \lambda_1) y_i^2 - \sum_{i=k+1}^{n} 2\hat{f}_i y_i + \lambda_1 r^2 - 2\|\boldsymbol{\alpha}\| \sqrt{r^2 - \boldsymbol{y}_\ell^T \boldsymbol{y}_\ell}. \quad (4.17)$$

The function $\Pi_\alpha^\ell(\boldsymbol{y}_\ell)$ is also strictly convex. Moreover, for any $\|\boldsymbol{y}_\ell\| < r$, we have $\Pi_\alpha^\ell(\boldsymbol{y}_\ell) < \Pi^\ell(\boldsymbol{y}_\ell)$, while for any $\|\boldsymbol{y}_\ell\| = r$, we have $\Pi_\alpha^\ell(\boldsymbol{y}_\ell) = \Pi^\ell(\boldsymbol{y}_\ell)$. The fact indicates that the unique stationary point of $\Pi_\alpha^\ell(\boldsymbol{y}_\ell)$ is in the interior of $\|\boldsymbol{y}_\ell\| \le r$. Thus the global solution $\bar{\boldsymbol{y}}_\ell^*$ is a stationary point of the problem (4.17) and then satisfies

$$\bar{y}_i^* = \frac{\hat{f}_i}{\lambda_i - \lambda_1 + \|\boldsymbol{\alpha}\|(r^2 - \bar{\boldsymbol{y}}_\ell^{*T} \bar{\boldsymbol{y}}_\ell^*)^{-\frac{1}{2}}}, i = k+1, \ldots, n.$$

and

$$|\bar{y}_i^*| < |\bar{y}_i|, i = k+1, \ldots, n. \quad (4.18)$$

We will prove that as $\|\boldsymbol{\alpha}\|$ approaches zero, $\bar{\boldsymbol{y}}^*$ will approach $\bar{\boldsymbol{y}}$. First, we have the following relation

$$\begin{aligned} \bar{\boldsymbol{y}}^{*T} \bar{\boldsymbol{y}} &= \sqrt{r^2 - \bar{\boldsymbol{y}}_\ell^{*T} \bar{\boldsymbol{y}}_\ell^*} \sqrt{r^2 - \bar{\boldsymbol{y}}_\ell^T \bar{\boldsymbol{y}}_\ell} + \bar{\boldsymbol{y}}_\ell^{*T} \bar{\boldsymbol{y}}_\ell \\ &\le \frac{1}{2}\left(r^2 - \bar{\boldsymbol{y}}_\ell^{*T} \bar{\boldsymbol{y}}_\ell^* + r^2 - \bar{\boldsymbol{y}}_\ell^T \bar{\boldsymbol{y}}_\ell\right) + \bar{\boldsymbol{y}}_\ell^{*T} \bar{\boldsymbol{y}}_\ell \\ &= r^2 - \frac{1}{2}\|\bar{\boldsymbol{y}}_\ell^* - \bar{\boldsymbol{y}}_\ell\|^2, \end{aligned}$$

where the first equality is derived from the definition of $\bar{\boldsymbol{y}}_k$ and the fact that $\bar{\boldsymbol{y}}^*$ locates on the surface of the sphere. Based on the relation

$$\|\bar{\boldsymbol{y}}^* - \bar{\boldsymbol{y}}\| \le r \arccos\left(\frac{\bar{\boldsymbol{y}}^{*T} \bar{\boldsymbol{y}}}{r^2}\right) \le r \arccos\left(\frac{r^2 - \frac{1}{2}\|\bar{\boldsymbol{y}}_\ell^* - \bar{\boldsymbol{y}}_\ell\|^2}{r^2}\right),$$

we will have $\|\bar{\boldsymbol{y}}^* - \bar{\boldsymbol{y}}\| \leq \varepsilon$, if $\|\bar{\boldsymbol{y}}_\ell^* - \bar{\boldsymbol{y}}_\ell\|^2 \leq 2r^2(1-\cos\frac{\varepsilon}{r})$. Then, it can be verified that

$$\|\bar{\boldsymbol{y}}_\ell^* - \bar{\boldsymbol{y}}_\ell\|^2 \leq \frac{r^2}{\left((\lambda_2-\lambda_1)\|\boldsymbol{\alpha}\|^{-1}\sqrt{r^2 - \bar{\boldsymbol{y}}_\ell^{*T}\bar{\boldsymbol{y}}_\ell^*} + 1\right)^2}. \tag{4.19}$$

If let the right side of equation (4.19) be less than or equal to $2r^2(1-\cos\frac{\varepsilon}{r})$, we obtain

$$\|\boldsymbol{\alpha}\|^2 \leq \frac{(\lambda_2-\lambda_1)^2(r^2 - \bar{\boldsymbol{y}}_\ell^{*T}\bar{\boldsymbol{y}}_\ell^*)}{(1/\sqrt{2(1-\cos\frac{\varepsilon}{r})} - 1)^2}.$$

Combining with relations in (4.18), we can state that $\|\bar{\boldsymbol{y}}^* - \bar{\boldsymbol{y}}\| \leq \varepsilon$ if the following inequality is true

$$\|\boldsymbol{\alpha}\|^2 \leq \frac{(\lambda_2-\lambda_1)^2(r^2 - \sum_{i=k+1}^{n}\frac{\hat{f}_i^2}{(\lambda_i-\lambda_1)^2})}{(1/\sqrt{2(1-\cos\frac{\varepsilon}{r})} - 1)^2}. \tag{4.20}$$

Since $\|\bar{\boldsymbol{x}}^* - \bar{\boldsymbol{x}}\| = \|\bar{\boldsymbol{y}}^* - \bar{\boldsymbol{y}}\|$, the equation (4.20) implies that $\|\bar{\boldsymbol{x}}^* - \bar{\boldsymbol{x}}\| \leq \varepsilon$. □

Theorem 20 shows that with a proper parameter $\boldsymbol{\alpha}$, the existence condition is guaranteed to hold true for the perturbed problem and the perturbation method can be used to solve the hard case approximately. As the perturbation parameters approach zero, the perturbed solutions will approach to one of the global solutions of $(\mathcal{P}_{sqp})$. By the projection theorem, the nearest points to $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{x}}^*$ in the subspace spanned by $\{\boldsymbol{U}_1, \ldots, \boldsymbol{U}_k\}$ are $\sum_{i=1}^{k}(\bar{\boldsymbol{x}}^T\boldsymbol{U}_i)\boldsymbol{U}_i$ and $\sum_{i=1}^{k}(\bar{\boldsymbol{x}}^{*T}\boldsymbol{U}_i)\boldsymbol{U}_i$, respectively. Then we have the following relation

$$\|\bar{\boldsymbol{x}}^* - \sum_{i=1}^{k}(\bar{\boldsymbol{x}}^{*T}\boldsymbol{U}_i)\boldsymbol{U}_i\|^2 < \|\bar{\boldsymbol{x}} - \sum_{i=1}^{k}(\bar{\boldsymbol{x}}^T\boldsymbol{U}_i)\boldsymbol{U}_i\|^2, \tag{4.21}$$

which means that the perturbed solution $\bar{\boldsymbol{x}}^*$ is closer to the subspace spanned by $\{\boldsymbol{U}_1, \ldots, \boldsymbol{U}_k\}$ than the solution $\bar{\boldsymbol{x}}$.

Furthermore, each solution of the problem $(\mathcal{P}_{sqp})$ can be approximated, if the perturbation parameter $\boldsymbol{\alpha}$ is properly chosen. When the multiplicity of λ_1 is equal to one, as stated in Theorem 16, there are exactly two global solutions. In this case, $\boldsymbol{\alpha}$ becomes a scalar and has exactly two possible directions, which are mutual opposite and respectively lead to the two global solutions (see Example 1). For general cases, there may be infinite number of global solutions for the problem $(\mathcal{P}_{sqp})$, and we will show that there is a one-to-one correspondence between solutions of the problem $(\mathcal{P}_{sqp})$ and directions of $\boldsymbol{\alpha}$. In the problem (4.17), variables $y_i, i = 1, \ldots, k$ are removed by solving the following minimization problem

$$\min\{-2\boldsymbol{\alpha}^T\boldsymbol{y}_k \mid \boldsymbol{y}_k^T\boldsymbol{y}_k = r^2 - \boldsymbol{y}_\ell^T\boldsymbol{y}_\ell,\ \boldsymbol{y}_k \in \mathbb{R}^k\}. \tag{4.22}$$

Its solution is

$$\boldsymbol{y}_k = h\boldsymbol{\alpha}, \quad h = \frac{1}{\|\boldsymbol{\alpha}\|}\sqrt{r^2 - \boldsymbol{y}_\ell^T \boldsymbol{y}_\ell}, \tag{4.23}$$

i.e., the point falls on the boundary of the sphere in (4.22) and has the same direction with $\boldsymbol{\alpha}$. If $\|\boldsymbol{\alpha}\|$ keeps unchanged, the problem (4.17) always has the same solution and the scalar h also keeps unchanged. Thus, each direction of $\boldsymbol{\alpha}$ is corresponding to a solution $\{y_i\}_{i=1}^k$, and all the solutions comprise the surface of a sphere centered at the original in $\mathbb{R}^k$. On the other hand, from the problem (4.16), we have $\bar{\boldsymbol{y}}_k^T \bar{\boldsymbol{y}}_k = r^2 - \bar{\boldsymbol{y}}_\ell^T \bar{\boldsymbol{y}}_\ell$, which means all global solutions of the problem $(\mathcal{P}_{sqp})$ also comprise the surface of a sphere. Combining Theorem 20, we then conclude that each solution of the problem $(\mathcal{P}_{sqp})$ can be approached as the direction of $\boldsymbol{\alpha}$ is properly chosen and $\|\boldsymbol{\alpha}\|$ approaches zero.

4.4 Canonical primal-dual algorithm

Based on the results obtained above, a *canonical primal-dual algorithm* is developed, which is matrix inverse free and the essential cost of calculation is only the matrix-vector multiplication.

The main step of this algorithm is to solve the following perturbed canonical dual problem:

$$(\mathcal{P}_\alpha^d) \qquad \max\left\{\Pi_\alpha^d(\sigma) = -\boldsymbol{p}^T \boldsymbol{G}(\sigma)^{-1}\boldsymbol{p} - r^2\sigma \mid \sigma \in \mathcal{S}_c^+\right\} \tag{4.24}$$

Let $\psi(\sigma)$ be its first-order derivative, i.e.,

$$\psi(\sigma) = (\Pi_\alpha^d(\sigma))' = \boldsymbol{p}^T \boldsymbol{G}(\sigma)^{-1}\boldsymbol{G}(\sigma)^{-1}\boldsymbol{p} - r^2.$$

Then the critical point of $\Pi_\alpha^d(\sigma)$ in $\mathcal{S}_c^+$ is corresponding to the solution of the equation $\psi(\sigma) = 0$ in $\mathcal{S}_c^+$. The first- and second-order derivatives of $\psi(\sigma)$ are

$$\begin{aligned}
\psi'(\sigma) &= -2\boldsymbol{p}^T \boldsymbol{G}(\sigma)^{-1}\boldsymbol{G}(\sigma)^{-1}\boldsymbol{G}(\sigma)^{-1}\boldsymbol{p}, \\
\psi''(\sigma) &= 6\boldsymbol{p}^T \boldsymbol{G}(\sigma)^{-1}\boldsymbol{G}(\sigma)^{-1}\boldsymbol{G}(\sigma)^{-1}\boldsymbol{G}(\sigma)^{-1}\boldsymbol{p}.
\end{aligned}$$

It is noticed that $\psi(\sigma)$ is strictly decreasing and strictly convex over $\mathcal{S}_c^+$, $\psi(\sigma)$ will approach $-r^2$ as σ approaches infinity and $\sigma = -\lambda_1$ is a pole of $\psi(\sigma)$.

We use the Lanczos method to compute an approximation for the smallest eigenvalue of $\boldsymbol{Q}$ and a corresponding eigenvector, denoted respectively by $\tilde{\lambda}_1$ and $\tilde{\boldsymbol{U}}_1$, where the latter is a unit vector. For choosing an effective perturbation, it is not necessary to calculate all eigenvectors of the smallest eigenvalue, since any one of which will be sufficient to divert the direction of $\boldsymbol{f}$. Here we use $\alpha\tilde{\boldsymbol{U}}_1$ as a perturbation to $\boldsymbol{f}$.

Although the perturbed canonical dual problem $(\mathcal{P}_\alpha^d)$ is strictly concave on $\mathcal{S}_c^+$, its derivative $\psi(\sigma)$ would become ill-conditioned when σ approaches to the pole. Therefore, instead of nonlinear optimization techniques, a bisection method is used

to find the root in $(-\lambda_1, +\infty)$ for $\psi(\sigma)$. Each time, as a dual solution $\sigma > -\lambda_1$ is obtained, the value of $\psi(\sigma)$ is calculated and checked to see whether it is equal to zero. For moderate-size problems, it is not hard to calculate $\boldsymbol{G}(\sigma)^{-1}\boldsymbol{p}$ by computing the inverse or decomposition of $\boldsymbol{G}(\sigma)$, but it is not possible for very large-size problems, especially when the memory is very limited. One alternative approach is to solve the following strictly convex minimization problem,

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \ \boldsymbol{x}^T\boldsymbol{G}(\sigma)\boldsymbol{x} - 2\boldsymbol{p}^T\boldsymbol{x}, \tag{4.25}$$

whose optimal solution is $\boldsymbol{x} = \boldsymbol{G}(\sigma)^{-1}\boldsymbol{p}$. Actually, during iterations, we do not need to calculate $\psi(\sigma)$ every time, especially when σ is on the left side of the root and close to the pole. It is discovered that for a given σ, the value of $\psi(\sigma)$ is equal to the optimal value of the following unconstrained concave maximization problem

$$\max_{\boldsymbol{z}\in\mathbb{R}^n} \ -\boldsymbol{z}^T\boldsymbol{G}(\sigma)\boldsymbol{G}(\sigma)\boldsymbol{z} + 2\boldsymbol{p}^T\boldsymbol{z} - r^2. \tag{4.26}$$

By the fact that the value of the objective function will increase during the iterations, we can stop solving the problem (4.26) if the target function is larger than a threshold, and then we claim that σ must be on the left side of the root. Thus, the ill-condition in computing $\psi(\sigma)$ can be prevented as σ approaches to the pole. Since the optimal value is equal to zero when σ is a root of $\psi(\sigma)$, any nonnegative value can be a threshold.

An uncertainty interval should be initialized before the bisection method is applied, and it is used to safeguard that the root is always in intervals of the bisection method. For the right end of the interval, any large enough number can be a candidate. An upper bound can be calculated and then be chosen to be the right end of the uncertainty interval. Let $\bar{\sigma}^* \in (-\lambda_1, +\infty)$ be the root of $\psi(\sigma)$. From the definition of $\psi(\sigma)$, we have

$$\frac{1}{(\lambda_1 + \bar{\sigma}^*)^2}\hat{\boldsymbol{p}}^T\hat{\boldsymbol{p}} - r^2 \geq 0.$$

Hence, $\sqrt{\hat{\boldsymbol{p}}^T\hat{\boldsymbol{p}}}/r = \|\boldsymbol{p}\|/r$ is an upper bound for the root $\bar{\sigma}^*$. However, the bound $\|\boldsymbol{p}\|/r$ may be not tight. A practical way is to let $\sigma = -\lambda_1$ as a starting point and then to update σ recursively by moving a certain step to its right each step. If the first σ that makes the value of $\psi(\sigma)$ be negative is smaller than the upper bound $\|\boldsymbol{p}\|/r$, it is a tighter right end for the uncertainty interval.

Algorithm 1 (Initialization)

Input: Coefficients $\boldsymbol{Q}$, $\boldsymbol{f}$ and r, and an error tolerance ε.

The smallest eigenvalue: Use Lanczos method to obtain $\tilde{\lambda}_1$ and $\tilde{\boldsymbol{U}}_1$.

Perturbation: If existence conditions do not hold, a perturbation is introduced and let

$$\boldsymbol{p} = \boldsymbol{f} + \alpha \tilde{\boldsymbol{U}}_1;$$

otherwise, let $\boldsymbol{p} = \boldsymbol{f}$.

Uncertainty interval: set a step size s_t and a threshold ε_t; let $\sigma = \sigma_\ell = -\tilde{\lambda}_1$.

step 1: Solve the problem (4.26). If the value of the target function is larger than the threshold ε_t, stop the iteration, let $\sigma = \sigma + s_t$ and go to step 1; otherwise, go to step 2.

step 2: Calculate the value of $\psi(\sigma)$. If $\psi(\sigma) > 0$, set $\sigma_\ell = \sigma$, $\sigma = \sigma + s_t$ and go to step 2; otherwise, let $\sigma_u = \sigma$ and stop.

As the uncertainty interval $[\sigma_\ell, \sigma_u]$ is obtained, the bisection method is applied to find the next iterate for σ, by setting σ be the middle point of the uncertainty interval. The main part of the algorithm is given as follows:

Algorithm 2 (Main)

Do

set $\sigma = (\sigma_\ell + \sigma_u)/2$ and calculate the value of $\psi(\sigma)$;

If $|\psi(\sigma)| < \varepsilon$, then STOP and return σ and $\boldsymbol{x}$;

Else if $\psi(\sigma) > 0$, update $\sigma_\ell = \sigma$;

Else update $\sigma_u = \sigma$;

End if

End do

4.5 Numerical experiments

First, three small-size examples are used to illustrate the application of the canonical duality theory. Then, randomly generated examples are presented to demonstrate the efficiency of our method.

4.5.1 Small-size examples

Example 1

The given coefficients are

$$\boldsymbol{Q} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \boldsymbol{f} = \begin{pmatrix} 0 \\ -1.8 \end{pmatrix}, \text{ and } r = 1.$$

The existence conditions do not hold true for this example. There are two global solutions, $\bar{\boldsymbol{x}}_1 = (0.437, -0.9)^T$ and $\bar{\boldsymbol{x}}_2 = (-0.437, -0.9)^T$, which are red points shown in Figure 4.2. In order to show how the perturbation method works, a big perturbation is firstly introduced to the linear coefficient $\boldsymbol{f}$ and let

$$\boldsymbol{p} = (0.5, -1.8)^T.$$

A critical point appears in the interior of $\mathcal{S}_c^+$, which is $\bar{\sigma} = 1.676$ (see Figure 4.2(b)). The corresponding optimal solution for the perturbed problem is $\bar{\boldsymbol{x}}_1^* = (0.74, -0.673)^T$, which is shown as a green point in Figure 4.2(a). As the perturbation becomes smaller, the solution of the perturbed problem should approach to that of the original problem. We then let

$$\boldsymbol{p} = (0.01, -1.8)^T.$$

The critial point now is $\bar{\sigma} = 1.022$ and the corresponding solution is $\bar{\boldsymbol{x}}_1^* = (0.456, -0.89)^T$ (see Figure 4.2(d) and 4.2(c)).

As pointed out above, the other global solution, $\bar{\boldsymbol{x}}_2$, can also be approximated by just choosing a perturbation with the opposite direction. Let $\boldsymbol{p} = (-0.5, -1.8)^T$ and $\boldsymbol{p} = (-0.01, -1.8)^T$. The critical point will be the same as that for $\bar{\boldsymbol{x}}_1^*$, $\bar{\sigma} = 1.676$ and $\bar{\sigma} = 1.022$, and their corresponding primal solutions are $\bar{\boldsymbol{x}}_2^* = (-0.74, -0.673)^T$ and $\bar{\boldsymbol{x}}_2^* = (-0.456, -0.89)^T$.

In Figure 4.2(b), we can see that there is no critical point between $-\lambda_2 = -1$ and $-\lambda_1 = 1$, which suggests that there will no local-nonglobal solution. While there is a critical point between $-\lambda_2 = -1$ and $-\lambda_1 = 1$ in Figure 4.2(d), by Theorem 18 there must be a local-nonglobal solution and it should locate near one of the global solutions, depending on the perturbation.

Example 2

The matrix $\boldsymbol{Q}$ and radius r are the same as that in Example 1 and $\boldsymbol{f}$ is changed to

$$\boldsymbol{f} = \begin{pmatrix} 0 \\ -3 \end{pmatrix},$$

which is in the same direction of that in Example 1 but has a larger length. We notice that though $\sum_{i=1}^{k} \hat{f}_i^2 \neq 0$ is violated, the condition $\sum_{i=k+1}^{n} \frac{\hat{f}_i^2}{(\lambda_i - \lambda_1)^2} > r^2$ holds true. Thus, the problem is not in the hard case. There is a critical point in the interior of $\mathcal{S}_c^+$, which is shown in Figure 4.3(b), and it is corresponding to the unique global solution of the primal problem, which is the green point in Figure 4.3(a).

Example 3

We consider a 4-dimensional problem with $\boldsymbol{Q}$, $\boldsymbol{f}$ and r being

$$\boldsymbol{Q} = \begin{pmatrix} -10 & 0 & 2 & -2 \\ 0 & -3 & -4 & 2 \\ 2 & -4 & 7 & -4 \\ -2 & 2 & -4 & 1 \end{pmatrix}, \quad \boldsymbol{f} = \begin{pmatrix} -10 \\ 6 \\ 10 \\ 9 \end{pmatrix}, \text{ and } r = 5.$$

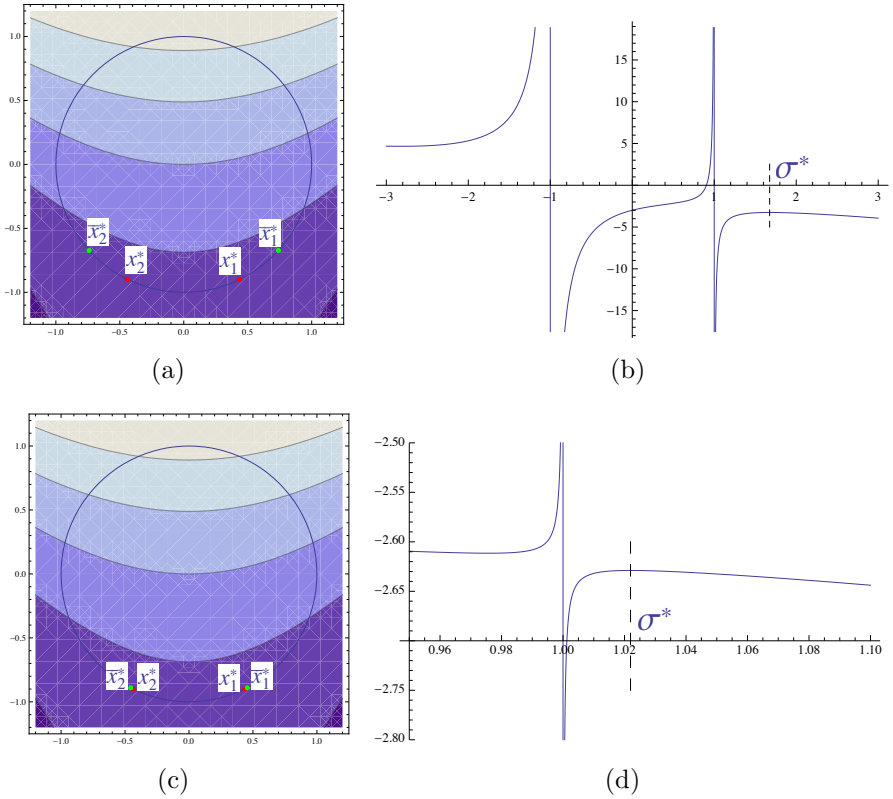

Figure 4.2: Example 1: (a) and (c) are contours of the primal function and the boundary of the sphere; (b) and (d) are the graphs of the dual function.

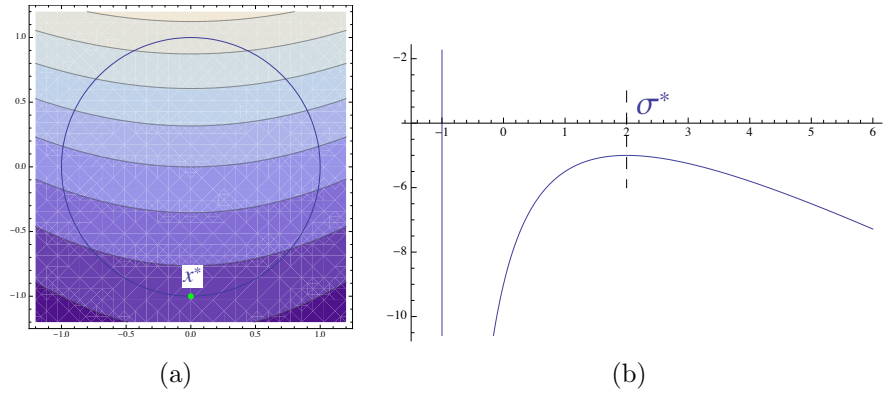

Figure 4.3: Example 2: (a) is the contour of the primal function and boundary of the sphere; (b) is the graph of the dual function.

	Succ.Solv.		Dist.Boun.		Numb.Iter.		Runn.Time.	
Dim	LD	QP	LD	QP	LD	QP	LD	QP
500	10	10	4.716e-09	5.245e-09	28.9	28.6	0.53	1.29
1000	10	10	4.261e-09	3.974e-09	27.1	27.5	1.67	6.25
2000	10	10	3.211e-09	3.822e-09	28.2	27.8	6.52	15.23
3000	10	10	5.674e-09	5.221e-09	26.1	26.4	20.90	72.43
5000	10	10	5.422e-09	3.873e-09	28.6	28.5	71.68	170.34

Table 4.1: General case and $\alpha = 1e-3$.

	Succ.Solv.		Dist.Boun.		Numb.Iter.		Runn.Time.	
Dim	LD	QP	LD	QP	LD	QP	LD	QP
500	10	10	4.532e-09	4.464e-09	28.9	28.9	0.43	1.16
1000	10	10	3.849e-09	5.931e-09	27.4	27.1	1.47	6.08
2000	10	10	2.648e-09	2.872e-09	27.9	28.5	6.26	15.82
3000	10	10	5.299e-09	5.137e-09	26.2	26.2	20.15	73.60
5000	10	10	3.188e-09	4.005e-09	28.7	28.5	65.71	171.92

Table 4.2: General case and $\alpha = 1e-4$.

As shown in Figure 4.1, the canonical dual function $\Pi^d(\sigma)$ has six critical points

$$\bar{\sigma}_6 = -11.1 < \bar{\sigma}_5 = -10.49 < \bar{\sigma}_4 = -1.84 < \bar{\sigma}_3 = 6.08 < \bar{\sigma}_2 = 8.23 < \bar{\sigma}_1 = 12.58.$$

It can be verified that $\bar{\sigma}_1$ belongs to $\mathcal{S}_c^+$, i.e., $\boldsymbol{G}(\bar{\sigma}_1) \succ 0$, which can also be observed from Figure 4.1 where all the vertical lines represent eigenvalues of matrix $\boldsymbol{Q}$. Thus the corresponding solution

$$\bar{\boldsymbol{x}}_1 = (-4.71, 1.11, 1.25, 0.18)^T$$

is the global solution of the primal problem. While $\bar{\sigma}_2 = 8.23$ is a local minimizer of $\Pi^d(\sigma)$ in $(-\lambda_2, -\lambda_1)$ and thus the corresponding solution

$$\bar{\boldsymbol{x}}_2 = (4.33, 1.05, 0.91, 2.08)^T$$

is the local-nonglobal minimizer.

4.5.2 Large-size examples

Examples with dimensions of 500, 1000, 2000, 3000 and 5000 are randomly generated, including both general and hard cases. For each given dimension, both cases are tested by ten examples, respectively. Thus, there are totally one hundred examples. All elements of the coefficients, $\boldsymbol{Q}$, $\boldsymbol{f}$ and r, are integer numbers in $[-100, 100]$. For each example of the hard case, in order to make $\boldsymbol{f}$ be easily chosen, we use a matrix

	Succ.Solv.		Dist.Boun.		Numb.Iter.		Runn.Time.	
Dim	LD	QP	LD	QP	LD	QP	LD	QP
500	10	10	4.340e-09	6.297e-09	36.0	34.9	0.48	1.11
1000	10	10	4.253e-09	4.904e-09	34.6	34.9	1.54	3.54
2000	10	10	2.808e-09	4.255e-09	35.9	35.8	7.15	15.11
3000	9	10	5.479e-09	4.466e-09	34.0	35.0	19.41	36.01
5000	10	10	3.755e-09	4.705e-09	35.2	35.5	74.79	121.41

Table 4.3: Hard case and $\alpha = 1e-3$.

	Succ.Solv.		Dist.Boun.		Numb.Iter.		Runn.Time.	
Dim	LD	QP	LD	QP	LD	QP	LD	QP
500	7	9	2.503e-09	4.488e-09	39.6	40.6	0.51	1.36
1000	9	9	3.148e-09	4.482e-09	37.4	38.3	1.56	3.81
2000	5	9	8.668e-09	5.785e-09	38.6	42.6	7.36	17.95
3000	5	10	6.003e-09	3.997e-09	38.4	40.6	20.43	41.06
5000	8	10	4.748e-09	2.814e-09	37.8	38.8	72.72	131.51

Table 4.4: Hard case and $\alpha = 1e-4$.

$\boldsymbol{Q}$ of whom the multiplicity of the smallest eigenvalue is equal to one. The vector $\boldsymbol{f}$ is constructed such that it is perpendicular to the eigenvector of the smallest eigenvalue, and then a proper radius r is selected such that the existence conditions are violated.

Two approaches are used to calculate the value of $\psi(\sigma)$, one using decomposition methods to calculate $\boldsymbol{G}(\sigma)^{-1}\boldsymbol{p}$, for which we use the "left division" in Matlab, and the other solving the problem (4.25), for which we use the function "quadprog" in Matlab. The tolerance parameter "TolFun" of "quadprog" is set to 1e-12. The Lanczos method is implemented by the function "eigs" of Matlab. The Matlab is of version 7.13 and runned in the platform with Linux 64-bit system and quad CPUs.

The step size s_t, the threshold ε_t and the termination tolerance ε are set to $\|\boldsymbol{p}\|/(200r)$, 0 and 1e-8, respectively. For the hard case, a perturbation $\alpha\boldsymbol{U}_1$ is added to the vector $\boldsymbol{f}$, and two values of α, 1e-3 and 1e-4, are tried.

Results are shown in Table 4.1, 4.2, 4.3 and 4.4, and they contain the number of examples which are successfully solved (Succ.Solv.), the distance of the optimal solution to the boundary of the sphere (Dist.Boun.), the number of iterations in Algorithm 2 (Main) (Numb.Iter.) and the running time (in second) of the algorithm (Runn.Time). The values in the columns of Dist.Boun., Numb.Iter. and Runn.Time are averages of the examples successfully solved. We compare the results of the algorithm adopting "left division" and that of the algorithm adopting "quadprog" in the same table, where LD denotes "left division" and QP denotes "quadprog".

We can see that the examples are solved very accurately with error allowance

being less than 1e-09. The failure in solving some examples is due to "left division" and "quadprog" being unable to handle very nearly singular matrices. For general cases, all the examples can be solved within no more than 30 iterations, while for hard cases, the number of iterations is around 40. From the running time, we notice that our method is capable of handling very large problems in reasonable time. The algorithms using "left division" and "quadprog" have similar performances in the accuracy and the number of iterations, whereas the one using "left division" needs much less time than that of the one using "quadprog". However, the one using "quadprog" is able to solve more examples successfully.

Chapter 5

Unconstrained Binary Quadratic Optimization

5.1 Introduction

In this chapter, the unconstrained binary quadratic optimization problem is discussed. The first application of the canonical duality to this problem appears in [45], and then the problem is revisited in [32, 51]. The existence and uniqueness conditions are then studied in [47], which helps in understanding the hardness and constructing unified solution methods.

5.1.1 formulations

The unconstrained binary quadratic optimization problem discussed here is defined as

$$(\mathcal{P}_{bqp}) \qquad \min_{\boldsymbol{x}} \ \Pi(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{f} \qquad \text{s.t.} \ \ \boldsymbol{x} \in \{-1,+1\}^n \tag{5.1}$$

where Q is a symmetric matrix in $\mathbb{R}^{n\times n}$, and $\boldsymbol{f}$ is a vector in $\mathbb{R}^n$. In papers, the unconstrained binary quadratic problem is also referred to 0-1 quadratic optimization problem,

$$\min_{\boldsymbol{x}} \ \frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{f} \qquad \text{s.t.} \ \ \boldsymbol{x} \in \{0,1\}^n \tag{5.2}$$

which can be easily converted into the problem $(\mathcal{P}_{bqp})$.

Because of properties $x_i^2 = 1$ for $x_i \in \{-1, 1\}$ and $x_i^2 - x_i = 0$ for $x_i \in \{0, 1\}$, we can make the diagonal entries of Q in $(\mathcal{P}_{bqp})$ and (5.2) be any values, without changing the shape of the objective function and hence the solutions. If the diagonal entries of Q are zeros, the objective function belongs to the group of so called pseudo-Boolean

functions and the problem (5.2) becomes a pseudo-Boolean optimization problem [19].

The objective functions in ($\mathcal{P}_{bqp}$) and (5.2) can also be written in homogeneous forms. For the problem (5.2), the linear item can be rewritten in the quadratic form, $\boldsymbol{x}^T\boldsymbol{f} = \boldsymbol{x}^T\mathrm{diag}(\boldsymbol{f})\boldsymbol{x}$, because of the property $x_i^2 = x_i$. For the problem ($\mathcal{P}_{bqp}$), an extra binary variable should be introduced, by which the linear item in $\Pi(\boldsymbol{x})$ is multiplied. The equivalence is obtained from the fact that the new variable will always be one for optimal solutions. Conversely, any homogeneous quadratic function can be converted into an equivalent one with a nonzero linear term. It is observed that $\frac{1}{2}\boldsymbol{x}^TQ\boldsymbol{x} = \frac{1}{2}(-\boldsymbol{x})^TQ(-\boldsymbol{x})$. Thus we can just fix the value of an arbitrary component of $\boldsymbol{x}$ and immediately get a nonhomogeneous quadratic function.

In principle, any linear or quadratic problem with linear constraints and bounded integer variables can be reformulated as a binary quadratic problem [50, 77]. For bounded integer constraints, for example $x_i \in \{u_{i1}, \ldots, u_{ik_i}\}$, variables $y_{ik} \in \{0, 1\}$ can be introduced and then x_i is replaced by $\sum_{j=1}^{k_i} u_{ij}y_{ij}$. For linear equality constraints, they can be removed by adding a quadratic infeasibility penalty function into the objective function, while, for linear inequality constraints, slack variables can be introduced, which have finite possible values (may not integer numbers) and are then transformed into binary variables. For some simple linear constraints, appropriate quadratic penalties are known in advance and can be used straight away [77].

5.1.2 Combinatorial problems and complexity

Many practical combinatorial optimization problems can be transformed into unconstrained binary quadratic optimization problems, such as problems of determining maximum cuts, maximum cliques, maximum vertex packing, minimum coverings, maximum independent sets and maximum independent weighted sets, and max 2-SAT problems [95, 97, 17, 78, 103, 79]. Here, we particularly present the relations with max-cut and maximum clique problems.

Max-cut problems

Let $G = (V, E)$ be an undirected graph with vertexes $V = \{1, 2, \ldots, n\}$ and edges $E = \{(i, j) \mid i, j \in V, i \neq j\}$. For each edge $(i, j) \in E$, there is a weight w_{ij} attached. A cut of graph G is defined as an edge set $E(S, T) = \{(i, j) \in E \mid i \in S, j \in T\}$, where (S, T) is a bipartition of the vertex set V, i.e., $S \cap T = \emptyset$ and $V = S \cup T$. The total weight of the cut is defined by

$$w(S, T) = \sum_{(i,j)\in E(S,T)} w_{ij}.$$

The max-cut problem is defined as finding a bipartition (S, T) such that $w(S, T)$ is maximized.

If we let $x_i = 1$ for $i \in S$ and $x_i = 0$ for $i \in T$, the total weight $w(S,T)$ can also be expressed as

$$w(S,T) = \sum_{(i,j)\in E} w_{ij} x_i (1 - x_j).$$

On the other hand, for any vector $\boldsymbol{x} \in \{0,1\}^n$, a cut can correspondingly be defined as $S = \{i \in V \mid x_i = 1\}$ and $T = \{i \in V \mid x_i = 0\}$. Thus there is a one-to-one correspondence between cuts of G and vectors in $\{0,1\}^n$. Let $Q = \{q_{ij}\}$, where $q_{ij} = -w_{ij}$ if $(i,j) \in E$ and otherwise $q_{ij} = 0$. Then the max-cut problem can be equivalently formulated as a 0-1 quadratic problem

$$\begin{aligned} \min \ & \boldsymbol{x}^T Q \boldsymbol{x} - \boldsymbol{x}^T Q \boldsymbol{e} \\ \text{s.t.} \ & \boldsymbol{x} \in \{0,1\}^n \end{aligned}$$

Reversely, consider the problem (5.2). Without loss of generality, we assume that all diagonal entries of matrix Q are zeros. Let

$$A = \{a_{ij}\} = \begin{bmatrix} 0 & \boldsymbol{f}^T - \frac{1}{2}\boldsymbol{e}^T Q \\ \boldsymbol{f} - \frac{1}{2} Q \boldsymbol{e} & \frac{1}{2} Q \end{bmatrix}.$$

Then a graph $G = (V,E)$, associated with the matrix A, can be constructed as follows: $V = \{0,1,2,\ldots,n\}$, $E = \{(i,j) \mid a_{ij} \neq 0\}$ and weight $w_{ij} = -a_{ij}$ for $(i,j) \in E$. If we fix vertex 0 in T for any cut (S,T) and let $\bar{\boldsymbol{x}} = (0, \boldsymbol{x})$ being the corresponding vector in $\{0,1\}^n$, the total weight for a cut will be

$$\bar{\boldsymbol{x}}^T A (\bar{\boldsymbol{x}} - \bar{\boldsymbol{e}}) = -\frac{1}{2} \boldsymbol{x}^T Q \boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{f},$$

where $\bar{\boldsymbol{e}}$ is an all-one vector. Thus minimizing a quadratic function over binary constraints becomes solving a max-cut problem.

Maximum clique problems

Let $G = (V,E)$ be an undirected graph with vertexes $V = \{1,2,\ldots,n\}$ and edges $E = \{(i,j) \mid i,j \in V, i \neq j\}$. For each vertex $i \in V$, a weight $w_i > 0$ is associated with it. Given a subset $S \subset V$, the subgraph induced by S is defined as $G(S) = (S, E(S))$, where $E(S) = \{(i,j) \mid (i,j) \in E, i,j \in S\}$. A clique C is a subset of V such that the subgraph $G(C) = (S, E(C))$ is complete, i.e., all vertexes of the subgraph $G(C)$ are pairwise adjacent. The maximum clique problem is to find a clique with maximal weight.

The maximum clique problem has many formulations [97, 17]. Let $\bar{G} = (V, \bar{E})$ be the complement graph of G, where $\bar{E} = \{(i,j) \mid i,j \in V, i \neq j, (i,j) \notin E\}$, and $A_{\bar{G}} = \{a_{ij}\}$ be the adjacency matrix of $\bar{G}$, i.e., $a_{ij} = 1$ if $(i,j) \in \bar{E}$ and $a_{ij} = 0$ if $(i,j) \notin \bar{E}$. The simplest formulation for the maximum clique problem is the following edge formulation:

$$\begin{aligned} \max \ & \boldsymbol{w}^T \boldsymbol{x} \\ \text{s.t.} \ & x_i + x_j \leq 1, \ \forall (i,j) \in \bar{E} \\ & \boldsymbol{x} \in \{0,1\}^n \end{aligned}$$

As mentioned before, because of the properties $x_i^2 - x_i = 0$, constraints $x_i + x_j \leq 1$ are equivalent to $x_i x_j = 0$, which can be removed by using a penalty function. Then the above formulation can be equivalently rewritten as a 0-1 quadratic problem

$$\begin{aligned} \min \quad & -\boldsymbol{w}^T \boldsymbol{x} + \alpha \sum_{(i,j)\in \bar{E}} x_i x_j = \alpha \boldsymbol{x}^T A_{\bar{G}} \boldsymbol{x} - \boldsymbol{w}^T \boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{x} \in \{0,1\}^n \end{aligned}$$

where α is a positive constant. Here, we notice that it is not necessary to make the parameter α approach to infinity, and any large enough positive number could guarantee the equivalent relation between the two problems.

Complexity

Both max-cut and maximum clique problems are NP-complete [74]. Thus, the problem QP is NP-complete. Moreover, the hardness is further investigated in [96] and it is proved that checking whether the problem has a unique solution, finding the unique solution and even just finding a discrete local minimum are all NP-hard.

There are some cases which can be solved by algorithms with polynomial time bounds. If Q is a diagonal matrix, then the objective function is separable and the problem can be solved analytically. If the matrix Q is of rank one, it is polynomial solvable since the solution can be found by inspection. For a matrix Q with bounded rank, the problem is also polynomial solvable, if off-diagonal entries of Q are nonpositive [14] or Q is positive semidefinite [4]. By reducing the binary quadratic problem to a max-cut problem in an associated graph, many other polynomial solvable cases are derived, including planar graphs [64, 95], graphs without K_5 minors [8], series-parallel graphs [11] and weakly bipartite graphs [62].

5.1.3 Algorithms

Numerous methods have been developed for solving the binary quadratic problem or the associated combinatorial problems. The exact methods are categorized into five groups: branch-and-bound methods, cutting plane methods, branch-and-cut methods, algebraic methods and continuous methods.

Branch and bound

The branch-and-bound method is based on the idea of implicitly and intelligently enumerating all the feasible solutions. Given n being the dimension of the binary quadratic problem, there are 2^n feasible solutions. It is hopeless to examine each solution to find the global solution even when n is moderate. The branch-and-bound method breaks the problem into a series of smaller problems that can be easily tackled, and then puts the information together again to obtain an optimal solution for the original problem. The construction of smaller problems is based on

successive partitioning of the solution set. The branch in branch-and-bound refers to this partitioning process; the bound refers to lower and upper bounds that are used to provide a proof of optimality.

There are many strategies in how a branch-and-bound algorithm is implemented. At each branching step of determining which node to branch from, the usual alternatives are least-lower-bound-next, depth-first and breadth-first. At each node, the upper bound for the corresponding subproblem can be provided by any feasible solutions, and the lower bound is normally obtained by relaxation, duality or some other methods, which will be surveyed below.

A variety of branch-and-bound methods have been introduced in literatures for the binary quadratic problem. They are equipped with different strategies of branching and lower bounding. The bounding techniques used in branch-and-bound methods are convex quadratic program relaxations [22, 16], linearization techniques [18, 63, 3], linear programming relaxations [9, 10], semidefinite programming relaxations [68, 67, 31, 60, 101], Second order cone programming relaxations [76, 91], roof duality [66, 18, 19], DC programming [125], one row relaxations [126, 127] and geometric property [84].

Cutting plane

For mixed-integer linear programming, the cutting plane methods work by first solving the linear relaxation, which is obtained by replacing the integer constraints with box constraints. The theory of the linear programming guarantees that under mild assumptions, one can always find an extreme point or a corner point that is an optimal solution. Then the obtained optimal solution is examined: if it satisfies the integer constraints, an optimal solution for the original problem is already found; if it not, a linear inequality can be constructed that separates the optimal solution from the convex hull of the feasible region of the original problem. Such an inequality is called a cut, which can then be added to the relaxation problem to cut off the non-integer optimal solution and tighten the feasible region. This process is repeated until an optimal integer solution is found. The two most prominent cuts for the mixed-integer linear programming are Gomory cut [6] and lift-and-project cut [5].

In [9, 132], cutting plane method are used to solve the binary quadratic problem, which is firstly transformed into a mixed-integer linear programming problem by linearization techniques (see [63] and references there). The most used linearization technique is to replace $x_i x_j$ with a new variable z_{ij}. Here, the diagonal entries of matrix Q are supposed to be zeros. Then the following problem is obtained:

$$\begin{aligned} \min_{\boldsymbol{x},\boldsymbol{z}} \quad & \sum_{i<j} q_{ij} z_{ij} - \sum_{i=1}^{n} f_i x_i \\ \text{s.t.} \quad & z_{ij} \le x_i,\ z_{ij} \le x_j,\ x_i + x_j - 1 \le z_{ij} \\ & z_{ij} \ge 0, 1 \le i < j \le n \\ & \boldsymbol{x} \in \{0,1\}^n \end{aligned} \tag{5.3}$$

In fact, z_{ij} can also be stated as 0-1 integer variables, since under other constraints they can only be zero or one. Thus, the problem (5.3) becomes a constrained 0-1 linear programming problem.

The convex hull of the feasible region of the problem (5.3) is called the Boolean quadric polytope [92], and three families of facets for this polytope is identified: the clique-inequality, the cut-inequality and the generalized cut inequality, which can be used to construct sufficient cuts.

Branch and cut

A combination of cutting planes and branch-and-bound search generates the so called branch-and-cut methods. At each node on the search tree, cutting planes are added to tighten the linear relaxations, and thus the lower bound is improved. Work in [93, 6, 126, 127] gives more details on the implementation of branch-and-cut methods.

Algebraic methods

In [30, 19], the binary quadratic problem is transformed into pseudo-Boolean optimization problems, where pseudo-Boolean functions are minimized over binary constraints $\boldsymbol{x} \in \{0,1\}^n$. A pseudo-boolean function is a real-valued function of 0-1 variables. Any pseudo-Boolean function can be written uniquely as a multi-linear polynomial:

$$g(\boldsymbol{x}) = a + \sum_i a_i x_i + \sum_{i<j} a_{ij} x_i x_j + \sum_{i<j<k} a_{ijk} x_i x_j x_k + \cdots .$$

Because of the property $x_i^2 = x_i$, the function $\Pi(\boldsymbol{x})$ can be written as a pseudo-Boolean function. The basic algorithm for pseudo-Boolean optimization determines the minimum of the pseudo-Boolean function by recursively eliminating variables, following the dynamic programming principle. However, computationally, the procedure could be very expensive. Thus, a branch-and-bound scheme is proposed to the variable elimination [30].

Continuous methods

In [94], a continuous approach is described for solving the problem ($\mathcal{P}_{bqp}$). Rather than using relaxations and bounding information in a tree search scheme, the authors employ Fischer-Burmeister nonlinear complementarity function to reformulate the problem as a continuous problem with equilibrium constraints. The binary constraint $x_i \in \{-1, 1\}$ is always equivalent to conditions:

$$-1 \le x_i \le 1, \quad (1 + x_i)(1 - x_i) = 0,$$

where, by employing the Fischer-Burmeister function, the latter complementarity condition is equivalent to the following equality

$$\phi_{FB}(1 + x_i, 1 - x_i) = \sqrt{2 + 2x_i^2} - 2 = 0.$$

Then, the quadratic penalty function and logarithmic barrier function are used to remove the equality and inequality constraints, and a global smoothing function is constructed, which promises convexity in a large subset of its domain.

5.2 Lagrangian relaxations

The feasible region of the problem (5.1) can be written as

$$\mathcal{X} = \{\boldsymbol{x} \mid x_i^2 = 1, i = 1, \ldots, n\},$$

which is a canonical transformation [45, 46, 55, 117, 52, 48]. Then, the Lagrangian is defined as

$$\begin{aligned} L(\boldsymbol{x}, \boldsymbol{\mu}) &= \frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x} + \boldsymbol{f}^T\boldsymbol{x} + \frac{1}{2}\sum_{i=1}^{n} \mu_i(x_i^2 - 1) \\ &= \frac{1}{2}\boldsymbol{x}^T(Q + \mathrm{diag}(\boldsymbol{\mu}))\boldsymbol{x} + \boldsymbol{f}^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{e}^T\boldsymbol{\mu}. \end{aligned}$$

and the Lagrangian dual problem is

$$\sup \ \{D(\boldsymbol{\mu}) \mid \boldsymbol{\mu} \in \mathrm{dom}h\}, \tag{5.4}$$

where $D(\boldsymbol{\mu})$ is the dual function

$$D(\boldsymbol{\mu}) = \inf\{L(\boldsymbol{x}, \boldsymbol{\mu}) \mid \boldsymbol{x} \in \mathbb{R}^n\}, \tag{5.5}$$

and $\mathrm{dom}D = \{\boldsymbol{\mu} \in \mathbb{R}^n \mid D(\boldsymbol{\mu}) > -\infty\}$.

The weak duality is always true

$$\Pi(\boldsymbol{x}) \geq D(\boldsymbol{\mu}) \ \ \forall \boldsymbol{x} \in \mathcal{X}, \forall \boldsymbol{\mu} \in \mathrm{dom}D.$$

Since the problem is nonconvex, the strong duality may not hold , i.e. $\Pi(\boldsymbol{x}^*) > D(\boldsymbol{\mu}^*)$ with $\boldsymbol{x}^*$ and $\boldsymbol{\mu}^*$ being optimal solutions for (5.1) and (5.4), respectively, and thus there may exists a duality gap. If this happens, the optimal value of the dual problem is only a lower bound for the primal problem, and the dual problem is then called the Lagrangian relaxation for the original problem. Whereas, if there exist $\bar{\boldsymbol{x}} \in \mathcal{X}$ and $\bar{\mu} \in \mathrm{dom}D$ such that $\Pi(\bar{\boldsymbol{x}}) = D(\bar{\mu})$, then $\bar{\boldsymbol{x}}$ and $\bar{\mu}$ must be global solutions of the primal and dual problems.

Before go to discuss the sufficient and necessary conditions for the global solutions, we have a look at what $\boldsymbol{\mu}$ will make the dual function $D(\boldsymbol{\mu}) > -\infty$. We have the following result.

Lemma 21 *Let $G \in \mathbb{S}^n$ and $\boldsymbol{f} \in \mathbb{R}^n$. Then, $\inf\{\frac{1}{2}\boldsymbol{x}^T G\boldsymbol{x} - \boldsymbol{f}^T\boldsymbol{x} : \ \boldsymbol{x} \in \mathbb{R}^n\} > -\infty$ if and only if the following two conditions hold:*

(i) $G \succeq 0$,

(ii) $\exists \boldsymbol{x} \in \mathbb{R}^n$ *such that* $G\boldsymbol{x} = \boldsymbol{f}$.

If $G \succeq 0$ does not hold, there will be a negative eigenvalue of the matrix G, say λ_i. Let $\boldsymbol{x}$ be on the same direction with an eigenvector of λ_i. The function will become $\frac{1}{2}\lambda_i\|\boldsymbol{x}\|^2 - \boldsymbol{f}^T\boldsymbol{x}$, which is unbounded below. So the first condition is necessary. The second condition can be easily understood if we apply the eigendecomposition to the matrix G. Geometrically, the second condition requires that the vector $\boldsymbol{f}$ should be perpendicular to the subspace generated by the eigenvectors of the zero eigenvalue of G, if there is any.

The following theorem presents a global sufficient optimality condition, proposed in [13], for the problem $(\mathcal{P}_{bqp})$.

Theorem 22 *Consider the problem* $(\mathcal{P}_{bqp})$*, and let* λ_1 *be the smallest eigenvalue of* Q*. If* $\boldsymbol{x} \in \mathcal{X}$ *satisfies*

$$\lambda_1 \boldsymbol{e} \geq diag(\boldsymbol{x})Q\boldsymbol{x} - diag(\boldsymbol{x})\boldsymbol{f}, \tag{5.6}$$

then $\boldsymbol{x}$ *is a global optimal solution.*

From the Lemma 21, we have $D(\boldsymbol{\mu}) = \inf\{L(\boldsymbol{x}, \boldsymbol{\mu}) : \ \boldsymbol{x} \in \mathbb{R}^n\} > -\infty$ if and only if there exists $\boldsymbol{x} \in \mathbb{R}^n$ such that

$$(Q + \mathrm{diag}(\boldsymbol{\mu}))\boldsymbol{x} = \boldsymbol{f}, \tag{5.7}$$

$$Q + \mathrm{diag}(\boldsymbol{\mu}) \succeq 0. \tag{5.8}$$

If $\boldsymbol{x} \in \mathcal{X}$, the condition (5.7) is equivalent to

$$\mathrm{diag}(\boldsymbol{x})(Q + \mathrm{diag}(\boldsymbol{\mu}))\boldsymbol{x} - \boldsymbol{f}) = 0,$$

from which we have

$$\boldsymbol{\mu} = -\mathrm{diag}(\boldsymbol{x})Q\boldsymbol{x} + \mathrm{diag}(\boldsymbol{x})\boldsymbol{f}. \tag{5.9}$$

For a vector $\boldsymbol{x} \in \mathcal{X}$, if $\boldsymbol{\mu}$ defined by (5.9) satisfies the condition (5.8), we have

$$\begin{aligned} D(\boldsymbol{\mu}) &= \inf\{\frac{1}{2}\boldsymbol{x}^T(Q + \mathrm{diag}(\boldsymbol{\mu}))\boldsymbol{x} - \boldsymbol{f}^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{e}^T\boldsymbol{\mu} \mid \boldsymbol{x} \in \mathbb{R}^n\} \\ &= -\frac{1}{2}\boldsymbol{f}^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{e}^T\boldsymbol{\mu} \\ &= -\frac{1}{2}\boldsymbol{f}^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{e}^T(-\mathrm{diag}(\boldsymbol{x})Q\boldsymbol{x} + \mathrm{diag}(\boldsymbol{x})\boldsymbol{f}) \\ &= \frac{1}{2}\boldsymbol{x}^TQ\boldsymbol{x} - \boldsymbol{f}^T\boldsymbol{x} = \Pi(\boldsymbol{x}) \end{aligned}$$

From the Lagrangian duality, we know that $\boldsymbol{x}$ is a global optimal solution. Because of $Q + \mathrm{diag}(\boldsymbol{\mu}) = Q + \min\{\mu_i\}I + \mathrm{diag}(\boldsymbol{\mu} - \min\{\mu_i\}\boldsymbol{e})$, the condition (5.8) is guaranteed if $\lambda_1 + \min\{\mu_i\} \geq 0$, which is true when the inequality (5.6) holds. Thus, for a

solution $\boldsymbol{x} \in \mathcal{X}$ satisfying the inequality (5.6), the vector $\boldsymbol{\mu}$ satisfying (5.9) is an optimal solution of the Lagrangian dual problem and there is no duality gap, which implies that $\boldsymbol{x}$ is a global optimal solution.

If let $\boldsymbol{x} = (-x_1^*, x_2^*, \ldots, x_n^*) \in \mathcal{X}$, we have

$$\frac{1}{2}\boldsymbol{x}^{*T}Q\boldsymbol{x}^* - \boldsymbol{f}^T\boldsymbol{x}^* \leq \frac{1}{2}\boldsymbol{x}^{*T}Q\boldsymbol{x}^* + 2q_{11} - 2x_1^*\boldsymbol{e}_1^TQ\boldsymbol{x}^* + 2x_1^*\boldsymbol{f}^T\boldsymbol{e}_1 - \boldsymbol{f}^T\boldsymbol{x}^*,$$

which reduces to

$$x_1^*\boldsymbol{e}_1^TQ\boldsymbol{x}^* + x_1^*\boldsymbol{f}^T\boldsymbol{e}_1 \leq q_{11},$$

where $\boldsymbol{e}_1 = (1, 0, \ldots, 0)$. Similarly, it is shown that for $i = 1, \ldots, n$,

$$x_j^*\boldsymbol{e}_j^TQ\boldsymbol{x}^* + x_j^*\boldsymbol{f}^T\boldsymbol{e}_j \leq q_{jj}.$$

Thus, we have the following global necessary optimality condition, which is proposed by [13].

Theorem 23 *Consider the problem ($\mathcal{P}_{bqp}$). If $\boldsymbol{x}^* \in \mathcal{X}$ is a global optimal solution, then*

$$diag(\boldsymbol{x}^*)Q\boldsymbol{x}^* - diag(\boldsymbol{x}^*)\boldsymbol{f} \leq diag(Q), \tag{5.10}$$

where $diag(Q) = (q_{11}, \ldots, q_{nn})$.

In the rest of this section, we assume that the objective function $\Pi(\boldsymbol{x})$ is homogeneous, and consider

$$p^* = \min\left\{\frac{1}{2}\boldsymbol{x}^TQ\boldsymbol{x} \mid \boldsymbol{x} \in \mathcal{X}\right\}. \tag{5.11}$$

The assumption will not make any loss of generality since any nonhomogeneous formulation can be equivalently transformed to a homogeneous one. Then the Lagrangian dual function can be explicitly written as

$$D(\boldsymbol{\mu}) = \inf_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\boldsymbol{x}^T(Q + \mathrm{diag}(\boldsymbol{\mu}))\boldsymbol{x} - \frac{1}{2}\boldsymbol{e}^T\boldsymbol{\mu} = \begin{cases} -\frac{1}{2}\boldsymbol{e}^T\boldsymbol{\mu} & Q + \mathrm{diag}(\boldsymbol{\mu}) \succeq 0 \\ -\infty & \text{otherwise} \end{cases}$$

and the Lagrangian dual problem becomes

$$\begin{aligned} d^* = \sup_{\boldsymbol{\mu}} \ & -\frac{1}{2}\boldsymbol{e}^T\boldsymbol{\mu} \\ & \text{s.t. } Q + \mathrm{diag}(\boldsymbol{\mu}) \succeq 0. \end{aligned} \tag{5.12}$$

The problem (5.12) is an SDP problem. Moreover, it has a unique optimal solution. The proof of the uniqueness can be found in [86].

Theorem 24 *The SDP problem (5.12) has a unique optimal solution.*

The dual problem of (5.12) is also a convex problem and shares the same optimal value with its primal problem. Formulating the Lagrangian for the problem (5.12)

$$\begin{aligned} L(\boldsymbol{\mu}, Z) &= -\frac{1}{2}\boldsymbol{e}^T\boldsymbol{\mu} + \frac{1}{2}(Q + \mathrm{diag}(\boldsymbol{\mu})) \cdot Z \\ &= \frac{1}{2}(\mathrm{diag}(Z) - \boldsymbol{e})^T\boldsymbol{\mu} + \frac{1}{2}Q \cdot Z, \end{aligned}$$

where $\mathrm{diag}(Z)$ denotes the column vector $\{z_{ii}\}$, the dual function is then defined as

$$D(Z) = \frac{1}{2}Q \cdot Z, \;\; \mathrm{dom}D = \{Z \mid \mathrm{diag}(Z) = \boldsymbol{e}, \; Z \succeq 0\}.$$

The Lagrangian dual problem for (5.12) is formulated as

$$\begin{aligned} \inf_Z \; & \frac{1}{2}Q \cdot Z \\ \text{s.t. } & \mathrm{diag}(Z) = \boldsymbol{e}, \\ & Z \succeq 0. \end{aligned} \tag{5.13}$$

Obviously, the Slater's condition hold for the problem (5.13), i.e., there exists Z such that $Z \succ 0$ and $\mathrm{diag}(Z) = \boldsymbol{e}$. Thus, the solution $\boldsymbol{\mu}$ and Z are optimal solutions for (5.12) and (5.13), respectively, if and only if the KKT conditions hold:

$$\begin{aligned} Q + \mathrm{diag}(\boldsymbol{\mu}) &\succeq 0 \\ Z &\succeq 0 \\ \mathrm{diag}(Z) &= \boldsymbol{e} \\ (Q + \mathrm{diag}(\boldsymbol{\mu})) \cdot Z &= 0 \end{aligned}$$

The sup in (5.12) and inf in (5.13) can then respectively be replaced by max and min.

Moreover, the optimal solution of the problem (5.12) is always on the boundary of the feasible region, as stated in the following result.

Lemma 25 *Let $\boldsymbol{\mu}^*$ be the optimal solution of the problem (5.12) and $\bar{\lambda}_1 \leq \bar{\lambda}_2 \leq \cdots \leq \bar{\lambda}_n$ be eigenvalues of the matrix $Q + diag(\boldsymbol{\mu}^*)$. Then*

$$\bar{\lambda}_1 = 0.$$

If $\bar{\lambda}_1 > 0$ and the solution $\boldsymbol{\mu}^*$ is an interior of the feasible region, we can always move a small distance towards the boundary such that the semidefinite constraint still holds. Then the new point will still satisfy the semidefinite constraint and but possess a larger function value, which contradicts the optimality of $\boldsymbol{\mu}^*$.

SDP relaxation

The SDP relaxation method belongs to the lift-and-project convex relaxation. With the homogeneous formulation, the problem is first lifted to an equivalent problem in the space $\mathbb{S}^n$. Let $X = \boldsymbol{x}\boldsymbol{x}^T$. Because of the equivalent relation

$$X = \boldsymbol{x}\boldsymbol{x}^T \Leftrightarrow X \succeq 0 \text{ and } \operatorname{rank}(X) = 1,$$

the problem $(\mathcal{P}_{bqp})$ with $\boldsymbol{f} = 0$ is equivalent to

$$\begin{aligned} \min_X \ & \frac{1}{2} Q \cdot X \\ \text{s.t. } & \operatorname{diag}(X) = \boldsymbol{e},\ X \succeq 0,\ \operatorname{rank}(X) = 1. \end{aligned}$$

Then, by removing the rank-1 constraint, the problem is relaxed and an SDP problem is obtained:

$$\begin{aligned} \min_X \ & \frac{1}{2} Q \cdot X \\ \text{s.t. } & \operatorname{diag}(X) = \boldsymbol{e},\ X \succeq 0. \end{aligned}$$

After the SDP problem is solved, the solution, which is in $\mathbb{S}^n$, is projected back into $\mathbb{R}^n$, and the projection provides an approximation solution for the original problem. Since the resulted SDP problem is an relaxation, its optimal value provides a lower bound for the original problem. Obviously, the SDP relaxation is equivalent to the Lagrangian relaxation.

The first SDP relaxation for the binary quadratic problem is given by [85], and then interests are poured into investigating its theoretical properties and applications. In [69], an interior-point method is proposed for the SDP. The most prominent result of applications of the SDP is that the SDP relaxation normally provides approximate solutions of very good quality. In [60], the Max Cut and Max 2SAT problems are considered, and the randomized approximation algorithms, which use the SDP relaxation, always deliver solutions of expected value at least 0.87856 times of the optimal value. The SDP is also employed to calculate lower bounds in brand-and-bound algorithms [68, 103].

Convex quadratic programming relaxations

The problem $(\mathcal{P}_{bqp})$ with $\boldsymbol{f} = 0$ can also be equivalently written as

$$\begin{aligned} \min_{\boldsymbol{x}} \ & \frac{1}{2} \boldsymbol{x}^T (Q + \operatorname{diag}(\boldsymbol{\mu})) \boldsymbol{x} - \frac{1}{2} \boldsymbol{e}^T \boldsymbol{\mu} \\ \text{s.t. } & x_i^2 = 1, i = 1, \ldots, n \end{aligned} \tag{5.14}$$

without changing the value of the objective function for each feasible solution.

If choose parameters $\boldsymbol{\mu}$ such that $Q+\text{diag}(\boldsymbol{\mu})$ is positive semidefinite and relax the integer constraints into box constraints, we obtain a convex quadratic programming relaxation

$$\begin{aligned} \min_{\boldsymbol{x}} \ & \frac{1}{2}\boldsymbol{x}^T(Q+\text{diag}(\boldsymbol{\mu}))\boldsymbol{x} - \frac{1}{2}\boldsymbol{e}^T\boldsymbol{\mu} \\ \text{s.t.} \ & x_i^2 \leq 1, i = 1, \ldots, n \end{aligned} \tag{5.15}$$

Obviously, its optimal value, which can be easily observed and is equal to $-\frac{1}{2}\boldsymbol{e}^T\boldsymbol{\mu}$, is a lower bound for the original problem. Naturally, we are interested in finding the best parameter $\boldsymbol{\mu}$, under the constraint of positive semidefiniteness, such that the lower bound is maximized. It results in the exact same problem (5.12).

Another convex quadratic programming relaxation is constructed by relaxing the integer constraint into a sphere constraint $\boldsymbol{x}^T\boldsymbol{x} \leq n$, on whose boundary the integer solutions locate,

$$\begin{aligned} \min_{\boldsymbol{x}} \ & \frac{1}{2}\boldsymbol{x}^T(Q+\text{diag}(\boldsymbol{\mu}))\boldsymbol{x} - \frac{1}{2}\boldsymbol{e}^T\boldsymbol{\mu} \\ \text{s.t.} \ & \boldsymbol{x}^T\boldsymbol{x} \leq n. \end{aligned} \tag{5.16}$$

Similarly, choosing the best parameter $\boldsymbol{\mu}$ is equivalent to solving the problem (5.12).

Improvements on the lower bound

Several improvements on the lower bound obtained from the SDP problem (5.12) have been proposed [86, 14].

As stated in Lemma 25, zeros are the smallest eigenvalues of $Q + \text{diag}(\boldsymbol{\mu}^*)$ with $\boldsymbol{\mu}^*$ being the optimal solution of the problem (5.12). Thus, the matrix $Q + \text{diag}(\boldsymbol{\mu}^*)$ can be decomposed into the following form

$$Q + \text{diag}(\boldsymbol{\mu}^*) = \bar{V}\bar{\Lambda}\bar{V}^T = \bar{V}_+\bar{\Lambda}_+\bar{V}_+^T \tag{5.17}$$

with

$$\bar{V} = [\bar{V}_0 \ \bar{V}_+], \text{ and } \bar{\Lambda} = \begin{bmatrix} 0_k & 0 \\ 0 & \bar{\Lambda}_+ \end{bmatrix}.$$

where k is the multiplicity of the smallest eigenvalue.

Lemma 26 *Let $\boldsymbol{\mu}^*$ be the optimal solution of the problem (5.12) and $Q + \text{diag}(\boldsymbol{\mu}^*)$ have the eigendecomposition in (5.17). Then, the following statements are equivalent*

1. $p^* = d^*$;
2. $\bar{V}_0\boldsymbol{z} \in \{-1, 1\}^n$ *for some* $\boldsymbol{z} \in \mathbb{R}^k$;
3. $n = \max\{\boldsymbol{x}^T\bar{V}_0\bar{V}_0^T\boldsymbol{x} \mid \boldsymbol{x} \in \{-1, 1\}^n\}$.

If there exists $\boldsymbol{x} \in \{-1,1\}^n$ such that $\boldsymbol{x} = \bar{V}_0 \boldsymbol{z}$ for some $\boldsymbol{z} \in \mathbb{R}^k$, we have $\boldsymbol{x}^T \bar{V}_0 \bar{V}_0^T \boldsymbol{x} = \boldsymbol{z}^T \bar{V}_0^T \bar{V}_0 \bar{V}_0^T \bar{V}_0 \boldsymbol{z} = \boldsymbol{z}^T \bar{V}_0^T \bar{V}_0 \boldsymbol{z} = \|\boldsymbol{x}\|^2 = n$. While for any $\boldsymbol{x}\{-1,1\}^n$, it is true that $\boldsymbol{x}^T \bar{V}_0 \bar{V}_0^T \boldsymbol{x} = \|\bar{V}_0^T \boldsymbol{x}\|^2 \leq \|\boldsymbol{x}\|^2 = n$. Thus, $\boldsymbol{x} = \bar{V}_0 \boldsymbol{z}$ is an optimal solution of the problem in the condition 3, with the optimal value of n. If there is a vector $\boldsymbol{x} \in \{-1,1\}^n$ such that the maximization problem in the condition 3 achieves the optimal value of n, we have $n = \|\bar{V}^T \boldsymbol{x}\|^2 = \|\bar{V}_0^T \boldsymbol{x}\|^2 + \|\bar{V}_+^T \boldsymbol{x}\|^2 = \|\bar{V}_0^T \boldsymbol{x}\|^2$. Thus, $\|\bar{V}_+^T \boldsymbol{x}\|^2 = 0$ and $\bar{V}_+^T \boldsymbol{x} = 0$, from which we have $\frac{1}{2}\boldsymbol{x}^T Q \boldsymbol{x} - (-\frac{1}{2}\boldsymbol{e}^T \boldsymbol{\mu}^*) = \frac{1}{2}\boldsymbol{x}^T \bar{V}_+ \bar{\Lambda}_+ \bar{V}_+^T \boldsymbol{x} = 0$. The vector $\boldsymbol{x}$ is an optimal solution and we have $p^* = d^*$. On the other hand, if $p^* = d^*$, we must have $\frac{1}{2}\boldsymbol{x}^T \bar{V}_+ \bar{\Lambda}_+ \bar{V}_+^T \boldsymbol{x} = 0$, which implies that $\bar{V}_+^T \boldsymbol{x} = 0$. Thus there exists a vector $\boldsymbol{z} \in \mathbb{R}^k$ such that $\boldsymbol{x} = \bar{V}_0 \boldsymbol{z}$.

After obtain an optimal solution of the problem (5.12), we can check whether there is a duality gap by solving the maximization problem

$$\bar{p} = \max\{\boldsymbol{x}^T \bar{V}_0 \bar{V}_0^T \boldsymbol{x} \mid \boldsymbol{x} \in \{-1,1\}^n\}. \tag{5.18}$$

By the convexity of the objective function in (5.18), we have

$$\bar{p} = \max_{\boldsymbol{x} \in [-1,1]^n} \boldsymbol{x}^T \bar{V}_0 \bar{V}_0^T \boldsymbol{x} = \max_{\boldsymbol{z} \in \mathcal{Z}} \boldsymbol{z}^T \boldsymbol{z},$$

where $\boldsymbol{z} = \bar{V}_0^T \boldsymbol{x}$ and

$$\mathcal{Z} = \{\bar{V}_0^T \boldsymbol{x} \mid \boldsymbol{x} \in [-1,1]^n\}.$$

Thus, the maximization problem becomes a problem of enumerating extreme points of the zonotope $\mathcal{Z}$. It is shown in [4] that the enumeration problem can be solved in $O(n^{k-1})$ for $k \geq 3$ and $O(n^k)$ for $k \leq 2$.

If $\bar{p} < n$, d^* only gives a lower bound. A method on how to construct a tighter lower bound is proposed in [86]. For $\boldsymbol{x} \in \{-1,1\}^n$, we have the following inequality

$$\begin{aligned} \frac{1}{2}\boldsymbol{x}^T Q \boldsymbol{x} &= d^* + \frac{1}{2}\boldsymbol{x}^T \bar{V}_+ \bar{\Lambda}_+ \bar{V}_+^T \boldsymbol{x} \\ &\geq d^* + \bar{\lambda}_{k+1} \frac{1}{2}\boldsymbol{x}^T \bar{V}_+ \bar{V}_+^T \boldsymbol{x} \\ &= d^* + \frac{1}{2}\bar{\lambda}_{k+1}(n - \bar{p}). \end{aligned}$$

Here, as it is defined, $\bar{\lambda}_{k+1}$ is the smallest positive eigenvalue. Thus, we have the following result.

Theorem 27 *Suppose $\bar{p} < n$. Then*

$$p^* \geq d^* + \frac{1}{2}(n - \bar{p})\bar{\lambda}_{k+1} > d^*. \tag{5.19}$$

A more complete analysis of the lower bound is then presented. Suppose that the matrix Q has the eigendecomposition

$$Q = V \Lambda V^T, \ \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n), \ V = [v_1, \ldots, v_n],$$

with

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n,$$

and v_i being the corresponding eigenvectors of λ_i. For each $j \in \{1, \ldots, n\}$, we denote the subspace spanned by a set of vectors $v_1, \ldots, v_j$ by $\mathcal{L}(v_1, \ldots, v_j)$, the distance between a vector $\boldsymbol{x}$ and the subspace $\mathcal{L}(v_1, \ldots, v_j)$ by $\mathrm{dist}(\boldsymbol{x}, \mathcal{L}(v_1, \ldots, v_j))$, and the distance between $\mathcal{X} = \{-1, 1\}^n$ and $\mathcal{L}(v_1, \ldots, v_j)$ by

$$d_j = \min\{\mathrm{dist}(\boldsymbol{x}, \mathcal{L}(v_1, \ldots, v_j)) \mid \boldsymbol{x} \in \mathcal{X}\}.$$

In [14], a strengthened lower bound is proposed.

Theorem 28 *The following inequality holds:*

$$p^* \geq \frac{1}{2}\left(n\lambda_1 + \sum_{j=1}^{n-1}(\lambda_{j+1} - \lambda_j)d_j^2\right). \tag{5.20}$$

For any vector $\boldsymbol{x} \in \mathcal{X}$, let $\boldsymbol{x} = V\boldsymbol{\alpha}$ with $\boldsymbol{\alpha} = \{\alpha_i\}$. The following inequality holds:

$$d_j^2 \leq \mathrm{dist}(\boldsymbol{x}, \mathcal{L}(v_1, \ldots, v_j)) = \sum_{i=j+1}^{n} \alpha_i^2,$$

which implies $\alpha_{j+1} \geq d_j^2 - \sum_{i=j+2}^{n} \alpha_i^2$. Hence we have

$$\begin{aligned}
\boldsymbol{x}^T Q \boldsymbol{x} &= n\lambda_1 + \sum_{i=2}^{n}(\lambda_i - \lambda_1)\alpha_i^2 \\
&\geq n\lambda_1 + (\lambda_2 - \lambda_1)d_1^2 - (\lambda_2 - \lambda_1)\sum_{i=3}^{n}\alpha_i^2 + \sum_{i=3}^{n}(\lambda_i - \lambda_1)\alpha_i^2 \\
&= n\lambda_1 + (\lambda_2 - \lambda_1)d_1^2 + \sum_{i=3}^{n}(\lambda_i - \lambda_2)\alpha_i^2
\end{aligned}$$

and, inductively, we get

$$\boldsymbol{x}^T Q \boldsymbol{x} \geq n\lambda_1 + \sum_{j=1}^{n-1}(\lambda_{j+1} - \lambda_j)d_j^2.$$

Thus the inequality (5.20) is proved.

Let $Q + \mathrm{diag}(\boldsymbol{\mu})$ replace Q and $\bar{\lambda}_1, \ldots, \bar{\lambda}_n$ be eigenvalues of $Q + \mathrm{diag}(\boldsymbol{\mu})$, as defined above. Similarly, it is true that

$$\boldsymbol{x}^T(Q + \mathrm{diag}(\boldsymbol{\mu}))\boldsymbol{x} \geq n\bar{\lambda}_1 + \sum_{j=1}^{n-1}(\bar{\lambda}_{j+1} - \bar{\lambda}_j)\bar{d}_j^2,$$

where $\bar{d}_j$ is the distance between $\mathcal{X}$ and $\mathcal{L}(\bar{v}_1, \ldots, \bar{v}_j)$, from which we have

$$p^* \geq \frac{1}{2}\left(n\bar{\lambda}_1 - \boldsymbol{e}^T\boldsymbol{\mu} + \sum_{j=1}^{n-1}(\bar{\lambda}_{j+1} - \bar{\lambda}_j)\bar{d}_j^2\right) = d^* + \frac{1}{2}\sum_{j=k}^{n-1}(\bar{\lambda}_{j+1} - \bar{\lambda}_j)\bar{d}_j^2. \tag{5.21}$$

On the other hand, we have

$$\begin{aligned}\bar{d}_j &= \min\{\mathrm{dist}(\boldsymbol{x}, \mathcal{L}(\bar{v}_1, \ldots, \bar{v}_j)) \mid \boldsymbol{x} \in \mathcal{X}\} \\ &= \min\{n - \sum_{i=1}^{j}\alpha_i^2 \mid \boldsymbol{x} = V\boldsymbol{\alpha} \in \mathcal{X}\} \\ &= n - \max\{\boldsymbol{x}^T\bar{V}_0\bar{V}_0^T\boldsymbol{x} \mid \boldsymbol{x} \in \mathcal{X}\}\end{aligned}$$

Hence, if omit the items with $j > k$ in the equation (5.21), we get

$$d^* + \frac{1}{2}\bar{\lambda}_{k+1}\bar{d}_k^2 = d^* + \frac{1}{2}\bar{\lambda}_{k+1}(n - \bar{p}),$$

which is the lower bound provided in Theorem 27.

The following result offers a necessary optimality condition.

Theorem 29 *If the vector $\boldsymbol{x}^* \in \mathcal{X}$ satisfies: $dist(\boldsymbol{x}^*, \mathcal{L}(v_1, \ldots, v_i)) = d_i$ for all indices $i \in \{1, \ldots, n-1\}$ such that $\lambda_{i+1} > \lambda_i$, then $\boldsymbol{x}$ is an optimal solution.*

Let $d_i^* = \mathrm{dist}(\boldsymbol{x}^*, \mathcal{L}(v_1, \ldots, v_i))$ and $s = \{i \mid \lambda_{i+1} > \lambda_i, i = 1, \ldots, n-1\}$. For each $i \in s$, we have $d_i = d_i^* = \sum_{j=i+1}^{n}\alpha_j^2$. Thus, in the proof of Theorem 28, all the inequalities becomes equalities, and we have the equality holds in (5.20), i.e.,

$$p^* = \frac{1}{2}\left(n\lambda_1 + \sum_{j\in s}(\lambda_{j+1} - \lambda_j)d_j^2\right) = \frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x},$$

which shows that $\boldsymbol{x}$ is an optimal solution.

If there is only one entry in the set s, supposed $s = \{m\}$, the left side of the equation (5.20) becomes $\frac{1}{2}(n\lambda_1 + (\lambda_{m+1} - \lambda_m)d_m^2)$. We know there may not exists a vector $\boldsymbol{x}$ such that $d_i = \mathrm{dist}(\boldsymbol{x}, \mathcal{L}(v_1, \ldots, v_i))$ and $d_j = \mathrm{dist}(\boldsymbol{x}, \mathcal{L}(v_1, \ldots, v_j))$, which makes the equality in the equation (5.20) fail to hold. But, since only one such distance, d_m for $s = \{m\}$, appears in the expression, there is always a vector $\boldsymbol{x}$ satisfying $d_m = \mathrm{dist}(\boldsymbol{x}, \mathcal{L}(v_1, \ldots, v_m))$. This vector $\boldsymbol{x}$ is an optimal solution and the equality in equation (5.20) holds.

5.3 Canonical duality for binary quadratic problems

5.3.1 Canonical dual problem

The integer constraints in the problem $(\mathcal{P}_{bqp})$ is treated as equality ones, $x_i^2 = 1, i = 1, \ldots, n$. Then follow the procedure of the canonical duality transformation discussed

in Chapter 3, we can easily get the canonical dual problem. Let

$$G(\boldsymbol{\sigma}) = Q + \text{diag}(\boldsymbol{\sigma}),$$

and

$$\mathcal{S}_a = \{\boldsymbol{\sigma} \in \mathbb{R}^n \,|\, \det(G(\boldsymbol{\sigma})) \neq 0\}.$$

The canonical dual function is formulated as

$$\Pi^d(\boldsymbol{\sigma}) = -\frac{1}{2}\boldsymbol{f}^T G(\boldsymbol{\sigma})^{-1}\boldsymbol{f} - \frac{1}{2}\boldsymbol{e}^T\boldsymbol{\sigma}, \tag{5.22}$$

and the canonical dual problem is defined as

$$(\mathcal{P}^d_{bqp}) \quad \text{ext}\{\Pi^d(\boldsymbol{\sigma}) \,|\, \boldsymbol{\sigma} \in \mathcal{S}_a\}. \tag{5.23}$$

Notice that if $\boldsymbol{f} = 0$, the function $\Pi^d(\boldsymbol{\sigma})$ is equal to the objective function in the problem (5.12), which is thus same to the problem of maximizing Π^d over the positive semidefinite region defined by

$$\mathcal{S}_c^+ = \{\boldsymbol{\sigma} \in \mathcal{S}_a \,|\, G(\boldsymbol{\sigma}) \succeq 0\}.$$

5.3.2 Global optimality conditions

The following theorem characterizes the primal-dual relation.

Theorem 30 (Complementary-dual principle) *If $\bar{\boldsymbol{\sigma}} \in \mathcal{S}_a$ is a critical point of $\Pi^d(\boldsymbol{\sigma})$, the vector*

$$\bar{\boldsymbol{x}} = G(\bar{\boldsymbol{\sigma}})^{-1}\boldsymbol{f} \tag{5.24}$$

is a feasible solution of ($\mathcal{P}_{bqp}$).

Let $\bar{\boldsymbol{x}} \in \mathcal{X}$ and

$$\bar{\boldsymbol{\sigma}} = \bar{\boldsymbol{x}} \circ \boldsymbol{f} - \bar{\boldsymbol{x}} \circ (Q\bar{\boldsymbol{x}}). \tag{5.25}$$

If $\bar{\boldsymbol{\sigma}} \in \mathcal{S}_a$, then $\bar{\boldsymbol{\sigma}}$ is a critical point of $\Pi^d(\boldsymbol{\sigma})$.

For both statements, we have

$$\Pi(\bar{\boldsymbol{x}}) = \Pi^d(\bar{\boldsymbol{\sigma}}). \tag{5.26}$$

Proof: From the assumption of $\bar{\boldsymbol{\sigma}}$ being a critical point and the definition of $\bar{\boldsymbol{x}}$, we have

$$\begin{aligned}\nabla_{\boldsymbol{\sigma}}\Pi^d(\bar{\boldsymbol{\sigma}}) &= \left\{\frac{1}{2}\boldsymbol{f}^T G(\bar{\boldsymbol{\sigma}})^{-1} I_i G(\bar{\boldsymbol{\sigma}})^{-1}\boldsymbol{f}\right\}_{i=1}^n - \frac{1}{2}\boldsymbol{e} \\ &= \frac{1}{2}\bar{\boldsymbol{x}} \circ \bar{\boldsymbol{x}} - \frac{1}{2}\boldsymbol{e} = 0,\end{aligned}$$

where I_i denotes an all-zero matrix except for the entry (i,i), which is equal to one. It is verified that $\bar{\boldsymbol{x}}$ is a feasible solution of the primal problem $(\mathcal{P}_{bqp})$. The definition of $\bar{\boldsymbol{x}}$ in equation (5.24) also implies that $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{\sigma}}$ satisfy the equilibrium equation

$$G(\bar{\boldsymbol{\sigma}})\bar{\boldsymbol{x}} - \boldsymbol{f} = 0, \tag{5.27}$$

from which we have

$$\bar{\boldsymbol{x}} \circ \bar{\boldsymbol{\sigma}} = \boldsymbol{f} - Q\bar{\boldsymbol{x}}. \tag{5.28}$$

The equality (5.28) is further equivalent to

$$\bar{\boldsymbol{x}} \circ \bar{\boldsymbol{x}} \circ \bar{\boldsymbol{\sigma}} = \bar{\boldsymbol{x}} \circ \boldsymbol{f} - \bar{\boldsymbol{x}} \circ (Q\bar{\boldsymbol{x}}).$$

Thus, we have

$$\bar{\boldsymbol{\sigma}} = \bar{\boldsymbol{x}} \circ \boldsymbol{f} - \bar{\boldsymbol{x}} \circ (Q\bar{\boldsymbol{x}}) = \mathrm{diag}(\bar{\boldsymbol{x}})\boldsymbol{f} - \mathrm{diag}(\bar{\boldsymbol{x}})Q\bar{\boldsymbol{x}}. \tag{5.29}$$

Now we are ready to prove the equation (5.26):

$$\begin{aligned}
\Pi^d(\bar{\boldsymbol{\sigma}}) &= -\frac{1}{2}\boldsymbol{f}^T G(\bar{\boldsymbol{\sigma}})^{-1}\boldsymbol{f} - \frac{1}{2}\boldsymbol{e}^T\bar{\boldsymbol{\sigma}} \\
&= -\frac{1}{2}\boldsymbol{f}^T\bar{\boldsymbol{x}} - \frac{1}{2}\boldsymbol{e}^T\left(\mathrm{diag}(\bar{\boldsymbol{x}})\boldsymbol{f} - \mathrm{diag}(\bar{\boldsymbol{x}})Q\bar{\boldsymbol{x}}\right) \\
&= \frac{1}{2}\bar{\boldsymbol{x}}^T Q\bar{\boldsymbol{x}} - \boldsymbol{f}^T\bar{\boldsymbol{x}} = \Pi(\bar{\boldsymbol{x}}).
\end{aligned}$$

Let $\bar{\boldsymbol{\sigma}}$ be defined by the equation (5.25). From the equivalence between (5.27) and (5.29), it can be easily proved that $\bar{\boldsymbol{\sigma}}$ is a critical point if $\bar{\boldsymbol{\sigma}} \in \mathcal{S}_a$.

The theorem is proved. □

Besides the positive semidefinite region $\mathcal{S}_c^+$, we also introduce the negative semidefinite region:

$$\mathcal{S}_c^- = \{\boldsymbol{\sigma} \in \mathcal{S}_a \mid G(\boldsymbol{\sigma}) \preceq 0\}.$$

First, $\mathcal{S}_c^+$ and $\mathcal{S}_c^-$ are convex sets. Second, from the expression of Hessian of the dual function $\Pi^d(\boldsymbol{\sigma})$, we notice that $\Pi^d(\bar{\boldsymbol{\sigma}})$ is concave on $\mathcal{S}_c^+$ and convex on $\mathcal{S}_c^-$. Hence, any critical point in $\mathcal{S}_c^+$ is a maximizer of Π^d over $\mathcal{S}_c^+$, and any critical point in $\mathcal{S}_c^-$ is a minimizer of Π^d over $\mathcal{S}_c^-$. We have the following result.

Theorem 31 *For any given matrix $Q \in \mathbb{S}^n$ and vector $\boldsymbol{f} \in \mathbb{R}^n$, suppose $\bar{\boldsymbol{\sigma}}$ is a critical point of the dual function $\Pi^d(\boldsymbol{\sigma})$ and $\bar{\boldsymbol{x}} = G(\bar{\boldsymbol{\sigma}})^{-1}\boldsymbol{f}$.*

1. *If $\bar{\boldsymbol{\sigma}} \in \mathcal{S}_c^+$, then $\bar{\boldsymbol{x}}$ is a global minimizer of $\Pi(\boldsymbol{x})$ on $\mathcal{X}$; we have*

$$\Pi(\bar{\boldsymbol{x}}) = \min_{\boldsymbol{x}\in\mathcal{X}} \Pi(\boldsymbol{x}) = \max_{\boldsymbol{\sigma}\in\mathcal{S}_c^+} \Pi^d(\boldsymbol{\sigma}) = \Pi^d(\bar{\boldsymbol{\sigma}}). \tag{5.30}$$

2. If $\bar{\boldsymbol{\sigma}} \in \mathcal{S}_c^-$, then $\bar{\boldsymbol{x}}$ is a global maximizer of $\Pi(\boldsymbol{x})$ on $\mathcal{X}$; we have

$$\Pi(\bar{\boldsymbol{x}}) = \max_{\boldsymbol{x}\in\mathcal{X}} \Pi(\boldsymbol{x}) = \min_{\boldsymbol{\sigma}\in\mathcal{S}_c^-} \Pi^d(\boldsymbol{\sigma}) = \Pi^d(\bar{\boldsymbol{\sigma}}). \tag{5.31}$$

Proof: The last equalities in (5.30) and (5.30) are obvious, since $\Pi^d(\boldsymbol{\sigma})$ is concave on $\mathcal{S}_c^+$ and convex on $\mathcal{S}_c^-$. For the minimizer, the weak duality in (3.30) shows that we always have

$$\Pi^d(\bar{\boldsymbol{\sigma}}) = \max_{\boldsymbol{\sigma}\in\mathcal{S}_c^+} \Pi^d(\boldsymbol{\sigma}) \le \min_{\boldsymbol{x}\in\mathcal{X}} \Pi(\boldsymbol{x}).$$

Thus, by Theorem 30, we have

$$\Pi(\bar{\boldsymbol{x}}) = \Pi^d(\bar{\boldsymbol{\sigma}}) = \min_{\boldsymbol{x}\in\mathcal{X}} \Pi(\boldsymbol{x}),$$

and it is proved that $\bar{\boldsymbol{x}}$ is a minimizer of $\Pi(\boldsymbol{x})$ on $\mathcal{X}$.

The second part of the theorem can be similarly proved by applying the fact that the maximization of $\Pi(\boldsymbol{x})$ is equivalent to the minimization of $-\Pi(\boldsymbol{x})$ since there are finite feasible solutions in $\mathcal{X}$. □

The equation (5.30) shows that the critical point $\bar{\boldsymbol{\sigma}}$ is the maximizer of the dual function $\Pi^d(\bar{\boldsymbol{\sigma}})$ over $\mathcal{S}_c^+$. On the other hand, if the maximizer is a critical point of $\Pi^d(\bar{\boldsymbol{\sigma}})$, we can claim that the corresponding $\bar{\boldsymbol{x}}$ defined by the equation (5.24) is a global optimal solution of the primal problem. Follow the discussion in Section 3.3.4, the maximizer can be found by solving the following SDP problem:

$$\begin{aligned} &\max_{\boldsymbol{\sigma},\tau} \ \tau \\ &\text{s.t.} \ \begin{pmatrix} 2G(\boldsymbol{\sigma}) & \boldsymbol{f} \\ \boldsymbol{f}^T & -\frac{1}{2}\boldsymbol{e}^T\boldsymbol{\sigma} - \tau \end{pmatrix} \succeq 0 \end{aligned} \tag{5.32}$$

Let $(\bar{\boldsymbol{\sigma}}, \bar{\tau})$ be a maximizer. Then, if $G(\bar{\boldsymbol{\sigma}}) \succ 0$, $\bar{\boldsymbol{\sigma}}$ must be a critical point of Π^d and $\bar{\boldsymbol{x}}$ must be a global solution. While if $\det(G(\bar{\boldsymbol{\sigma}})) = 0$, $\bar{\boldsymbol{\sigma}}$ may not be a critical point and $\Pi^d(\bar{\boldsymbol{\sigma}})$ is only a lower bound for the primal problem. It shows that the integer problem can be converted into a convex optimization problem, which can be solved efficiently by well-developed convex optimization methods.

Corollary 32 *Suppose that $\bar{\boldsymbol{\sigma}}$ is a maximizer of the problem (5.32) and $\bar{\boldsymbol{x}} = G(\bar{\boldsymbol{\sigma}})^{-1}\boldsymbol{f}$. If $G(\bar{\boldsymbol{\sigma}}) \succ 0$, then $\bar{\boldsymbol{x}}$ is a global optimal solution of the problem ($\mathcal{P}_{bqp}$).*

5.3.3 Existence and uniqueness

The following result gives a criterion of existence and uniqueness of a critical point in $\mathcal{S}_c^+$.

Theorem 33 *If, for any $\boldsymbol{\sigma}_0$ with $det(G(\boldsymbol{\sigma}_0)) = 0$ and $G(\boldsymbol{\sigma}_0) \succeq 0$ and any $\boldsymbol{\sigma} \in \mathcal{S}_c^+$, we have*

$$\lim_{t\to 0^+} \Pi^d(\boldsymbol{\sigma}_0 + t\boldsymbol{\sigma}) = -\infty, \tag{5.33}$$

then the canonical dual problem ($\mathcal{P}_{bqp}^d$) has a unique critical point $\bar{\boldsymbol{\sigma}}$ in $\mathcal{S}_c^+$.

Proof: The assumption in the equation (5.33), plus the fact that $\Pi^d(\boldsymbol{\sigma})$ approaches to minus infinity as any entry of $\boldsymbol{\sigma}$ increases infinitely, implies that the function $\Pi^d(\boldsymbol{\sigma})$ is coercive on the convex set $\mathcal{S}_c^+$. Since $\Pi^d(\boldsymbol{\sigma})$ is concave over $\mathcal{S}_c^+$, it has at least one maximizer, which must be a critical point. We use $\bar{\boldsymbol{\sigma}}$ to denote one of the maximizers. The uniqueness results from the fact that the Hessian of $\Pi^d(\boldsymbol{\sigma})$ is negative definite at the critical point $\bar{\boldsymbol{\sigma}}$. □

5.3.4 Examples

Example 1 Let

$$Q = \begin{bmatrix} -2 & -3 \\ -3 & -1 \end{bmatrix}, \text{ and } \boldsymbol{f} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}.$$

In this case, the dual function has four critical points,

$$\bar{\boldsymbol{\sigma}}_1 = (4, 6),\ \bar{\boldsymbol{\sigma}}_2 = (6, 2),\ \bar{\boldsymbol{\sigma}}_3 = (0, 0),\ \text{and } \bar{\boldsymbol{\sigma}}_4 = (-2, -4),$$

with function values

$$\Pi^d(\bar{\boldsymbol{\sigma}}_1) = -5.5 < \Pi^d(\bar{\boldsymbol{\sigma}}_2) = -3.5 < \Pi^d(\bar{\boldsymbol{\sigma}}_3) = -1.5 < \Pi^d(\bar{\boldsymbol{\sigma}}_4) = 4.5.$$

The corresponding solutions of the primal problem are

$$\bar{\boldsymbol{x}}_1 = (-1, -1),\ \bar{\boldsymbol{x}}_2 = (1, 1),\ \bar{\boldsymbol{x}}_3 = (1, -1),\ \text{and } \bar{\boldsymbol{x}}_4 = (-1, 1).$$

By checking the eigenvalues of $Q+\text{diag}(\boldsymbol{\sigma})$, we find $Q+\text{diag}(\boldsymbol{\sigma}_1) \succeq 0$, $Q+\text{diag}(\boldsymbol{\sigma}_4) \preceq 0$, and $Q+\text{diag}(\boldsymbol{\sigma}_2)$ and $Q+\text{diag}(\boldsymbol{\sigma}_3)$ are indefinite. Thus, Theorem 30 and Theorem 31 are demonstrated.

Example 2 Let

$$Q = \begin{bmatrix} -22 & 9 & 1 \\ 9 & -140 & 6 \\ 1 & 6 & -80 \end{bmatrix}, \text{ and } \boldsymbol{f} = \begin{pmatrix} -2 \\ -6 \\ -1 \end{pmatrix}.$$

The dual function has eight critical points:

$$\begin{aligned} &\bar{\boldsymbol{\sigma}}_1 = (12, 128, 73), && \bar{\boldsymbol{\sigma}}_2 = (10, 119, 72), && \bar{\boldsymbol{\sigma}}_3 = (11, 134, 7), \\ &\bar{\boldsymbol{\sigma}}_4 = (11, 125, 8), && \bar{\boldsymbol{\sigma}}_5 = (19, 21, 78), && \bar{\boldsymbol{\sigma}}_6 = (3, 12, 79), \\ &\bar{\boldsymbol{\sigma}}_7 = (20, 15, 2). && \bar{\boldsymbol{\sigma}}_8 = (2, 6, 1) \end{aligned}$$

The corresponding primal solutions are

$$\begin{aligned} &\bar{\boldsymbol{x}}_1 = (0, 1, 1), && \bar{\boldsymbol{x}}_2 = (1, 1, 1), && \bar{\boldsymbol{x}}_3 = (0, 1, 0), \\ &\bar{\boldsymbol{x}}_4 = (1, 1, 0), && \bar{\boldsymbol{x}}_5 = (1, 0, 1), && \bar{\boldsymbol{x}}_6 = (0, 0, 1), \\ &\bar{\boldsymbol{x}}_7 = (1, 0, 0). && \bar{\boldsymbol{x}}_8 = (0, 0, 0) \end{aligned}$$

with function values

$$\Pi(\bar{\boldsymbol{x}}_1) = -97, \Pi(\bar{\boldsymbol{x}}_2) = -96, \Pi(\bar{\boldsymbol{x}}_3) = -64, \Pi(\bar{\boldsymbol{x}}_4) = -64,$$
$$\Pi(\bar{\boldsymbol{x}}_5) = -47, \Pi(\bar{\boldsymbol{x}}_6) = -39, \Pi(\bar{\boldsymbol{x}}_7) = -9, \Pi(\bar{\boldsymbol{x}}_8) = 0.$$

It can be verified that $Q + \text{diag}(\bar{\boldsymbol{\sigma}}_1)$ is positive definite, $Q + \text{diag}(\bar{\boldsymbol{\sigma}}_8)$ is negative definite and all $Q + \text{diag}(\bar{\boldsymbol{\sigma}}_i)$ for $i = 2, \ldots, 7$ are indefinite. The function values show that $\bar{\boldsymbol{x}}_1$ is the minimizer and $\bar{\boldsymbol{x}}_8$ is the maximizer. Thus, Theorem 30 and Theorem 31 are explained.

5.4 Perturbed problems

5.4.1 Canonical duality for perturbed problems

As mentioned in the previous section, the problem ($\mathcal{P}_{bqp}$) is equivalent to the following perturbed problem

$$\begin{aligned} \min_{\boldsymbol{x}} \ & \Pi_\alpha(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T Q_\alpha \boldsymbol{x} - \boldsymbol{f}^T \boldsymbol{x} \\ \text{s.t. } & \boldsymbol{x} \in \mathcal{X}, \end{aligned} \tag{5.34}$$

where $Q_\alpha = Q - \text{diag}(\boldsymbol{\alpha})$ and $\alpha \geq 0$. For any given indefinite matrix $Q \in \mathbb{S}^n$, there exist vectors $\boldsymbol{\alpha} \in \mathbb{R}^n$ which can make Q_α be either positive definite or negative definite. Here, we choose $\boldsymbol{\alpha}$ such that $Q_\alpha \prec 0$. Then the function $\Pi_\alpha(\boldsymbol{x})$ is strictly concave.

By Legendre-Fenchel transformation, the problem (5.34) is equivalent to

$$\min\{-\boldsymbol{x}^T\boldsymbol{z} - \frac{1}{2}(\boldsymbol{z} - \boldsymbol{f})^T Q_\alpha^{-1}(\boldsymbol{z} - \boldsymbol{f}) \mid \boldsymbol{x} \in \mathcal{X},\ \boldsymbol{z} \in \mathbb{R}^n\}. \tag{5.35}$$

Given $\boldsymbol{x} \in \mathcal{X}$, the objective function is strictly convex with respect to $\boldsymbol{z}$, where the stationary point and the minimizer is $\boldsymbol{z} = -Q_\alpha \boldsymbol{x} + \boldsymbol{f}$ and the function value is equal to $\Pi_\alpha(\boldsymbol{x})$. Thus, if $(\boldsymbol{x}, \boldsymbol{z})$ is an optimal solution of the problem 5.35, the vector $\boldsymbol{x}$ must be an optimal solution of the problem 5.34.

The total complementary function for the problem (5.35) is defined as

$$\Xi(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\sigma}) = -\boldsymbol{x}^T\boldsymbol{z} - \frac{1}{2}(\boldsymbol{z} - \boldsymbol{f})^T Q_\alpha^{-1}(\boldsymbol{z} - \boldsymbol{f}) + \frac{1}{2}\boldsymbol{x}^T \text{diag}(\boldsymbol{\sigma})\boldsymbol{x} - \frac{1}{2}\boldsymbol{e}^T\boldsymbol{\sigma}.$$

For any given $\boldsymbol{z}$, let derivatives of $\Xi(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\sigma})$ with respect to $\boldsymbol{x}$ and $\boldsymbol{\sigma}$ be equal to zero,

$$\text{diag}(\boldsymbol{\sigma})\boldsymbol{x} - \boldsymbol{z} = 0 \tag{5.36}$$
$$\boldsymbol{x} \circ \boldsymbol{x} = \boldsymbol{e} \tag{5.37}$$

We first notice in the problem (5.35) that if $(\boldsymbol{x}, \boldsymbol{z})$ is an optimal solution, $x_i = \text{sign}(z_i)$ for $z_i \neq 0$; otherwise, x_i could be either positive or negative. Here, we can

actually assume that the variable $\boldsymbol{\sigma}$ is nonnegative, and the assumption will not make optimal solutions of the problem (5.35) violate equations (5.36) and (5.37). Under this assumption, the value of $\boldsymbol{\sigma}$ satisfying (5.36) is equal to the absolute value of the given $\boldsymbol{z}$. The item $-\boldsymbol{x}^T\boldsymbol{z} + \frac{1}{2}\boldsymbol{x}^T\mathrm{diag}(\boldsymbol{\sigma})\boldsymbol{x} - \frac{1}{2}\boldsymbol{e}^T\boldsymbol{\sigma}$ in the function $\Xi(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\sigma})$ is then equal to $-\sum_{i=1}^n |z_i|$ for any $\boldsymbol{x}$ and $\boldsymbol{\sigma}$ satisfying (5.36) and (5.37) with a given $\boldsymbol{z}$. Then the following canonical dual problem can be defined

$$\min \ \mathrm{ext} \left\{ \Pi_\alpha^d(\boldsymbol{z}) = -\frac{1}{2}(\boldsymbol{z} - \boldsymbol{f})^T Q_\alpha^{-1}(\boldsymbol{z} - \boldsymbol{f}) - \sum_{i=1}^n |z_i| \,|\, \boldsymbol{z} \in \mathbb{R}^n \right\}. \tag{5.38}$$

Because of the negative definiteness of Q_α, the quadratic term in the function $\Pi_\alpha^d(\boldsymbol{z})$ is convex. While, the term $-\sum_{i=1}^n |z_i|$ is concave and nonsmooth. Thus, $\Pi_\alpha^d(\boldsymbol{z})$ is a nonconvex and nonsmooth function. First, we have the following result, whose proof can be found in [47].

Theorem 34 *Given $Q \in \mathbb{S}^n$ and $\boldsymbol{f} \in \mathbb{R}^n$, the problem (5.38) is canonically dual to the primal problem ($\mathcal{P}_{bqp}$) in the sense that if $\bar{\boldsymbol{z}}$ is a stationary point of the function $\Pi_\alpha^d(\boldsymbol{z})$, then the vector $\bar{\boldsymbol{x}}$ defined by*

$$\bar{x}_i = sign(\bar{z}_i), \tag{5.39}$$

where $\bar{x}_i$ can be either 1 or -1 if $\bar{z}_i = 0$, is a feasible solution of ($\mathcal{P}_{bqp}$), and $\Pi(\bar{\boldsymbol{x}}) = \Pi_\alpha^d(\bar{\boldsymbol{z}})$.

By fixing the sign of variable $\boldsymbol{z}$, the symbol of absolute value can be removed, and then $\Pi_\alpha^d(\boldsymbol{z})$ becomes a convex function. There are 2^n possible signs of $\boldsymbol{z}$, each of which is corresponding to confining the variable $\boldsymbol{z}$ in a hyperoctant. We can use vectors $\boldsymbol{x} \in \mathcal{X}$ to label the corresponding hyperoctants. In the following, when we say $\boldsymbol{z}$ is confined in the hyperoctant $\boldsymbol{x}$, it means that $\mathrm{sign}(z_i) = x_i$ for $i = 1, \ldots, n$. In each hyperoctant, we can use the corresponding vector $\boldsymbol{x} \in \mathcal{X}$ to remove the symbol of absolute value and write the function $\Pi_\alpha^d(\boldsymbol{z})$ as

$$\Pi_\alpha^d(\boldsymbol{z}) = -\frac{1}{2}(\boldsymbol{z} - \boldsymbol{f})^T Q_\alpha^{-1}(\boldsymbol{z} - \boldsymbol{f}) - \boldsymbol{x}^T\boldsymbol{z}, \tag{5.40}$$

which is a strictly convex function. If we ignore the confinement of the variable $\boldsymbol{z}$, the convex function $\Pi_\alpha^d(\boldsymbol{z})$ in the equation (5.40) has a minimizer, which is also a stationary point,

$$\boldsymbol{z} = \boldsymbol{f} - Q_\alpha \boldsymbol{x}. \tag{5.41}$$

If the stationary point $\boldsymbol{z}$ is in the hyperoctant $\boldsymbol{x}$, it is also a stationary point of the function $\Pi_\alpha^d(\boldsymbol{z})$ in (5.38). On the other hand, each stationary point in Theorem 34 must be the stationary point of the function $\Pi_\alpha^d(\boldsymbol{z})$ in (5.40) in a certain hyperoctant. Because the matrix Q_α is nonsingular, the equation (5.41) defines a one-to-one relation between hyperoctants and stationary points of all such possible functions in

(5.40). Some stationary points are in their corresponding hyperoctants, but others not. Nevertheless, for each hyperoctant $\boldsymbol{x}$ and the stationary point $\boldsymbol{z}$, we have

$$\Pi(\boldsymbol{x}) = \Pi_\alpha(\boldsymbol{z}). \tag{5.42}$$

The following result claims that if a hyperoctant does not contain its stationary point, it will not be an optimal solution of the primal problem ($\mathcal{P}_{bqp}$).

Lemma 35 *Suppose $Q \in \mathbb{S}^n$ and $\boldsymbol{f} \in \mathbb{R}^n$ are given, and $\boldsymbol{\alpha} \geq 0$ is any vector such that $Q_\alpha = Q - diag(\boldsymbol{\alpha}) \prec 0$. If $\bar{\boldsymbol{x}} \in \mathcal{X}$ is an optimal solution of the primal problem ($\mathcal{P}_{bqp}$), then there exists a stationary point $\bar{\boldsymbol{z}}$ in the canonical dual problem (5.38) such that $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{z}}$ satisfy the equation (5.41).*

Proof: As mentioned above, a vector $\boldsymbol{z}$ defined by the equation (5.41) for an $\boldsymbol{x} \in \mathcal{X}$ is a stationary point of the problem (5.38) if and only if $\boldsymbol{z}$ is in the hyperoctant $\boldsymbol{x}$. If $\bar{\boldsymbol{z}} = \boldsymbol{f} - Q_\alpha \bar{\boldsymbol{x}}$ is not a stationary point, the vector $\bar{\boldsymbol{z}}$ will be in a hyperoctant $\boldsymbol{x}_1$, which must be different to $\bar{\boldsymbol{x}}$. Let $\boldsymbol{z}_1 = \boldsymbol{f} - Q_\alpha \boldsymbol{x}_1$. Immediately, we have the following inequalities

$$\begin{aligned} -\frac{1}{2}(\bar{\boldsymbol{z}} - \boldsymbol{f})^T Q_\alpha^{-1}(\bar{\boldsymbol{z}} - \boldsymbol{f}) - \bar{\boldsymbol{x}}^T \bar{\boldsymbol{z}} &\geq -\frac{1}{2}(\bar{\boldsymbol{z}} - \boldsymbol{f})^T Q_\alpha^{-1}(\bar{\boldsymbol{z}} - \boldsymbol{f}) - \boldsymbol{x}_1^T \bar{\boldsymbol{z}} \\ &> -\frac{1}{2}(\boldsymbol{z}_1 - \boldsymbol{f})^T Q_\alpha^{-1}(\boldsymbol{z}_1 - \boldsymbol{f}) - \boldsymbol{x}_1^T \boldsymbol{z}_1. \end{aligned}$$

The first inequality results from the fact that $-\boldsymbol{x}_1^T \bar{\boldsymbol{z}} = -\sum_{i=1}^n |\bar{z}_i| \leq -\bar{\boldsymbol{x}}^T \bar{\boldsymbol{z}}$, while the second one is because of the strictly convexity. Then, combining with the equation (5.42), we have $\Pi(\bar{\boldsymbol{x}}) > \Pi(\boldsymbol{x}_1)$, which contradicts $\bar{\boldsymbol{x}}$ being an optimal solution of the primal problem ($\mathcal{P}_{bqp}$). □

By Lemma 35, there always exist stationary points for $\Pi_\alpha^d(\boldsymbol{z})$ in the problem (5.38), and hence the problem (5.38) is always feasible. We can remove ext in the problem (5.38) and redefine the dual problem as

$$\min \left\{ \Pi_\alpha^d(\boldsymbol{z}) = -\frac{1}{2}(\boldsymbol{z} - \boldsymbol{f})^T Q_\alpha^{-1}(\boldsymbol{z} - \boldsymbol{f}) - \sum_{i=1}^n |z_i| \mid \boldsymbol{z} \in \mathbb{R}^n \right\}. \tag{5.43}$$

We then have the following results which illustrates the relation between optimal solutions of the primal problem ($\mathcal{P}_{bqp}$) and the canonical dual problem (5.43).

Theorem 36 *Suppose $Q \in \mathbb{S}^n$ and $\boldsymbol{f} \in \mathbb{R}^n$ are given, and $\boldsymbol{\alpha} \geq 0$ is any vector such that $Q_\alpha = Q - diag(\boldsymbol{\alpha}) \prec 0$. If $\bar{\boldsymbol{z}}$ is an optimal solution of the canonical dual problem (5.43), then $\bar{\boldsymbol{x}}$ that satisfies the equation (5.41) with $\bar{\boldsymbol{z}}$ is an optimal solution of the primal problem ($\mathcal{P}_{bqp}$), and*

$$\Pi(\bar{\boldsymbol{x}}) = \min_{\boldsymbol{x} \in \mathcal{X}} \Pi(\boldsymbol{x}) = \min_{\boldsymbol{z} \in \mathbb{R}^n} \Pi_\alpha^d(\boldsymbol{z}) = \Pi_\alpha^d(\bar{\boldsymbol{z}}). \tag{5.44}$$

Proof: It must be true that the optimal solution $\bar{\boldsymbol{z}}$ is a stationary point of $\Pi_\alpha^d(\boldsymbol{z})$. Suppose that the optimal solution $\bar{\boldsymbol{z}}$ is in the hyperoctant $\boldsymbol{x} \in \mathcal{X}$. If $\bar{\boldsymbol{z}}$ is not a stationary point, then the vector $\boldsymbol{z} = \boldsymbol{f} - Q_\alpha \boldsymbol{x}$ is a stationary point of the function $\Pi_\alpha^d(\boldsymbol{z})$ in the equation (5.40), and $\boldsymbol{z}$ is not in the hyperoctant $\boldsymbol{x}$ since $\bar{\boldsymbol{z}}$ is a minimizer of $\Pi_\alpha^d(\boldsymbol{z})$ over the hyperoctant $\boldsymbol{x}$. Thus, we have $\Pi_\alpha^d(\boldsymbol{z}) < \Pi_\alpha^d(\bar{\boldsymbol{z}})$, and $\boldsymbol{x}$ will not be an optimal solution. From Lemma 35, we know there must be a stationary point of the function $\Pi_\alpha^d(\boldsymbol{z})$ for each optimal solution, whose value is strictly less than $\Pi(\boldsymbol{x}) = \Pi_\alpha^d(\boldsymbol{z})$ and is thus less than $\Pi_\alpha^d(\bar{\boldsymbol{z}})$. It contradicts the assumption that $\bar{\boldsymbol{z}}$ is a minimizer of the problem (5.43). Thus, $\bar{\boldsymbol{z}}$ must be a stationary point. Consequently, the vector $\bar{\boldsymbol{x}}$ is a feasible solution of the problem $(\mathcal{P}_{bqp})$. Because $\bar{\boldsymbol{z}}$ is a minimizer, its function value is less than or equal to all other stationary points. Therefore, $\bar{\boldsymbol{x}}$ is an optimal solution of the primal problem $(\mathcal{P}_{bqp})$ and the equation (5.44) is proved true. □

From the proof of Lemma 35, it is shown that a vector $\boldsymbol{x} \in \mathcal{X}$ will not be an optimal solution if the vector $\boldsymbol{z}$ defined by the equation (5.41) is not in the hyperoctant $\boldsymbol{x}$. We can use this conclusion to cut off many solutions which are not optimal ones. But if the parameter $\boldsymbol{\alpha}$ is not chosen properly, this conclusion will become useless.

Lemma 37 *Given $Q \in \mathbb{S}^n$ and $\boldsymbol{f} \in \mathbb{R}^n$, there exists a vector $\bar{\boldsymbol{\alpha}}$ such that for any $\boldsymbol{\alpha} \geq \bar{\boldsymbol{\alpha}}$ the function $\Pi_\alpha^d(\boldsymbol{z})$ in the problem (5.43) will have a stationary point in each hyperoctant $\boldsymbol{x} \in \mathcal{X}$.*

Proof: The function $\Pi_\alpha^d(\boldsymbol{z})$ has a stationary point in the hyperoctant $\boldsymbol{x}$ is equivalent to

$$\boldsymbol{z} \circ \boldsymbol{x} = (-\boldsymbol{f} - Q\boldsymbol{x}) \circ \boldsymbol{x} + \boldsymbol{\alpha} \geq 0. \tag{5.45}$$

If let $\bar{\alpha}_i = |-f_i - \sum_{j=1}^n q_{ij} x_j|$, the equation (5.45) is always true for any $\boldsymbol{\alpha} \geq \bar{\boldsymbol{\alpha}}$ and any $\boldsymbol{x} \in \mathcal{X}$. Thus, the lemma is proved. □

5.4.2 Analytically solvable cases

If Q is a diagonal matrix, i.e., $Q = \mathrm{diag}(\boldsymbol{q})$ with $\boldsymbol{q} = \{q_i\} \in \mathbb{R}^n$, the canonical dual function $\Pi^d(\boldsymbol{\sigma})$ has a simple form

$$\Pi^d(\boldsymbol{\sigma}) = -\sum_{i=1}^n \left(\frac{f_i^2}{2(q_i + \sigma_i)} + \frac{1}{2}\sigma_i \right).$$

Obviously, the condition (5.33) holds for $\Pi^d(\boldsymbol{\sigma})$ if $f_i \neq 0$. By the criticality condition $\nabla \Pi^d(\boldsymbol{\sigma}) = 0$, we have

$$|q_i + \sigma_i| = |f_i|, \ i = 1, \ldots, n,$$

and then have, for $i = 1, \ldots, n$,

$$x_i = \frac{f_i}{q_i + \sigma_i} = \begin{cases} \frac{f_i}{|f_i|}, & \text{if } q_i + \sigma_i > 0, \\ -\frac{f_i}{|f_i|}, & \text{if } q_i + \sigma_i < 0. \end{cases}$$

Therefore, by Theorem 31, we have the following result.

Corollary 38 *Suppose $f_i \neq 0$ for $i = 1, \ldots, n$ and $\bar{\boldsymbol{\sigma}}$ is a critical point of $\Pi^d(\boldsymbol{\sigma})$.*

1. *If $q_i + \sigma_i > 0,\ i = 1, \ldots, n$, then $\bar{\boldsymbol{x}} = \{\frac{f_i}{|f_i|}\}$ is a global solution of the primal problem ($\mathcal{P}_{bqp}$).*
2. *If $q_i + \sigma_i < 0,\ i = 1, \ldots, n$, then $\bar{\boldsymbol{x}} = \{-\frac{f_i}{|f_i|}\}$ is a global maximizer of $\Pi(\boldsymbol{x})$ over $\mathcal{X}$.*

Given any $\bar{\boldsymbol{x}} \in \mathcal{X}$, a vector $\bar{\boldsymbol{\sigma}}$ is defined by the equation (5.29), i.e.,

$$\bar{\sigma}_i = \bar{x}_i f_i - \sum_{j=1}^{n} q_{ij} \bar{x}_i \bar{x}_j,\ i = 1, \ldots, n.$$

The inequality holds

$$|f_i| - \sum_{j=1}^{n} q_{ij} \bar{x}_i \bar{x}_j \geq |f_i| - \sum_{j=1}^{n} |q_{ij}|.$$

Then, if it is true that $|f_i| - \sum_{j=1}^{n} |q_{ij}| \geq -\lambda_1, i = 1, \ldots, n$, we always have $\bar{\sigma}_i \geq -\lambda_1, i = 1, \ldots, n$ by choosing $\bar{x}_i = \text{sign}(f_i)$. By Theorem 30 and Theorem 31, we know that $\bar{\boldsymbol{\sigma}}$ is a critical point in $\mathcal{S}_c^+$. Thus, such a chosen vector $\bar{\boldsymbol{x}}$ must be an optimal solution of the problem ($\mathcal{P}_{bqp}$). We have proven the following result.

Theorem 39 *Given $Q \in \mathbb{S}^n$ and $\boldsymbol{f} \in \mathbb{R}^n$, if the following condition holds*

$$|f_i| - \sum_{j=1}^{n} |q_{ij}| \geq -\lambda_1, \tag{5.46}$$

where λ_1 is the smallest eigenvalue of the matrix Q, the vector $\bar{\boldsymbol{x}}$ defined by

$$\bar{x}_i = sign(f_i), i = 1, \ldots, n, \tag{5.47}$$

is an optimal solution of the primal problem ($\mathcal{P}_{bqp}$), and the vector $\bar{\boldsymbol{\sigma}}$ defined by

$$\bar{\boldsymbol{\sigma}} = \bar{\boldsymbol{x}} \circ (\boldsymbol{f} - Q\bar{\boldsymbol{x}}) \tag{5.48}$$

is the critical point of $\Pi^d(\boldsymbol{\sigma})$ in $\mathcal{S}_c^+$.

If start from the perturbed problem (5.34), we have $G(\boldsymbol{\sigma}) = Q_\alpha + \text{diag}(\boldsymbol{\sigma})$ in the dual function. Let $\boldsymbol{\alpha}$ be a vector such that the matrix Q_α is positive definite. Then, for a solution $\boldsymbol{x} \in \mathcal{X}$, if $\boldsymbol{\sigma}$ defined by the equation (5.29) is nonnegative, we immediately have $G(\boldsymbol{\sigma}) \succeq 0$ and thus $\boldsymbol{x}$ is an optimal solution. If the following inequality holds

$$|f_i| \geq (q_{ii} - \alpha_i) + \sum_{j \neq i} |q_{ij}|,$$

we will have

$$\begin{cases} f_i \geq (q_{ii} - \alpha_i) + \sum_{j \neq i} |q_{ij}| \geq (q_{ii} - \alpha_i)x_i + \sum_{j \neq i} q_{ij}x_j, & \text{if } f_i \geq 0, \\ f_i \leq -(q_{ii} - \alpha_i) - \sum_{j \neq i} |q_{ij}| \leq (q_{ii} - \alpha_i)x_i + \sum_{j \neq i} q_{ij}x_j, & \text{if } f_i < 0. \end{cases}$$

The vector $\boldsymbol{\sigma}$ will be nonnegative if the entries of $\boldsymbol{x}$ have values of

$$x_i = \begin{cases} 1, & f_i \geq (q_{ii} - \alpha_i) + \sum_{j \neq i} |q_{ij}|, \\ -1, & f_i \leq -(q_{ii} - \alpha_i) - \sum_{j \neq i} |q_{ij}|. \end{cases} \tag{5.49}$$

Then $\boldsymbol{x}$ will be an optimal solution for the problem $(\mathcal{P}_{bqp})$, and $\boldsymbol{\sigma}$ will be a critical point of the dual function $\Pi_\alpha^d(\boldsymbol{\sigma})$ in $\mathcal{S}_c^+$.

A symmetric matrix is also guaranteed to be positive semidefinite if it is a diagonally dominant matrix with nonnegative diagonal entries [70]. Hence, for the matrix $G(\boldsymbol{\sigma}) = Q + \text{diag}(\boldsymbol{\sigma})$, the positive semidefiniteness can be achieved if $\boldsymbol{\sigma}$ makes $G(\boldsymbol{\sigma})$ be diagonally dominant, i.e.,

$$q_{ii} + \sigma_i \geq \sum_{j \neq i} |q_{ij}|.$$

By replacing σ_i with $x_i(f_i - \sum_{j=1}^n q_{ij}x_j)$, it becomes

$$x_i(f_i - \sum_{j \neq i} q_{ij}x_j) \geq \sum_{j \neq i} |q_{ij}|,$$

which is always true if

$$|f_i| \geq 2 \sum_{j \neq i} |q_{ij}|$$

and the vector $\boldsymbol{x}$ is accordingly chosen. Thus, we have the following result.

Theorem 40 *Given a matrix $Q \in \mathbb{S}^n$ and a vector $\boldsymbol{f} \in \mathbb{R}^n$, if the following condition holds*

$$|f_i| \geq 2 \sum_{j \neq i} |q_{ij}|, \ i = 1, \ldots, n \tag{5.50}$$

the vector $\bar{\boldsymbol{x}}$ determined by

$$\bar{x}_i = \begin{cases} 1, & f_i \geq 2 \sum_{j \neq i} |q_{ij}|, \\ -1, & f_i \leq -2 \sum_{j \neq i} |q_{ij}|, \end{cases} \tag{5.51}$$

is an optimal solution of the primal problem $(\mathcal{P}_{bqp})$, and the vector $\bar{\boldsymbol{\sigma}}$ defined by

$$\bar{\boldsymbol{\sigma}} = \bar{\boldsymbol{x}} \circ (\boldsymbol{f} - Q_\alpha \bar{\boldsymbol{x}}) \tag{5.52}$$

is the critical point of $\Pi_\alpha^d(\boldsymbol{\sigma})$ in $\mathcal{S}_c^+$.

It should be pointed out that the conditions (5.46) and (5.50) can not replace each other in checking the analytical solvability, because the diagonal dominance is only a sufficient condition for the positive semidefiniteness.

Chapter 6

Nonconvex optimization of log-sum-exp functions and quartic polynomials

6.1 Introduction

In this chapter, we are interested in the following nonconvex global optimization problem:

$$(\mathcal{P}_{lse}) \qquad \min_{\boldsymbol{x}\in\mathbb{R}^n} \Pi(\boldsymbol{x}) = T(\boldsymbol{x}) + W(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x} - \boldsymbol{p}^T\boldsymbol{x} \tag{6.1}$$

in which $Q \in \mathbb{S}^n$ and $\boldsymbol{p} \in \mathbb{R}^n$. The quartic (or 4th-order) polynomial function $W(\boldsymbol{x})$ and the log-sum-exp function $T(\boldsymbol{x})$ are defined as

$$W(\boldsymbol{x}) = \sum_{i=1}^{r_1} \frac{\alpha_i}{2}\left(\frac{1}{2}\boldsymbol{x}^T A_i\boldsymbol{x} - \boldsymbol{g}_i^T\boldsymbol{x} - c_i\right)^2,$$

$$T(\boldsymbol{x}) = \frac{1}{\beta}\log\left[1 + \sum_{i=1}^{r_2}\exp\left(\beta\left(\frac{1}{2}\boldsymbol{x}^T B_i\boldsymbol{x} - \boldsymbol{h}_i^T\boldsymbol{x} - d_i\right)\right)\right],$$

where $A_i, B_i \in \mathbb{S}^n$, $\boldsymbol{g}_i, \boldsymbol{h}_i \in \mathbb{R}^n$, $c_i, d_i \in \mathbb{R}$, and $\alpha_i, \beta \in \mathbb{R}_{++}$.

The quartic polynomial $W(\boldsymbol{x})$ is the so-called double-well potential if $A_i \succeq 0,\ c_i > 0$ for $i = 1, \ldots, r_1$. This function has extensive applications in mathematical physics, for example, in [39] it was used to model post-buckling of beams. While $T(\boldsymbol{x})$ is one of the fundamental functions in engineering sciences, which arises broadly in regions including plasticity theory [123], nonsmooth variational problems [36], structural optimization problems [7], information theory [27], network communication systems [26, 20, 28], and robot manipulator designing [1, 2, 105]. In numerical analysis, the function $T(\boldsymbol{x})$ is often used to deal with minimax problems [98, 99, 100, 109].

The rest of this chapter is arranged as follows. We first show in Section 6.2 how a canonical dual problem can be constructed by the standard canonical duality transformation. Then in Section 6.3, the triality theory is presented and proved. A

special case of one-dimensional dual problem is discussed in Section 6.4, where we present a necessary and sufficient condition for the existence of a critical point in the positive semidefinite region. Finally, examples are provided in Section 6.5 to explain the canonical duality theory.

6.2 Canonical dual problem

Following the standard procedure of the canonical duality transformation, we first introduce a geometrical operator from $\mathbb{R}^n$ to $\mathcal{E}_a \subseteq \mathbb{R}^m$:

$$(\boldsymbol{\xi}(\boldsymbol{x}), \boldsymbol{\eta}(\boldsymbol{x})) = \left(\left\{ \frac{1}{2}\boldsymbol{x}^T A_i \boldsymbol{x} - \boldsymbol{g}_i^T \boldsymbol{x} \right\}_{i=1}^{r_1}, \left\{ \frac{1}{2}\boldsymbol{x}^T B_i \boldsymbol{x} - \boldsymbol{h}_i^T \boldsymbol{x} \right\}_{i=1}^{r_2} \right)$$

where $m = r_1 + r_2$. Generally speaking, for any given m and symmetrical matrices A_i and B_i, the range $\mathcal{E}_a$ may not be convex. But, the range $\mathcal{E}_a$ is always convex when $m = 2$: without loss of generality, we assume that $\boldsymbol{g}_1 = 0$ and $\boldsymbol{h}_1 = 0$; As proved in [21], we will have $\mathcal{E}_a = \{(\frac{1}{2}\mathrm{tr}(A_1\boldsymbol{X}), \frac{1}{2}\mathrm{tr}(B_1\boldsymbol{X})) \mid \boldsymbol{X} \in \mathbb{S}_+^n\}$, where the right side of the equation is the range of all positive semidefinite matrix under a linear transformation, and thus it is a convex set. We assume here that $\mathcal{E}_a$ is a convex set. Therefore, a canonical function can be defined on $\mathcal{E}_a$:

$$V(\boldsymbol{\xi}, \boldsymbol{\eta}) = V_1(\boldsymbol{\xi}) + V_2(\boldsymbol{\eta})$$

where

$$V_1(\boldsymbol{\xi}) = \sum_{i=1}^{r} \frac{\alpha_i}{2}(\xi_i - c_i)^2,$$
$$V_2(\boldsymbol{\eta}) = \frac{1}{\beta}\log\left[1 + \sum_{i=1}^{p}\exp\left(\beta(\eta_i - d_i)\right)\right].$$

Here ξ_i and η_i denote the ith entry of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$, respectively. Its conjugate function is

$$V^*(\boldsymbol{\sigma}, \boldsymbol{\tau}) = V_1^*(\boldsymbol{\sigma}) + V_2^*(\boldsymbol{\tau})$$

in which $V_1^*(\boldsymbol{\sigma})$ and $V_2^*(\boldsymbol{\tau})$ are conjugate functions of $V_1(\boldsymbol{\xi})$ and $V_2(\boldsymbol{\eta})$, with expressions

$$V_1^*(\boldsymbol{\sigma}) = \frac{1}{2}\boldsymbol{\sigma}^T \mathrm{diag}(\boldsymbol{\alpha})^{-1}\boldsymbol{\sigma} + \boldsymbol{\sigma}^T\boldsymbol{c},$$
$$V_2^*(\boldsymbol{\tau}) = \frac{1}{\beta}\left[\sum_{i=1}^{p}\tau_i\log(\tau_i) + (1 - \sum_{i=1}^{p}\tau_i)\log(1 - \sum_{i=1}^{p}\tau_i)\right] + \boldsymbol{\tau}^T\boldsymbol{d},$$

where $\boldsymbol{d} = \{d_i\}_{i=1}^p$, $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1}^r$ and $\boldsymbol{c} = \{c_i\}_{i=1}^r$. Obviously, V_1^* and V_2^* are twice differentiable in their domains. Let

$$\mathcal{E}_a^* = \{(\boldsymbol{\sigma}, \boldsymbol{\tau}) \in \mathbb{R}^m \mid \boldsymbol{\tau} > 0, \boldsymbol{\tau}^T\boldsymbol{e} < 1\}.$$

By the Legendre-Fenchel transformation (see Section 2.1.3), we have the following relation

$$(\boldsymbol{\xi}, \boldsymbol{\eta}) = (\nabla V_1^*(\boldsymbol{\sigma}), \nabla V_2^*(\boldsymbol{\tau})) \Longleftrightarrow (\boldsymbol{\sigma}, \boldsymbol{\tau}) = (\nabla V_1(\boldsymbol{\xi}), \nabla V_2(\boldsymbol{\eta})), \tag{6.2}$$

which is further equivalent to

$$V(\boldsymbol{\xi}, \boldsymbol{\eta}) + V^*(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \boldsymbol{\xi}^T \boldsymbol{\sigma} + \boldsymbol{\eta}^T \boldsymbol{\tau}. \tag{6.3}$$

Then the total complementary function $\Xi : \mathbb{R}^n \times \mathcal{E}_a^* \to \mathbb{R}$ can be defined by

$$\Xi(\boldsymbol{x}, \boldsymbol{\sigma}, \boldsymbol{\tau}) = \frac{1}{2} \boldsymbol{x}^T G(\boldsymbol{\sigma}, \boldsymbol{\tau}) \boldsymbol{x} - \boldsymbol{f}(\boldsymbol{\sigma}, \boldsymbol{\tau})^T \boldsymbol{x} - V_1^*(\boldsymbol{\sigma}) - V_2^*(\boldsymbol{\tau}), \tag{6.4}$$

where

$$G(\boldsymbol{\sigma}, \boldsymbol{\tau}) = Q + \sum_{i=1}^{r} \sigma_i A_i + \sum_{i=1}^{p} \tau_i B_i \text{ and } \boldsymbol{f}(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \boldsymbol{p} + \sum_{i=1}^{r} \sigma_i \boldsymbol{g}_i + \sum_{i=1}^{p} \tau_i \boldsymbol{h}_i.$$

In the following, we use G and $\boldsymbol{f}$ to abbreviate $G(\boldsymbol{\sigma}, \boldsymbol{\tau})$ and $\boldsymbol{f}(\boldsymbol{\sigma}, \boldsymbol{\tau})$. From the total complementary function, the canonical dual function $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau})$ is defined by

$$\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \text{ext} \left\{ \Xi(\boldsymbol{x}, \boldsymbol{\sigma}, \boldsymbol{\tau}) \mid \boldsymbol{x} \in \mathbb{R}^n \right\}. \tag{6.5}$$

Notice that for any given $(\boldsymbol{\sigma}, \boldsymbol{\tau})$, the total complementary function $\Xi(\boldsymbol{x}, \boldsymbol{\sigma}, \boldsymbol{\tau})$ is a quadratic function of $\boldsymbol{x}$ and its stationary points are the solutions of the following equation system

$$\nabla_{\boldsymbol{x}} \Xi(\boldsymbol{x}, \boldsymbol{\sigma}, \boldsymbol{\tau}) = G\boldsymbol{x} - \boldsymbol{f} = 0. \tag{6.6}$$

If $\det(G) \neq 0$ for a given $(\boldsymbol{\sigma}, \boldsymbol{\tau})$, $\boldsymbol{x}$ can be solved analytically and uniquely, and the canonical dual function $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau})$ is well defined and can be written as

$$\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau}) = -\frac{1}{2} \boldsymbol{f}^T G^{-1} \boldsymbol{f} - V_1^*(\boldsymbol{\sigma}) - V_2^*(\boldsymbol{\tau}). \tag{6.7}$$

The dual function $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau})$ is twice differentiable.

Let

$$\mathcal{S}_a = \left\{ (\boldsymbol{\sigma}, \boldsymbol{\tau}) \mid (\boldsymbol{\sigma}, \boldsymbol{\tau}) \in \mathcal{E}^*, \ \det(G) \neq 0 \right\}.$$

The canonical dual problem is defined as

$$(\mathcal{P}_{lse}^d) \quad \text{ext} \left\{ \Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau}) \mid (\boldsymbol{\sigma}, \boldsymbol{\tau}) \in \mathcal{S}_a \right\}. \tag{6.8}$$

Immediately, we have the following result about the relations between the problems $(\mathcal{P}_{lse})$ and $(\mathcal{P}_{lse}^d)$ on extreme points. It shows that there is no duality gap between the primal problem and the canonical dual problem.

Theorem 41 (**Complementary-Dual Principle**) *The problem $(\mathcal{P}_{lse}^d)$ is canonically dual to the problem $(\mathcal{P}_{lse})$ in the sense that:*

1. *If* $\bar{\boldsymbol{x}}$ *is a critical point of* $\Pi(\boldsymbol{x})$ *and*

$$(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) = (\nabla V_1(\bar{\boldsymbol{\xi}}), \nabla V_2(\bar{\boldsymbol{\eta}})) \in \mathcal{S}_a, \tag{6.9}$$

where $\bar{\boldsymbol{\xi}} = \boldsymbol{\xi}(\bar{\boldsymbol{x}})$ *and* $\bar{\boldsymbol{\eta}} = \boldsymbol{\eta}(\bar{\boldsymbol{x}})$*, then* $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ *is a critical point of the dual function* $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau})$*.*

2. *If* $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ *is a critical point of* $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau})$ *in* $\mathcal{S}_a$*, then*

$$\bar{\boldsymbol{x}} = G^{-1}\boldsymbol{f} \tag{6.10}$$

is a critical point of $\Pi(\boldsymbol{x})$*.*

Moreover, for both statements, $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ *is a critical point of* $\Xi(\boldsymbol{x}, \boldsymbol{\sigma}, \boldsymbol{\tau})$ *and*

$$\Pi(\bar{\boldsymbol{x}}) = \Xi(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) = \Pi^d(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}). \tag{6.11}$$

Proof: First, assume that $\bar{\boldsymbol{x}}$ is a critical point of $\Pi(\boldsymbol{x})$ and $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) \in \mathcal{S}_a$ is defined by (6.9). We have

$$\begin{aligned} 0 = \nabla\Pi(\bar{\boldsymbol{x}}) &= \nabla\boldsymbol{\eta}(\bar{\boldsymbol{x}})\nabla V_2(\bar{\boldsymbol{\eta}}) + \nabla\boldsymbol{\xi}(\bar{\boldsymbol{x}})\nabla V_1(\bar{\boldsymbol{\xi}}) + Q\boldsymbol{x} - \bar{\boldsymbol{f}} \\ &= \nabla\boldsymbol{\eta}(\bar{\boldsymbol{x}})\bar{\boldsymbol{\tau}} + \nabla\boldsymbol{\xi}(\bar{\boldsymbol{x}})\bar{\boldsymbol{\sigma}} + Q\bar{\boldsymbol{x}} - \bar{\boldsymbol{f}} \\ &= \bar{G}\bar{\boldsymbol{x}} - \bar{\boldsymbol{f}} \end{aligned} \tag{6.12}$$

where $\bar{G} = G(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ and $\bar{\boldsymbol{f}} = \boldsymbol{f}(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$. From the assumption, it is true that $\det(\bar{G}) \neq 0$, hence $\bar{\boldsymbol{x}}$ is the unique point that satisfies the equation (6.12). By the relation (6.2), the definition of $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ implies that $\nabla V_1^*(\bar{\boldsymbol{\sigma}}) = \bar{\boldsymbol{\xi}}$ and $\nabla V_2^*(\bar{\boldsymbol{\tau}}) = \bar{\boldsymbol{\eta}}$, where $\bar{\xi}_i$ and $\bar{\eta}_i$ are now equal to

$$\bar{\xi}_i = \frac{1}{2}\bar{\boldsymbol{f}}^T\bar{G}^{-1}A_i\bar{G}^{-1}\bar{\boldsymbol{f}}, \text{ and } \bar{\eta}_i = \frac{1}{2}\bar{\boldsymbol{f}}^T\bar{G}^{-1}B_i\bar{G}^{-1}\bar{\boldsymbol{f}}.$$

By the expression of the gradient of Π^d,

$$\nabla\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \begin{pmatrix} \left\{\frac{1}{2}\boldsymbol{f}^T G^{-1}A_i G^{-1}\boldsymbol{f} - \boldsymbol{g}_i^T G^{-1}\boldsymbol{f} - \partial V_1^*(\boldsymbol{\sigma})/\partial\sigma_i\right\}_{i=1}^{r_1} \\ \left\{\frac{1}{2}\boldsymbol{f}^T G^{-1}B_i G^{-1}\boldsymbol{f} - \boldsymbol{h}_i^T G^{-1}\boldsymbol{f} - \partial V_2^*(\boldsymbol{\tau})/\partial\tau_i\right\}_{i=1}^{r_2} \end{pmatrix}, \tag{6.13}$$

it shows that $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ is a critical point of $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau})$.

Conversely, we then assume that $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ is a critical point of $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau})$ in $\mathcal{S}_a$ and $\bar{\boldsymbol{x}} = \bar{G}^{-1}\bar{\boldsymbol{f}}$. From (6.13) and the definition of $\bar{\boldsymbol{x}}$, we have $\bar{\boldsymbol{\xi}} = \partial V_1^*(\bar{\boldsymbol{\sigma}})/\partial\sigma_i$ and $\bar{\boldsymbol{\eta}} = \partial V_2^*(\bar{\boldsymbol{\tau}})/\partial\tau_i$, which, combining the relation (6.2), imply that $\bar{\boldsymbol{\sigma}} = \nabla V_1(\bar{\boldsymbol{\xi}})$ and $\bar{\boldsymbol{\tau}} = \nabla V_2(\bar{\boldsymbol{\eta}})$. Thus, the equation (6.12) is proved true and $\bar{\boldsymbol{x}}$ is a critical point.

The rest of the theorem is obvious. Therefore, the theorem is proved. □

6.3 Triality theory

In this section we will study the optimality conditions for global and local solutions of the primal and dual problems. Let

$$\mathcal{S}_c^+ = \{(\boldsymbol{\sigma}, \boldsymbol{\tau}) \in \mathcal{S}_a \mid G \succeq 0\}, \text{ and } \mathcal{S}_c^- = \{(\boldsymbol{\sigma}, \boldsymbol{\tau}) \in \mathcal{S}_a \mid G \preceq 0\}.$$

It is easy to prove that both $\mathcal{S}_c^+$ and $\mathcal{S}_c^-$ are convex sets.

Now, we present the main result, which illustrates the relations between the primal and canonical dual problems on global and local solutions.

Theorem 42 (**Triality Theorem**) *Suppose that* $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ *is a critical point of* $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau})$, *and* $\bar{\boldsymbol{x}} = \bar{G}^{-1}\bar{\boldsymbol{f}}$.

1. *If* $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) \in \mathcal{S}_c^+$, *then the min-max duality holds in the form of*

$$\Pi(\bar{\boldsymbol{x}}) = \min_{\boldsymbol{x} \in \mathbb{R}^n} \Pi(\boldsymbol{x}) = \max_{(\boldsymbol{\sigma}, \boldsymbol{\tau}) \in \mathcal{S}_c^+} \Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \Pi^d(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}). \tag{6.14}$$

2. *If* $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) \in \mathcal{S}_c^-$, *the double-max duality holds in the form that if* $\bar{\boldsymbol{x}}$ *is a local maximizer of* $\Pi(\boldsymbol{x})$ *or* $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ *is a local maximizer of* $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau})$, *we have*

$$\Pi(\bar{\boldsymbol{x}}) = \max_{\boldsymbol{x} \in \mathcal{X}_0} \Pi(\boldsymbol{x}) = \max_{(\boldsymbol{\sigma}, \boldsymbol{\tau}) \in \mathcal{S}_0} \Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \Pi^d(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) \tag{6.15}$$

   for some neighborhood [1] $\mathcal{X}_0 \times \mathcal{S}_0 \subset \mathbb{R}^n \times \mathcal{S}_c^-$ *of* $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$.

3. *If* $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) \in \mathcal{S}_c^-$, *then the double-min duality holds conditionally as:*

   (a) *If* $m = n$, $\bar{\boldsymbol{x}}$ *being a local minimizer of* $\Pi(\boldsymbol{x})$ *is equivalent to* $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ *being a local minimizer of* $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau})$, *and we have*

$$\Pi(\bar{\boldsymbol{x}}) = \min_{\boldsymbol{x} \in \mathcal{X}_0} \Pi(\boldsymbol{x}) = \min_{(\boldsymbol{\sigma}, \boldsymbol{\tau}) \in \mathcal{S}_0} \Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \Pi^d(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) \tag{6.16}$$

   for some neighborhood $\mathcal{X}_0 \times \mathcal{S}_0 \subset \mathbb{R}^n \times \mathcal{S}_c^-$ *of* $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$.

   (b) *If* $m < n$, $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ *being a local minimizer or a saddle point of* $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau})$ *is equivalent to* $\bar{\boldsymbol{x}}$ *being a saddle point of* $\Pi(\boldsymbol{x})$.

   (c) *If* $m > n$, $\bar{\boldsymbol{x}}$ *being a local minimizer or a saddle point of* $\Pi(\boldsymbol{x})$ *is equivalent to* $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ *being a saddle point of* $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau})$.

Proof:

[1] We use the same definition of the neighborhood as defined in [40] (Note 1 on page 306), i.e., a subset $\mathcal{X}_0$ is said to be the neighborhood of the critical point $\bar{\boldsymbol{x}}$ if $\bar{\boldsymbol{x}}$ is the only critical point in $\mathcal{X}_0$.

1. Since $\bar{G} \succ 0$ when $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) \in \mathcal{S}_c^+$ and $D \succ 0$, the Hessian of the dual function is negative definite, $\nabla^2 \Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau}) \prec 0$, which implies that $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau})$ is strictly concave over $\mathcal{S}_c^+$. Hence, it is true that

$$\Pi^d(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) = \max_{(\boldsymbol{\sigma}, \boldsymbol{\tau}) \in \mathcal{S}_c^+} \Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau}).$$

The total complementary function $\Xi(\boldsymbol{x}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ is a convex function with respect to $\boldsymbol{x}$ in $\mathbb{R}^n$, which, plus the fact that $\bar{\boldsymbol{x}}$ is a critical point of $\Xi(\boldsymbol{x}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$, means that we have $\Xi(\boldsymbol{x}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) \geq \Xi(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ for any $\boldsymbol{x} \in \mathbb{R}^n$. Therefore, for any $\boldsymbol{x} \in \mathbb{R}^n$, we have

$$\Pi(\boldsymbol{x}) \geq \Xi(\boldsymbol{x}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) \geq \Xi(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) = \Pi(\bar{\boldsymbol{x}}),$$

in which the first inequality is due to the Fenchel-Young inequality. The equation (6.14) is proved.

2. Assume that $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ is a critical point of $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau})$ in $\mathcal{S}_c^-$. From the equaion (6.13), we have

$$0 = \nabla \Pi^d(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) = \begin{pmatrix} \boldsymbol{\xi}(\bar{\boldsymbol{x}}) - \nabla V_1^*(\bar{\boldsymbol{\sigma}}) \\ \boldsymbol{\eta}(\bar{\boldsymbol{x}}) - \nabla V_2^*(\bar{\boldsymbol{\tau}}) \end{pmatrix},$$

which is equivalent to

$$\bar{\boldsymbol{\sigma}} = \nabla V_1(\bar{\boldsymbol{\xi}}), \text{ and } \bar{\boldsymbol{\tau}} = \nabla V_2(\bar{\boldsymbol{\eta}}).$$

Then, the Hessian matrices of $\Pi(\boldsymbol{x})$ at $\bar{\boldsymbol{x}}$ and $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau})$ at $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ can be expressed as

$$\nabla^2 \Pi(\bar{\boldsymbol{x}}) = \bar{G} + \bar{F} \bar{M} \bar{F}^T, \tag{6.17}$$

$$\nabla^2 \Pi^d(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) = -\bar{F}^T \bar{G}^{-1} \bar{F} - \bar{M}^{-1}, \tag{6.18}$$

where $\bar{F} \in \mathbb{R}^{n \times m}$ and $\bar{M} \in \mathbb{R}^{m \times m}$ are defined by

$$\bar{F} = \left[A_1 \bar{\boldsymbol{x}} - \boldsymbol{g}_1, \ldots, A_{r_1} \bar{\boldsymbol{x}} - \boldsymbol{g}_{r_1}, B_1 \bar{\boldsymbol{x}} - \boldsymbol{h}_1, \ldots, B_{r_2} \bar{\boldsymbol{x}} - \boldsymbol{h}_{r_2} \right],$$
$$\bar{M} = \begin{bmatrix} \beta(\operatorname{diag}(\bar{\boldsymbol{\tau}}) - \bar{\boldsymbol{\tau}} \bar{\boldsymbol{\tau}}^T) & 0 \\ 0 & \operatorname{diag}(\boldsymbol{\alpha}) \end{bmatrix}.$$

By Lemma 44, we have

$$\nabla^2 \Pi(\bar{\boldsymbol{x}}) \preceq 0 \Longleftrightarrow \nabla^2 \Pi^d(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) \preceq 0.$$

Thus, $\bar{\boldsymbol{x}}$ being a local maximizer of $\Pi(\boldsymbol{x})$ is equivalent to $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ being a local maximizer of $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau})$.

If $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ is a critical point, since $\bar{G} \prec 0$ and $\bar{M} \prec 0$, we have $\nabla^2 \Pi^d(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) \prec 0$. Hence, there exists a neighborhood $\mathcal{S}_0 \subset \mathcal{S}_c^-$ around $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ such that for all $(\boldsymbol{\sigma}, \boldsymbol{\tau}) \in \mathcal{S}_0$, $\nabla^2 \Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau}) \preceq 0$. Since the map $\boldsymbol{x} = G^{-1} \boldsymbol{f}$ is continuous over $\mathcal{S}_a$, the image of the map over $\mathcal{S}_0$ is a neighborhood of $\bar{\boldsymbol{x}}$, which we denote as $\mathcal{X}_0$. It shows that $\bar{\boldsymbol{x}}$ and $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ are local maximizers.

3. We then prove the double-min duality.

   (a) Assume that $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ is a local minimizer of $\Pi^d(\boldsymbol{\sigma}, \boldsymbol{\tau})$ in $\mathcal{S}_c^-$. From

   $$\nabla^2 \Pi^d(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) = -\bar{F}^T \bar{G}^{-1} \bar{F} - \bar{M}^{-1} \succ 0,$$

   we have $-\bar{F}^T \bar{G}^{-1} \bar{F} \succ \bar{M}^{-1} \succ 0$, which implies that the matrix F is invertible. Immediately, we have

   $$-\bar{G}^{-1} \succeq (\bar{F}^T)^{-1} \bar{M}^{-1} \bar{F}^{-1},$$

   which is further equivalent to

   $$-\bar{G} \preceq \bar{F} \bar{M} \bar{F}^T.$$

   Thus, it is proved that $\nabla^2 \Pi(\bar{\boldsymbol{x}}) = \bar{G} + \bar{F} \bar{M} \bar{F}^T \succeq 0$ and $\bar{\boldsymbol{x}}$ is a local minimizer of $\Pi(\boldsymbol{x})$. The converse can be proved similarly.

   (b) We claim that $\bar{\boldsymbol{x}}$ is not a local minimizer of $\Pi(\boldsymbol{x})$. If $\bar{\boldsymbol{x}}$ is a local minimizer of $\Pi(\boldsymbol{x})$, we would have $\nabla^2 \Pi(\bar{\boldsymbol{x}}) = \bar{G} + \bar{F} \bar{M} \bar{F}^T \succeq 0$, which is equivalent to $\bar{F} \bar{M} \bar{F}^T \succeq -\bar{G}$. Since $-\bar{G} \succ 0$, it is true that matrix $\bar{F}$ has full rank and

   $$n = \operatorname{rank}(-\bar{G}) = \operatorname{rank}(\bar{F} \bar{M} \bar{F}^T) \leq \min \left\{ \operatorname{rank}(\bar{F}), \operatorname{rank}(\bar{D}) \right\} = m,$$

   which is contradictory to the assumption that $m < n$. Therefore, if $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ is a local minimizer or a saddle point, $\bar{\boldsymbol{x}}$ must be a saddle point. The converse is also true.

   (c) The proof is similar to that of case (b).

The theorem is proved. □

6.4 One-dimensional dual problem

Here, we consider a special case of the problem $(\mathcal{P}_{lse})$:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \Pi_1(\boldsymbol{x}) = T(\boldsymbol{x}) + \frac{1}{2} \boldsymbol{x}^T Q \boldsymbol{x} - \boldsymbol{p}^T \boldsymbol{x}. \tag{6.19}$$

where

$$T(\boldsymbol{x}) = \frac{1}{\beta} \log \left[1 + \exp \left(\beta \left(\frac{1}{2} \boldsymbol{x}^T A \boldsymbol{x} - \boldsymbol{b}^T \boldsymbol{x} - c \right) \right) \right].$$

We confine our discussion to the case where A is positive definite. Thus, without loss of generality, we can let $A = I$. The canonical dual function is a univariate function,

$$\Pi_1^d(\tau) = -\frac{1}{2} \boldsymbol{f}(\tau)^T G(\tau)^{-1} \boldsymbol{f}(\tau) - c\tau - \frac{1}{\beta} \left(\tau \log(\tau) + (1 - \tau) \log(1 - \tau) \right) \tag{6.20}$$

in which

$$G(\tau) = Q + \tau I, \text{ and } \boldsymbol{f}(\tau) = \boldsymbol{p} + \tau \boldsymbol{b}.$$

Assume that the matrix Q has a eigendecomposition of $Q = U\Lambda U^T$. The diagonal entities of Λ are the eigenvalues of the matrix Q in nondecreasing order,

$$\lambda_1 = \cdots = \lambda_k < \lambda_{k+1} \leq \cdots \leq \lambda_n.$$

The columns of U are the corresponding eigenvectors. If we let $\hat{\boldsymbol{p}} = U^T\boldsymbol{p}$ and $\hat{\boldsymbol{b}} = U^T\boldsymbol{b}$, the dual function can be rewritten as

$$\Pi_1^d(\tau) = -\frac{1}{2}\sum_{i=1}^{n} \frac{(\hat{p}_i + \tau\hat{b}_i)^2}{\lambda_i + \tau} - c\tau - \frac{1}{\beta}\left(\tau\log(\tau) + (1-\tau)\log(1-\tau)\right). \tag{6.21}$$

The first- and second-order derivatives of $\Pi_1^d(\tau)$ are

$$\nabla\Pi_1^d(\tau) = \frac{1}{2}\sum_{i=1}^{n} \frac{(\hat{p}_i - \lambda_i\hat{b}_i)^2}{(\lambda_i + \tau)^2} + \frac{1}{2}\sum_{i=1}^{n}\hat{b}_i^2 - c - \frac{1}{\beta}\log(\frac{\tau}{1-\tau}), \tag{6.22}$$

$$\nabla^2\Pi_1^d(\tau) = \sum_{i=1}^{n} \frac{(\hat{p}_i - \lambda_i\hat{b}_i)^2}{(\lambda_i + \tau)^3} - \frac{1}{\beta}\frac{1}{\tau(1-\tau)}, \tag{6.23}$$

and the set $\mathcal{S}_c^+$ is

$$\mathcal{S}_c^+ = \{\tau \mid 0 < \tau < 1, \tau > -\lambda_1\}.$$

Notice that if $\lambda_1 \geq 0$, i.e., the matrix Q is positive semidefinite, the matrix G is always positive definite for $\tau \in \mathcal{S}_c^+ = \{\tau \mid 0 < \tau < 1\}$. There exists a unique critical point of the dual function in $\mathcal{S}_c^+$. If $\lambda_1 \leq -1$, the set $\mathcal{S}_c^+$ will be empty. It can be proved that the minimization problem (6.19) is not lower bounded. For the case where $-1 < \lambda_1 < 0$, we have the following existence conditions for $\Pi_1^d(\tau)$ having a critical point in $\mathcal{S}_c^+$.

Theorem 43 *Suppose that $\lambda_i, i = 1, \ldots, n$ are defined as above and $-1 < \lambda_1 < 0$. Then there exists a critical point of $\Pi_1^d(\tau)$ in $\mathcal{S}_c^+$ if and only if*

$$\sum_{i=1}^{k}(\hat{p}_i - \lambda_1\hat{b}_i)^2 \neq 0, \text{ or} \tag{6.24}$$

$$\frac{1}{2}\sum_{i=k+1}^{n}(\hat{p}_i - \lambda_i\hat{b}_i)^2/(\lambda_i - \lambda_1)^2 + \frac{1}{2}\sum_{i=1}^{n}\hat{b}_i^2 - c - \frac{1}{\beta}\log(\frac{-\lambda_1}{1+\lambda_1}) > 0. \tag{6.25}$$

If $\Pi_1^d(\tau)$ has a critical point in $\mathcal{S}_c^+$, the critical point is unique. Let $\bar{\tau}$ denote the critical point. Then $\bar{\boldsymbol{x}} = G(\bar{\tau})^{-1}\boldsymbol{f}(\bar{\tau})$ is a global solution of the problem (6.19).

From the expression in (6.25), it shows clearly that the condition in (6.25) always holds if it is true that

$$\frac{1}{2}\sum_{i=1}^{n}\hat{b}_i^2 - c - \frac{1}{\beta}\log(\frac{-\lambda_1}{1+\lambda_1}) > 0.$$

Because of the log item, this criteria can only be used when $-1 < \lambda_1 < 0$.

x	$\Pi(x)$	optimality	(σ,τ)	$\Pi^d(\sigma,\tau)$	$\mathcal{S}_a$	optimality
$\bar{x}_1$	0.113	global minimizer	$(\bar{\sigma}_1,\bar{\tau}_1)$	0.113	$\mathcal{S}_c^+$	local maximizer
$\bar{x}_2$	1.688	local minimizer	$(\bar{\sigma}_2,\bar{\tau}_2)$	1.688	$\mathcal{S}_c^-$	saddle point
$\bar{x}_3$	5.661	local maximizer	$(\bar{\sigma}_3,\bar{\tau}_3)$	5.661	$\mathcal{S}_c^-$	local maximizer

Table 6.1: Dualities for Example 2.

6.5 Examples

In this section, two examples are provided to illustrate the canonical duality theory. By examining the critical points of the primal and dual functions, we show how the dualities in Theorem 42 are verified.

Example 1

Consider the one-dimensional problem:

$$\min_{x\in\mathbb{R}} \quad \Pi(x) = \log\left[1+\exp\left(0.5x^2-0.1\right)\right] + 5\left(x^2-1\right)^2 - 0.8x.$$

The corresponding canonical dual function is

$$\Pi^d(\sigma,\tau) = -\frac{0.32}{\tau+2\sigma} - \sigma - 0.05\sigma^2 - 0.1\tau - \left[\tau\log(\tau) + (1-\tau)\log(1-\tau)\right].$$

The graph of function $\Pi(x)$ and the contour of $\Pi^d(\sigma,\tau)$ are shown in Figure 6.1.

There are three critical points of the dual function $\Pi^d(\tau,\sigma)$:

$$\begin{pmatrix}\bar{\sigma}_1\\ \bar{\tau}_1\end{pmatrix} = \begin{pmatrix}0.098\\ 0.6\end{pmatrix}, \quad \begin{pmatrix}\bar{\sigma}_2\\ \bar{\tau}_2\end{pmatrix} = \begin{pmatrix}-0.71\\ 0.59\end{pmatrix} \begin{pmatrix}\bar{\sigma}_3\\ \bar{\tau}_3\end{pmatrix} = \begin{pmatrix}-9.983\\ 0.475\end{pmatrix},$$

which are corresponding to the solutions of the primal problem:

$$\bar{x}_1 = 1.005, \ \bar{x}_2 = -0.964, \ \text{and} \ \ \bar{x}_3 = -0.041.$$

The duality relations are shown in Table 6.1. The min-max duality is validated by x_1 and (σ_1,τ_1), and the double-max duality appears as both x_3 and (σ_1,τ_1) being local maximizers. Whereas, for the double-min duality, since $n<m$, a local minimizer of the primal problem must be corresponding to a saddle point of the dual problem, which is the case of x_3 and (σ_1,τ_1).

Example 2

We consider a nonconvex and nonsmooth optimization problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^2} \max\left\{x_1^2+x_2^2-x_2, \ \ -x_1^2-x_2^2+3x_2\right\}.$$

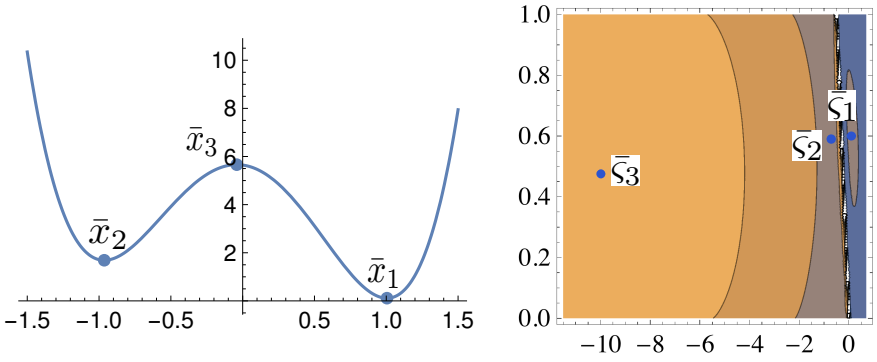

Figure 6.1: Example 1: the graph of function $\Pi(x)$ and the contour of function $\Pi^d(\sigma, \tau)$.

It's easy to verify that the optimal solution is $(0, 0)$ with function value of 0. Here, we use the log-sum-exp function to approximate the function $\max\{\cdot, \cdot\}$. We then get the following smooth optimization problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^2} \Pi(\boldsymbol{x}) = \frac{1}{\beta}\log\left[1 + \exp\left(\beta\left(2x_1^2 + 2x_2^2 - 4x_2\right)\right)\right] - x_1^2 - x_2^2 + 3x_2.$$

Its canonical dual function is

$$\Pi^d(\tau) = -\frac{1}{2}\frac{(4\tau - 3)^2}{4\tau - 2} - \frac{1}{\beta}\left[\tau\log\tau + (1-\tau)\log(1-\tau)\right].$$

First, we let β be a small value,

$$\beta = 2.5.$$

The primal function $\Pi(\boldsymbol{x})$ has three critical points:

$$\bar{\boldsymbol{x}}_1 = (0, -0.094),\ \bar{\boldsymbol{x}}_2 = (0, 1.856),\ \text{and}\ \bar{\boldsymbol{x}}_3 = (0, 1.528),$$

which are corresponding to the three critical points of $\Pi^d(\tau)$,

$$\bar{\tau}_1 = 0.728,\ \bar{\tau}_2 = 0.208,\ \text{and}\ \bar{\tau} = 0.026.$$

Then, we increase the value of β and let

$$\beta = 100.$$

The primal function $\Pi(\boldsymbol{x})$ still has three critical points, which are

$$\bar{\boldsymbol{x}}_1 = (0, -0.0027),\ \bar{\boldsymbol{x}}_2 = (0, 1.997),\ \text{and}\ \bar{\boldsymbol{x}}_3 = (0, 1.5).$$

The first two solutions are corresponding to

$$\bar{\tau}_1 = 0.749,\ \text{and}\ \bar{\tau}_2 = 0.249.$$

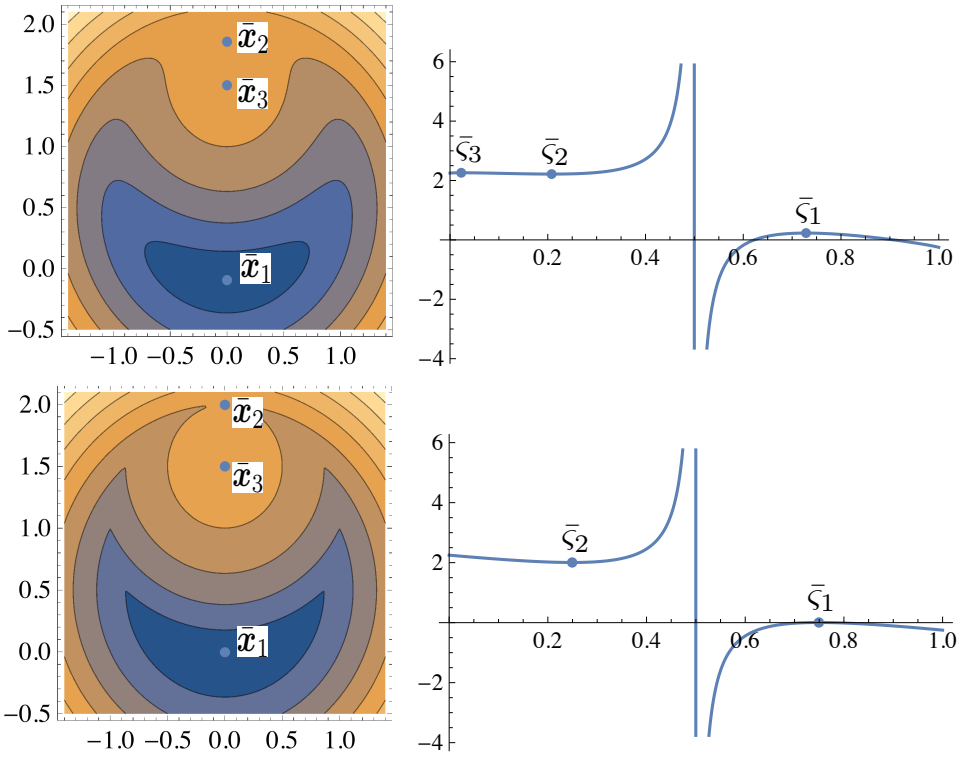

Figure 6.2: Example 2: the contour of $\Pi(\boldsymbol{x})$ and the graph of $\Pi^d(\tau)$; The two above are for the case of $\beta = 2.5$, and the two below are for the case of $\beta = 100$.

Whereas, for the third solution, it is clear that, from the formula (see the definition of (6.9) in Theorem 41)

$$\bar{\tau}_3 = \nabla V_2(\xi(\bar{\boldsymbol{x}}_3)) = \frac{\exp(-1.5\beta)}{1 + \exp(-1.5\beta)},$$

the value of $\bar{\tau}_3$ will approach 0, as β increases unboundedly. This feature is also shown in the graphs (see Figure 6.2). Here, the duality relation is mentioned only for the case of $\beta = 100$. It is clear, as shown in Figure 6.2, that $\bar{\tau}_1 \in \mathcal{S}_c^+$ and $\bar{\tau}_2 \in \mathcal{S}_c^-$. The min-max duality is true by the fact that $\bar{\boldsymbol{x}}_1$ is the global solution. The double-min duality is also verified, because of the fact that $n > m$ and $\boldsymbol{x}_2$ is a saddle point. Moreover, we have

$$\Pi(\bar{\boldsymbol{x}}_1) = \Pi^d(\bar{\tau}_1) = 0.006, \text{ and } \Pi(\bar{\boldsymbol{x}}_2) = \Pi^d(\bar{\tau}_2) = 2.006.$$

Chapter 7

Conclusions

In this thesis, the canonical duality theory is studied. Its three main parts, canonical dual transformation, complementary-dual principle and triality theory, are developed for a general optimization problem, which is required to satisfy only certain decomposition conditions. This general problem covers a wide range of optimization problems arising in the real world. It is shown, in the new result on the complementary-dual principle, that there is a one-to-one correspondence between all the KKT points of the canonical dual problem and of the primal problem; moreover, each pair of corresponding KKT points share the same function value. Then, the case where all the operators are quadratic is specifically studied. The triality theory reveals insightful information about global and local solutions. It is shown that as long as the global optimality condition holds true, the primal problem is equivalent to a convex problem in the dual space, which can be solved efficiently by existing convex methods; even if the condition does not hold, the convex problem still provides a lower bound that is at least as good as that by the Lagrangian relaxation method. It is also shown that through examining the canonical dual problem, the hidden convexity of the primal problem is easily observable, which is demonstrated using three examples.

For the spherically constrained quadratic minimization problem, a detailed study has been presented. It is shown that by the canonical duality, this nonconvex optimization is equivalent to a unified concave maximization problem over the positive semidefinite region in the dual space. Based on this canonical dual problem, sufficient and necessary conditions are obtained that separate problems into hard case and easy case. A perturbation method and an associated polynomial algorithm are proposed. Numerical results demonstrate that the proposed approach is able to solve large-size problems efficiently, including hard-case problems. Combining with the trust region method, this approach should be able to effectively solve general global optimization problems.

The lower bounds and analytically solvable cases are comprehensively studied for the binary quadratic problem, a fundamental problem in discrete optimization. The canonical duality is able to provide relaxations and lower bounds which are at least as good as that by Lagrangian relaxations. Combining with methods for finding critical points, the relaxations can be used to equip enumeration methods.

The perturbation technique plays a key role in exploring the analytically solvable cases; it is also effective in computing critical points in the dual space, especially when a critical point falls on the boundary of an interested region.

The third application studied here involves a very general nonlinear global optimization problem with 4th-order polynomials and log-sum-exp functions. By the canonical duality, it is concluded that if there is a critical point in the positive semidefinite region, the primal problem is equivalent to a concave maximization problem and its global solution can be calculated analytically from this critical point. Then two special cases are discussed: a fourth-order polynomial minimization problem and a minimax problem. For these two special cases, the min-max duality in the triality theory is reinforced and sufficient and necessary conditions for the existence of a critical point in the positive semidefinite region are discovered. The examples clearly demonstrate the perfect duality relations between the primal problem and the canonical dual problem. The study shows that some optimization problems, even if they are nonconvex and nonsmooth, can still be efficiently solved via the canonical duality approach.

The contributions of the canonical duality theory to mathematical optimization are manifested: insightful relations have been revealed, which might not otherwise be observable, and unified efficient solution methods can be developed for a wide range of problems. The relations revealed by the triality theory are also true for many other global optimization problems under certain conditions (see [41, 43, 44, 45, 52, 49, 55]). They lead to construction of unified solution methods, which start from computing a solution for the dual problem in the positive semidefinite region in the dual space. For any nonconvex problem, if the positive semidefinite region is not empty, the corresponding convex problem in the dual space always has a solution. Then if the solution is a critical point, it will correspond to a global solution of the primal problem; otherwise, its function value provides a lower bound for the primal problem. If the latter happens, by the new results on the complementary-dual principle, there must be a critical point outside of the positive semidefinite region which provides a global solution to the primal problem. Approximation and heuristic methods can then be employed to search for desired critical points.

Appendix A

Linear algebra

A brief review of some basic concepts from linear algebra and matrix calculus is given in this appendix. They form a part of preliminary knowledge for the discussion in the previous chapters.

A.1 Column space, nullspace and rank

Let $A \in \mathbb{R}^{m \times n}$. The *column space* of A, defined as

$$\mathcal{C}(A) = \{A\boldsymbol{x} \,|\, \boldsymbol{x} \in \mathbb{R}^n\},$$

consists of all linear combinations of the columns. The *nullspace* of A, defined as

$$\mathcal{N}(A) = \{\boldsymbol{x} \,|\, A\boldsymbol{x} = 0\},$$

consists of all solutions to the system of linear equations $A\boldsymbol{x} = 0$. The column space is a subspace of $\mathbb{R}^m$, while the nullspace is a subspace of $\mathbb{R}^n$.

The dimension of the column space is equal to the *rank* of A, which is denoted by $\text{rank}(A)$. It is true that $\text{rank}(A) = \text{rank}(A^T)$ and the rank can never be greater than the minimum of m and n. If $\text{rank}(A) = k$, there exist k linearly independent columns of A, which form a basis of the subspace $\mathcal{C}(A)$, and k linearly independent rows of A, which indicate that the dimension of $\mathcal{N}(A)$ is equal to $n-k$. If $\text{rank}(A) = \min\{m, n\}$, the matrix A is said to have *full rank*. For a square matrix, i.e., $m = n$, it is also said to be *nonsingular* if it has full rank; otherwise, it is *singular*. A nonsingular matrix is *invertible*.

A.2 Orthogonality

A set of vectors $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ in $\mathbb{R}^n$ is *orthogonal* if $\boldsymbol{x}_i^T \boldsymbol{x}_j = 0$ whenever $i \neq j$. A collection of subspaces $S_1, \ldots, S_k$ in $\mathbb{R}^n$ is *mutually orthogonal* if $\boldsymbol{x}^T \boldsymbol{y} = 0$ whenever $\boldsymbol{x} \in S_i$ and $\boldsymbol{y} \in S_j$ for $i \neq j$. For a subspace $S \subseteq \mathbb{R}^n$, the set

$$S^\perp = \{\boldsymbol{y} \in \mathbb{R}^n \,|\, \boldsymbol{y}^T \boldsymbol{x} = 0 \text{ for all } \boldsymbol{x} \in S\}$$

is called the *orthogonal complement* of S. It can be shown that $\mathcal{C}(A)^\perp = \mathcal{N}(A^T)$.

A matrix $Q \in \mathbb{R}^{m\times m}$ is said to be *orthogonal* if $Q^TQ = I$, i.e., columns of A are unit vectors and form an orthogonal set of vectors. The 2-norm is invariant under orthogonal transformation, for $\|Q\boldsymbol{x}\|_2^2 = \boldsymbol{x}^TQ^TQ\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{x} = \|\boldsymbol{x}\|_2^2$ if Q is orthogonal.

A.3 Eigenvalues and eigenvectors

Let $A \in \mathbb{R}^{n\times n}$. If a scalar λ and a nonzero vector $\boldsymbol{x} \in \mathbb{R}^n$ happen to satisfy the equation

$$A\boldsymbol{x} = \lambda\boldsymbol{x},$$

then λ is called an *eigenvalue* of A and $\boldsymbol{x}$ is called an *eigenvector* of A associated with λ. Notice that any vector of the form $\alpha\boldsymbol{x}$ with $\alpha \in \mathbb{R}$ and $\alpha \neq 0$ satisfy the equation with λ and is an eigenvector of A associated with λ. Thus, for each eigenvalue, there are infinite number of associated eigenvectors.

The set of all eigenvalues of A is called the *spectrum* of A, and it consists of all roots of the so called *characteristic polynomial* $\det(tI - A)$. From the fundamental theorem of algebra, the matrix A has n eigenvalues, but not necessary all real numbers.

The determinant and trace can be expressed in terms of the eigenvalues,

$$\mathrm{tr}(A) = \sum_{i=1}^{n} \lambda_i, \text{ and } \det(A) = \prod_{i=1}^{n} \lambda_i.$$

It is clear that A is nonsingular if and only if all eigenvalues are nonzero.

A.4 Symmetric matrices and eigenvalue decomposition

A matrix $A \in \mathbb{R}^{n\times n}$ is said to be *symmetric* if $A = A^T$. We use $\mathbb{S}^n$ to denote the set of all symmetric matrices in $\mathbb{R}^{n\times n}$. A very important property about symmetric matrices is that all the n eigenvalues are real numbers and the *eigenvalue decomposition* always exists. If $A \in \mathbb{S}^n$, then A can be factored as

$$A = U\Lambda U^T,$$

where $U \in \mathbb{R}^{n\times n}$ is orthogonal and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. The real values $\lambda_1, \ldots, \lambda_n$ are the *eigenvalues* of A, and the columns of U are the corresponding eigenvectors. The eigenvalue decomposition is also referred to as *spectral decomposition*. The decomposition can also be expressed as

$$A = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$$

where $\boldsymbol{u}_i$ are columns of U.

A matrix $A \in \mathbb{S}^n$ is *positive definite*, denoted as $A \succ 0$, if $\boldsymbol{x}^T A\boldsymbol{x} > 0$ for all nonzero $\boldsymbol{x} \in \mathbb{R}^n$. By the eigenvalue decomposition, we have

$$\boldsymbol{x}^T A\boldsymbol{x} = \sum_{i=1}^{n} \lambda_i (\boldsymbol{u}_i^T \boldsymbol{x})^2.$$

Thus, A is positive definite if and only if all its eigenvalues are positive. If $-A$ is positive definite, we say A is *negative definite*, denoted as $A \prec 0$. We use $\mathbb{S}^n_{++}$ to denote the set of all positive definite matrices in $\mathbb{S}^n$.

If the strict inequality is weakened to $\boldsymbol{x}^T A\boldsymbol{x} \geq 0$, then A is said to be *positive semidefinite*, denoted as $A \succeq 0$. If $-A$ is positive semidefinite, A is called *negative semidefinite*, denoted as $A \preceq 0$. We use $\mathbb{S}^n_{+}$ to denote the set of all positive semidefinite matrices in $\mathbb{S}^n$.

If $A \in \mathbb{R}^{n\times n}$ is symmetric and positive definite, then it can be factored as

$$A = LL^T$$

where L is lower triangular and nonsingular with positive diagonal entries. This is called the *Cholesky factorization* of A, and the matrix L is called the *Cholesky factor*, which is uniquely determined by A.

A.5 Singular value decomposition

Suppose $A \in \mathbb{R}^{m\times n}$ has rank k. Then A may be written in the form

$$A = U\Sigma V^T \tag{A.1}$$

where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices, i.e., $U^TU = I$ and $V^TV = I$. The matrix $\Sigma = \{\sigma_{ij}\} \in \mathbb{R}^{m\times n}$ has $\sigma_{ij} = 0$ for $i \neq j$ and

$$\sigma_{11} \geq \sigma_{22} \geq \cdots \geq \sigma_{kk} > \sigma_{k+1,k+1} = \cdots = \sigma_{qq} = 0,$$

where $q = \min\{m, n\}$. The entries $\sigma_{ii}, i = 1, \ldots, q$, simply denoted as σ_i, are known as the *singular values* of A, the columns of U are the *left singular vectors*, and the columns of V are the *right singular vectors*. The factorization is called the *singular value decomposition* (SVD) of A. The equation (A.1) can also be written

$$A = \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$$

from which we see clearly that $\sigma_i, i = 1, \ldots, q$ can always be nonnegative for arbitrary matrices.

Notice that the singular value decomposition could be seen as a generalization to arbitirary matrices of the eigenvalue decomposition of symmetric matrices. Actually, it is closely related to the eigenvalue decomposition. We have

$$AA^T = U\Lambda_1 U^T \tag{A.2}$$

$$A^T A = V\Lambda_2 V^T, \tag{A.3}$$

with $\Lambda_1 = \Sigma\Sigma^T \in \mathbb{R}^{m\times m}$ and $\Lambda_2 = \Sigma^T\Sigma \in \mathbb{R}^{n\times n}$. It is obvious that equations (A.2) and (A.3) are eigenvalue decompositions: both AA^T and A^TA are positive semidefinite and have nonzero eigenvalues $\sigma_1^2, \ldots, \sigma_k^2$; the columns of U are eigenvectors of AA^T and the columns of V are eigenvectors of A^TA. So positive singular values of A are positive square roots of eigenvalues of AA^T or A^TA. If A is symmetric, the singular values are the absolute values of its eigenvalues.

A.6 Moore-Penrose pseudo-inverse

For $A \in \mathbb{R}^{m\times n}$, the Moore-Penrose pseudo-inverse or simply *pseudo-inverse* of A is defined as a matrix $A^\dagger \in \mathbb{R}^{n\times m}$ satisfying all of the following equations:

1. $(AA^\dagger)^T = AA^\dagger$
2. $(A^\dagger A)^T = A^\dagger A$
3. $A^\dagger AA^\dagger = A^\dagger$
4. $AA^\dagger A = A$

If A has singular value decomposition $A = U\Sigma V$, its pseudo-inverse can be expressed as

$$A^\dagger = V\Sigma^\dagger U^T,$$

where $\Sigma^\dagger$ is the transpose of Σ with all positive singular values being replaced by their reciprocals. The pseudo-inverse exists for any matrix, even for a singular square matrix and for a nonsquare matrix.

The pseudo-inverse comes up in many problems. Here, we discuss the application in solving a system of linear equations,

$$A\boldsymbol{x} = \boldsymbol{b}$$

with arbitrary A and b. The system may have no solution, unique solution, or infinite solutions. If there is no solution, the vector $\boldsymbol{x}^* = A^\dagger\boldsymbol{b}$ gives a least-squares solution, i.e., $\boldsymbol{x}^*$ is a solution of the least-squares problem

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

It is also the projection of the vector $\boldsymbol{b}$ in the column space of A. If there are infinite solutions, all the solutions can be expressed by

$$\boldsymbol{x} = A^{\dagger}\boldsymbol{b} + (I - A^{\dagger}A)\boldsymbol{w}$$

with $\boldsymbol{w} \in \mathbb{R}^n$, and thus $\boldsymbol{x}^* = A^{\dagger}\boldsymbol{b}$ is a solution with the minimal norm to the linear system.

A.7 Schur lemma

Let $X \in \mathbb{S}^n$ be in the block form

$$X = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix},$$

and assume $\det(A) \neq 0$. The matrix

$$S = C - B^T A^{-1} B$$

is called the *Schur complement* of A in X. Schur complement is a key tool in matrix analysis.

The determinant of X can be written in the formula

$$\det(X) = \det(A)\det(S),$$

which generalizes the familiar formula for the determinant of a 2×2 matrix. This formula can be verified by the fact that X can be expressed as

$$X = \begin{bmatrix} I & 0 \\ -B^T A^{-1} & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & C - B^T A^{-1} B \end{bmatrix} \begin{bmatrix} I & -A^{-1}B \\ 0 & I \end{bmatrix}. \tag{A.4}$$

From this fact, we also have the following conditions for positive definiteness of X.

- $X \succ 0$ if and only if $A \succ 0$ and $S \succ 0$.
- If $A \succ 0$, then $X \succeq 0$ if and only if $S \succeq 0$.

These results can be generalized into the situation when A is singular. Replacing the inverse with the pseudo-inverse of A, we still have the equation (A.4) if $(I - AA^{\dagger})B = 0$. Thus, we have

- $X \succeq 0$ if and only if $A \succeq 0$, $(I - AA^{\dagger})B = 0$ and $C - B^T A^{-1} B \succeq 0$.
- If $A \succeq 0$, then $X \succeq 0$ if and only if $(I - AA^{\dagger})B = 0$ and $C - B^T A^{-1} B \succeq 0$.

The following lemma, which generalizes the Lemma 6 in [56], plays a key role in the proof of Triality Theorem.

Lemma 44 *Assume that $P \in \mathbb{S}^n_{++}$, $U \in \mathbb{S}^m_{++}$ and $D \in \mathbb{R}^{n\times m}$. Then, we have*

$$-P - DUD^T \succeq 0 \Longleftrightarrow U^{-1} + D^T P^{-1} D \succeq 0. \tag{A.5}$$

Proof: From the assumption, we know that $-P \succ 0$ and $U^{-1} \succ 0$. By the results above, the statement $-P - DUD^T \succeq 0$ is equivalent to

$$\begin{bmatrix} U^{-1} & D^T \\ D & -P \end{bmatrix} \succeq 0,$$

which is also equivalent to $U^{-1} - D^T(-P)^{-1}D = U^{-1} + D^T P^{-1} D \succeq 0$. Thus, the lemma is proved. □

A.8 Inverse of the sum of matrices and inverse of block matrix

The two identities are first presented,

$$(I+P)^{-1} = I - (I+P)^{-1}P, \text{ and } (I+PQ)^{-1}P = P(I+QP)^{-1},$$

where $I+P$ and $I+PQ$ are nonsingular, and they will play key roles in deriving formulas for the inverse of the sum of matrices. Suppose that $A+BCD$ is invertible with A being nonsingular and B, C and D being general matrices. By repeatedly using the two identities, we then have

$$\begin{aligned}
(A+BCD)^{-1} &= (A(I+A^{-1}BCD))^{-1} \\
&= A^{-1} - (I+A^{-1}BCD)^{-1}A^{-1}BCDA^{-1} \\
&= A^{-1} - A^{-1}(I+BCDA^{-1})^{-1}BCDA^{-1} \\
&= A^{-1} - A^{-1}B(I+CDA^{-1}B)^{-1}CDA^{-1} \\
&= A^{-1} - A^{-1}BC(I+DA^{-1}BC)^{-1}DA^{-1} \\
&= A^{-1} - A^{-1}BCD(I+A^{-1}BCD)^{-1}A^{-1} \\
&= A^{-1} - A^{-1}BCDA^{-1}(I+BCDA^{-1})^{-1}
\end{aligned}$$

Suppose that X is nonsingular and is partitioned into the block form

$$X = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

in which A and D are square matrices. Obviously, A and D are also nonsingular. Then we have the following formulas

$$\begin{aligned}
X^{-1} &= \begin{bmatrix} (A-BD^{-1}C)^{-1} & -A^{-1}B(D-CA^{-1}B)^{-1} \\ -D^{-1}C(A-BD^{-1}C)^{-1} & (D-CA^{-1}B)^{-1} \end{bmatrix} \\
&= \begin{bmatrix} (A-BD^{-1}C)^{-1} & -(A-BD^{-1}C)^{-1}BD^{-1} \\ -(D-CA^{-1}B)^{-1}CA^{-1} & (D-CA^{-1}B)^{-1} \end{bmatrix}
\end{aligned}$$

If X is symmetric, i.e., $A^T = A$, $D^T = D$ and $C = B^T$, then the inverse formula can be written as

$$X^{-1} = \begin{bmatrix} A^{-1} + A^{-1}BS^{-1}B^TA^{-1} & -A^{-1}BS^{-1} \\ -S^{-1}B^TA^{-1} & S^{-1} \end{bmatrix},$$

where S is the Schur complement of A in X.

Appendix B

Matrix differentiation

In this section, we list some useful formulas of matrix differentiation that appear in the previous chapters.

B.1 Derivatives with vectors

Let $\boldsymbol{x} = \{x_i\} \in \mathbb{R}^n$, and let

$$\boldsymbol{f}(\boldsymbol{x}) = \{f_i(\boldsymbol{x})\} : \mathbb{R}^n \to \mathbb{R}^m,$$

where each component $f_i(\boldsymbol{x})$ is a scalar function of $\boldsymbol{x}$. We then present definitions and notations of differentiation with vectors.

Derivatives

If $m = 1$, $\boldsymbol{f}(\boldsymbol{x})$ reduces to a scalar, which is denoted by $f(\boldsymbol{x})$. We are familar with its derivative, which is also referred to as the gradient,

$$\nabla f(\boldsymbol{x}) = \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \begin{pmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{x_n} \end{pmatrix}.$$

While if $n = 1$, the derivative is written in the row form:

$$\nabla \boldsymbol{f}(x) = \frac{\partial \boldsymbol{f}}{\partial x} = (f_1'(x), \cdots, f_n'(x)).$$

For general cases, the derivative of $\boldsymbol{f}(\boldsymbol{x})$ with respect to $\boldsymbol{x}$ is the $n \times m$ matrix:

$$\nabla \boldsymbol{f}(\boldsymbol{x}) = \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_n(\boldsymbol{x})}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} & \cdots & \frac{\partial f_n(\boldsymbol{x})}{\partial x_n} \end{bmatrix}.$$

If the function $\boldsymbol{f}(\boldsymbol{x})$ is twice differentiable, we can further take derivative of $\nabla \boldsymbol{f}(\boldsymbol{x})$. Here we only give the notation for the case when $m = 1$. The second derivative of $f(\boldsymbol{x})$ is also called the Hessian matrix, defined by

$$\nabla^2 f(\boldsymbol{x}) = \frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x}^2} = \frac{\partial}{\partial \boldsymbol{x}} \left(\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} \right).$$

Jacobian matrix

The matrix $\nabla \boldsymbol{f}(\boldsymbol{x})$ is called the *Jacobian matrix* of the vector function $\boldsymbol{f}(\boldsymbol{x})$. In the case the Jacobian matrix is square, i.e., $n = m$, its determinant,

$$J = \det(\nabla \boldsymbol{f}(\boldsymbol{x})) = \begin{vmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_n(\boldsymbol{x})}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} & \cdots & \frac{\partial f_n(\boldsymbol{x})}{\partial x_n} \end{vmatrix}$$

is called the *Jacobian determinant* (or simply, the *Jacobian*) of the function.

The chain rule

Let $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}^m$ and $\boldsymbol{z} = \boldsymbol{z}(\boldsymbol{y}) : \mathbb{R}^m \to \mathbb{R}^l$. By the chain rule for scalar functions, each entry of the matrix $\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}$ may be expanded as

$$\frac{\partial z_i}{\partial x_j} = \sum_{k=1}^{m} \frac{\partial z_i}{\partial y_k} \frac{\partial y_k}{\partial x_j} = \frac{\partial \boldsymbol{y}}{\partial x_j} \frac{\partial z_i}{\partial \boldsymbol{y}}.$$

Then, we have the identity

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{y}}, \tag{B.1}$$

which is the chain rule for vector functions. Furthermore, if $\boldsymbol{w}$ is a vector function of $\boldsymbol{z}$, thus a function of $\boldsymbol{x}$, then the derivative of $\boldsymbol{w}$ with respect to $\boldsymbol{x}$ can be expressed as

$$\frac{\partial \boldsymbol{w}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{y}} \frac{\partial \boldsymbol{w}}{\partial \boldsymbol{z}}.$$

Notice that we build the product of matrices to the left, in comparison with the conventional chain rule of calculus where one builds the chain to the right. For example, if here all vectors reduce to scalars, the identity (B.1) becomes

$$\frac{\partial z}{\partial x} = \frac{\partial y}{\partial x} \frac{\partial z}{\partial y} = \frac{\partial z}{\partial y} \frac{\partial y}{\partial x},$$

which is the conventional chain rule of calculus.

Implicit function theorem

The implicit function theorem is fundamentally important to the discussion in Chapter 3. For a proof, see [129].

Let the functions $F_i(y_1, \ldots, y_m, x_1, \ldots, x_n), i = 1, \ldots, m$ all be defined in a neighborhood of the point $P_0 : (y_1^0, \ldots, y_m^0, x_1^0, \ldots, x_n^0)$ and have continuous first partial derivatives in this neighborhood. Let the equations

$$F_i(y_1, \ldots, y_m, x_1, \ldots, x_n) = 0, i = 1, \ldots, m,$$

be satisfied at P_0. If at P_0 the Jacobian is not equal to zero, i.e.,

$$\frac{\partial(F_1, \ldots, F_m)}{\partial(y_1, \ldots, y_m)} \neq 0,$$

then in an appropriate neighborhood of $(x_1^0, \ldots, x_n^0)$, there is a unique set of continuous functions

$$y_i = f_i(x_1, \ldots, x_n), \ i = 1, \ldots, m,$$

such that $y_i^0 = f_i(x_1^0, \ldots, x_n^0)$ for $i = 1, \ldots, m$ and for all i

$$F_i(f_1(x_1, \ldots, x_n), \ldots, f_m(x_1, \ldots, x_n), x_1, \ldots, x_n) = 0$$

in the neighborhood. Furthermore, the f_i have continuous partial derivatives satisfying

$$\frac{\partial y_i}{\partial x_j} = \frac{\frac{\partial(F_1, \ldots, F_m)}{\partial(y_1, \ldots, y_{i-1}, x_j, y_{i+1}, \ldots, y_m)}}{\frac{\partial(F_1, \ldots, F_m)}{\partial(y_1, \ldots, y_m)}}, i = 1, \ldots, m, j = 1, \ldots, n.$$

B.2 Derivatives with matrices

Let $X = \{x_{ij}\} \in \mathbb{R}^{m \times n}$, and let

$$Y = F(X) : \mathbb{R}^{m \times n} \to \mathbb{R}^{p \times q},$$

where $Y = y_{ij}$ and each entry y_{ij} is a function of X. If X is a scalar, denoted by x, the derivative of the matrix function Y of x is given by

$$\frac{\partial Y}{\partial x} = \begin{bmatrix} \frac{\partial y_{11}}{\partial x} & \cdots & \frac{\partial y_{1q}}{\partial x} \\ \vdots & & \vdots \\ \frac{\partial y_{p1}}{\partial x} & \cdots & \frac{\partial y_{pq}}{\partial x} \end{bmatrix},$$

which is known as the tangent matrix. While if Y is a scalar, denoted by y, the derivative is given by

$$\frac{\partial y}{\partial X} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \cdots & \frac{\partial y}{\partial x_{1n}} \\ \vdots & & \vdots \\ \frac{\partial y}{\partial x_{m1}} & \cdots & \frac{\partial y}{\partial x_{mn}} \end{bmatrix}.$$

Derivatives of matrix trace, determinant and inverse

Let

$$y = \mathrm{tr}(X) = \sum_{i=1}^{n} x_{ii}.$$

Obviously, all non-diagonal entries of the derivative vanish whereas the diagonal entries equal one, thus

$$\frac{\partial y}{\partial X} = I.$$

Let $Y = F(X)$ be a matrix-valued function of the matrix X. We want to find the derivative of the determinant of Y with respect to X, i.e.,

$$\frac{\partial \mathrm{det}(Y)}{\partial X}.$$

The chain rule gives

$$\frac{\partial \mathrm{det}(Y)}{\partial x_{ij}} = \sum_k \sum_l \frac{\partial \mathrm{det}(Y)}{\partial y_{kl}} \frac{\partial y_{kl}}{\partial x_{ij}}.$$

From the expression of the determinant

$$\mathrm{det}(Y) = \sum_l y_{kl} C_{kl},$$

where C_{kl} is the cofactor of the entry y_{kl}, we have

$$\frac{\partial \mathrm{det}(Y)}{\partial y_{kl}} = C_{kl}.$$

It then follows that

$$\frac{\partial \mathrm{det}(Y)}{\partial x_{ij}} = \sum_k \sum_l C_{kl} \frac{\partial y_{kl}}{\partial x_{ij}} = \mathrm{det}(Y)\mathrm{tr}(Y^{-1} \frac{\partial Y}{\partial x_{ij}}).$$

We then want to find the derivative of the inverse of Y with respect to X, i.e.,

$$\frac{\partial Y^{-1}}{\partial X}.$$

By the aid of the identity $Y^{-1}Y = I$, we have

$$\frac{\partial Y^{-1}}{\partial x_{ij}} Y + Y^{-1} \frac{\partial Y}{\partial x_{ij}} = 0,$$

from which the derivative $\partial Y^{-1}/\partial x_{ij}$ is expressed as

$$\frac{\partial Y^{-1}}{\partial x_{ij}} = Y^{-1} \frac{\partial Y}{\partial x_{ij}} Y^{-1}.$$

Bibliography

[1] H. Abdi and S. Nahavandi. Designing optimal fault tolerant jacobian for robotic manipulators. In *Advanced Intelligent Mechatronics (AIM), 2010 IEEE/ASME International Conference on*, pages 426–431. IEEE, 2010.

[2] H. Abdi, S. Nahavandi, and A. A. Maciejewski. Optimal fault-tolerant jacobian matrix generators for redundant manipulators. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 4688–4693. IEEE, 2011.

[3] W. P. Adams and H. D. Sherali. A tight linearization and an algorithm for zero-one quadratic programming problems. *Manage. Sci.*, 32(10):1274–1290, 1986.

[4] K. Allemand, K. Fukuda, T. M. Liebling, and E. Steiner. A polynomial case of unconstrained zero-one quadratic optimization. *Math. Program.*, 91(1):49–52, 2001.

[5] E. Balas, S. Ceria, and G. Cornuéjols. Mixed 0-1 programming by lift-and-project in a branch-and-cut framework. *Manage. Sci.*, 42(9):1229–1246, 1996.

[6] E. Balas, S. Ceria, G. Cornuéjols, and N. Natraj. Gomory cuts revisited. *Oper. Res. Lett.*, 19(1):1–9, 1996.

[7] N. V. Banichuk. Minimax approach to structural optimization problems. *J. Optimiz. Theory App.*, 20(1):111–127, 1976.

[8] F. Barahona. The max-cut problem on graphs not contractible to K5. *Oper. Res. Lett.*, 2(3):107–111, 1983.

[9] F. Barahona, M. Jünger, and G. Reinelt. Experiments in quadratic 0–1 programming. *Math. Program.*, 44(1-3):127–137, 1989.

[10] F. Barahona and L. Ladanyi. Branch and cut based on the volume algorithm: Steiner trees in graphs and max-cut. *RAIRO-Oper. Res.*, 40(1):53–73, 2006.

[11] F. Barahona and A. R. Mahjoub. On the cut polytope. *Math. Program.*, 36(2):157–173, 1986.

[12] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty. *Nonlinear programming: theory and algorithms*. John Wiley & Sons, 2013.

[13] A. Beck and M. Teboulle. Global optimality conditions for quadratic optimization problems with binary constraints. *SIAM J. Optimiz.*, 11(1):179–188, 2000.

[14] W. Ben-Ameur and J. Neto. Spectral bounds for unconstrained (-1,1)-quadratic optimization problems. *Eur. J. Oper. Res.*, 207(1):15–24, 2010.

[15] A. Ben-Tal and M. Teboulle. Hidden convexity in some nonconvex quadratically constrained quadratic programming. *Math. Program.*, 72(1):51–63, 1996.

[16] A. Billionnet and S. Elloumi. Using a mixed integer quadratic programming solver for the unconstrained quadratic 0-1 problem. *Math. Program.*, 109(1):55–68, 2007.

[17] I. M. Bomze, M. Budinich, P. M. Pardalos, and M. Pelillo. The maximum clique problem. In *Handbook of combinatorial optimization*, pages 1–74. Springer, 1999.

[18] E. Boros, Y. Crama, and P. L. Hammer. Chvatal cuts and odd cycle inequalities in quadratic 0-1 optimization. *SIAM J. Discrete Math.*, 5(2):163–177, 1992.

[19] E. Boros and P. L. Hammer. Pseudo-boolean optimization. *Discrete Appl. Math.*, 123(1):155–225, 2002.

[20] S. P. Boyd, S. J. Kim, L. Vandenberghe, and A. Hassibi. A tutorial on geometric programming. *Optim. Eng.*, 8(1):67–127, 2007.

[21] S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[22] M. W. Carter. The indefinite zero-one quadratic problem. *Discrete Appl. Math.*, 7(1):23–44, 1984.

[23] Y. Chen and D. Y. Gao. Global solutions to nonconvex optimization of 4th-order polynomial and log-sum-exp functions. *J. Global Optim.*, pages 1–15, 2014.

[24] Y. Chen and D. Y. Gao. Canonical dual approach for minimizing a nonconvex quadratic function over a sphere. In *Advances in Global Optimization*, pages 149–156. Springer, 2015.

[25] Y. Chen and D. Y. Gao. Global solutions to spherically constrained quadratic minimization via canonical duality theory. *Math. Mech. Solids*, 2015.

[26] M. Chiang. *Geometric programming for communication systems*. Now Publishers Inc, 2005.

[27] M. Chiang and S. Boyd. Geometric programming duals of channel capacity and rate distortion. *IEEE T. Inform. Theory*, 50(2):245–258, 2004.

[28] M. Chiang, C. W. Tan, D. P. Palomar, D. O'Neill, and D. Julian. Power control by geometric programming. *IEEE T. Wirel. Commun.*, 6(7):2640–2651, 2007.

[29] A. R. Conn, N. I. M. Gould, and P. L. Toint. *Trust-region methods*. SIAM, Philadelphia, PA, 2000.

[30] Y. Crama, P. Hansen, and B. Jaumard. The basic algorithm for pseudo-boolean programming revisited. *Discrete Appl. Math.*, 29(2):171–185, 1990.

[31] C. Delorme and S. Poljak. Laplacian eigenvalues and the maximum cut problem. *Math. Program.*, 62(1-3):557–574, 1993.

[32] S. C. Fang, D. Y. Gao, R. L. Sheu, and S. Y. Wu. Canonical dual approach to solving 0-1 quadratic programming problems. *J. Ind. Manag. Optim.*, 4(1):125–142, 2008.

[33] C. Fortin and H. Wolkowicz. The trust region subproblem and semidefinite programming. *Optim. Method Softw.*, 19(1):41–67, 2004.

[34] D. Y. Gao. On the complementary bounding theorems for limit analysis. *Int. J. Solids Struct.*, 24(6):545–556, 1988.

[35] D. Y. Gao. Panpenalty finite element programming for limit analysis. *Comput. Struct.*, 28(6):749–755, 1988.

[36] D. Y. Gao. Minimax and triality theory in nonsmooth variational problems. In *Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods*, pages 161–179. Springer, 1999.

[37] D. Y. Gao. Canonical dual transformation method and generalized triality theory in nonsmooth global optimization. *J. Global Optim.*, 17(1/4):127–160, 2000.

[38] D. Y. Gao. *Duality principles in nonconvex systems: theory, methods, and applications*. Springer Netherlands, 2000.

[39] D. Y. Gao. Finite deformation beam models and triality theory in dynamical post-buckling analysis. *Int. J. Nonlin. Mech.*, 35(1):103–131, 2000.

[40] D. Y. Gao. Nonconvex semi-linear problems and canonical duality solutions. In *Advances in mechanics and mathematics*, pages 261–312. Springer, 2003.

[41] D. Y. Gao. Perfect duality theory and complete solutions to a class of global optimization problems. *Optimization*, 52(4-5):467–493, 2003.

[42] D. Y. Gao. Canonical duality theory and solutions to constrained nonconvex quadratic programming. *J. Global Optim.*, 29(4):377–399, 2004.

[43] D. Y. Gao. Sufficient conditions and perfect duality in nonconvex minimization with inequality constraints. *J. Ind. Manag. Optim.*, 1(1):53–63, 2005.

[44] D. Y. Gao. Complete solutions and extremality criteria to polynomial optimization problems. *J. Global Optim.*, 35(1):131–143, 2006.

[45] D. Y. Gao. Solutions and optimality criteria to box constrained nonconvex minimization problems. *J. Ind. Manag. Optim.*, 3(2):293–304, 2007.

[46] D. Y. Gao. Canonical duality theory: unified understanding and generalized solution for global optimization problems. *Comput. Chem. Eng.*, 33(12):1964–1972, 2009.

[47] D. Y. Gao and N. Ruan. Solutions to quadratic minimization problems with box and integer constraints. *J. Global Optim.*, 47(3):463–484, 2010.

[48] D. Y. Gao, N. Ruan, and V. Latorre. Canonical duality-triality theory: bridge between nonconvex analysis/mechanics and global optimization in complex systems. *Math. Mech. Solids*, 2015.

[49] D. Y. Gao, N. Ruan, and P. M. Pardalos. Canonical dual solutions to sum of fourth-order polynomials minimization problems with applications to sensor network localization. In P. M. Pardalos, Y. Y. Ye, V. Boginski, and C. Commander, editors, *Sensors: Theory, Algorithms, and Applications*, pages 37–54. Springer, 2010.

[50] D. Y. Gao, N. Ruan, and P. M. Pardalos. Canonical dual solutions to sum of fourth-order polynomials minimization problems with applications to sensor network localization. In *Sensors: Theory, Algorithms, and Applications*, pages 37–54. Springer, 2012.

[51] D. Y. Gao, N. Ruan, and H. D. Sherali. Solutions and optimality criteria for nonconvex constrained global optimization problems with connections between canonical and Lagrangian duality. *J. Global Optim.*, 45(3):473–497, 2009.

[52] D. Y. Gao, N. Ruan, and H. D. Sherali. Canonical dual solutions for fixed cost quadratic programs. In A. Chinchuluun, P. M. Pardalos, R. Enkhbat, and I. Tseveendorj, editors, *Optimization and Optimal Control*, pages 139–156. Springer, 2010.

[53] D. Y. Gao and H. D. Sherali. Canonical duality theory: Connection between nonconvex mechanics and global optimization. In D. Y. Gao and H. D. Sherali, editors, *Advances in Applied Mathematics and Global Optimization*, pages 249–316. Springer, 2009.

[54] D. Y. Gao and G. Strang. Geometric nonlinearity: Potential energy, complementary energy, and the gap function. *Q. Appl. Math*, 47(3):487–504, 1989.

[55] D. Y. Gao, L. T. Watson, D. R. Easterling, W. I. Thacker, and S. C. Billups. Solving the canonical dual of box- and integer-constrained nonconvex quadratic programs via a deterministic direct search algorithm. *Optim. Method. Softw.*, 26(1):1–14, 2011.

[56] D. Y. Gao and C. Wu. On the triality theory for a quartic polynomial optimization problem. *J. Ind. Manag. Optim.*, 8:229–242, 2012.

[57] D. Y. Gao and H. Yu. Multi-scale modelling and canonical dual finite element method in phase transitions of solids. *Int. J. Solids and Struct.*, 45(13):3660–3673, 2008.

[58] Y. Gao and G. Strang. Geometric nonlinearity: potential energy, complementary energy, and the gap function. *Quart. Appl. Math.*, 47(3):487–504, 1989.

[59] D. M. Gay. Computing optimal locally constrained steps. *SIAM J. Sci. Stat. Comp.*, 2(2):186–197, 1981.

[60] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42(6):1115–1145, 1995.

[61] N. I. M. Gould, S. Lucidi, M. Roma, and P. L. Toint. Solving the trust-region subproblem using the lanczos method. *SIAM J. Optimiz.*, 9(2):504–525, 1999.

[62] M. Grötschel, L. Lovász, and A. Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, 1981.

[63] S. Gueye and P. Michelon. A linearization framework for unconstrained quadratic (0-1) problems. *Discrete Appl. Math.*, 157(6):1255–1266, 2009.

[64] F. Hadlock. Finding a maximum cut of a planar graph in polynomial time. *SIAM J. Comput.*, 4(3):221–225, 1975.

[65] W. W. Hager. Minimizing a quadratic over a sphere. *SIAM J. Optimiz.*, 12(1):188–208, 2001.

[66] P. L. Hammer, P. Hansen, and B. Simeone. Roof duality, complementation and persistency in quadratic 0–1 optimization. *Math. Program.*, 28(2):121–155, 1984.

[67] C. Helmberg. *Semidefinite programming for combinatorial optimization*. Konrad-Zuse-Zentrum für Informationstechnik Berlin, 2000.

[68] C. Helmberg and F. Rendl. Solving quadratic (0, 1)-problems by semidefinite programs and cutting planes. *Math. Program.*, 82(3):291–315, 1998.

[69] C. Helmberg, F. Rendl, R. J. Vanderbei, and H. Wolkowicz. An interior-point method for semidefinite programming. *SIAM J. Optimiz.*, 6(2):342–361, 1996.

[70] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, 1990.

[71] A. Jaffe. Constructive quantum field theory. In *Mathematical physics 2000*, pages 111–127. Imp. Coll. Press, London, 2000.

[72] Q. Jin, S.-C. Fang, and W. Xing. On the global optimality of generalized trust region subproblems. *Optimization*, 59(8):1139–1151, 2010.

[73] N. Jorge and S. J. Wright. *Numerical optimization*, volume 2. Springer New York, 1999.

[74] R. M. Karp. *Reducibility among combinatorial problems*. Springer, 1972.

[75] T. Kibble. Phase transitions and topological defects in the early universe. *Aust. J. Phys.*, 50(4):697–722, 1997.

[76] S. Kim and M. Kojima. Second order cone programming relaxation of nonconvex quadratic optimization problems. *Optim. Method. Softw.*, 15(3-4):201–224, 2001.

[77] G. Kochenberger, J.-K. Hao, F. Glover, M. Lewis, Z. Lü, H. Wang, and Y. Wang. The unconstrained binary quadratic programming problem: a survey. *J. Comb. Optim.*, 28(1):58–81, 2014.

[78] G. A. Kochenberger, F. Glover, B. Alidaee, and C. Rego. An unconstrained quadratic binary programming approach to the vertex coloring problem. *Ann. Oper. Res.*, 139(1):229–241, 2005.

[79] G. A. Kochenberger, J.-K. Hao, Z. Lü, H. Wang, and F. Glover. Solving large scale max cut problems via tabu search. *J. Heuristics*, 19(4):565–571, 2013.

[80] L. D. Landau and E. M. Lifshitz. *Course of theoretical physics Vol 1: Mechanics*. Higher Education Press, Beijing China, 2007.

[81] V. Latorre and D. Y. Gao. Canonical duality for radial basis neural networks. In *Proceedings of The Eighth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA), 2013*, pages 1189–1197. Springer, 2013.

[82] V. Latorre and D. Y. Gao. Canonical dual solutions to nonconvex radial basis neural network optimization problem. *Neurocomputing*, 134:189–197, 2014.

[83] V. Latorre and D. Y. Gao. Canonical duality for solving general nonconvex constrained problems. *Optim. Lett.*, pages 1–17, 2015.

[84] D. Li, X. Sun, and C. L. Liu. An exact solution method for unconstrained quadratic 0–1 programming: a geometric approach. *J. Global Optim.*, 52(4):797–829, 2012.

[85] L. Lovász and A. Schrijver. Cones of matrices and set-functions and 0-1 optimization. *SIAM J. Optimiz.*, 1(2):166–190, 1991.

[86] U. Malik, I. M. Jaimoukha, G. D. Halikias, and S. K. Gungah. On the gap between the quadratic integer programming problem and its semidefinite relaxation. *Math. Program.*, 107(3):505–515, 2006.

[87] J. M. Martínez. Local minimizers of quadratic functions on euclidean balls and spheres. *SIAM J. Optimiz.*, 4(1):159–176, 1994.

[88] D. M. Morales Silva and D. Y. Gao. Canonical duality theory and triality for solving general nonconstrained global optimization problems. *arXiv preprint arXiv:1210.0180*, 2012.

[89] D. M. Morales Silva and D. Y. Gao. Complete solutions and triality theory to a nonconvex optimization problem with double-well potential in R^n. *Numer. Algebra Contr. Optim.*, 3(2):271–282, 2013.

[90] J. J. Moré and D. C. Sorensen. Computing a trust region step. *SIAM J. Sci. Stat. Comp.*, 4(3):553–572, 1983.

[91] M. Muramatsu and T. Suzuki. A new second-order cone programming relaxation for max-cut problems. *Journal of Operations Research of Japan*, 43:164–177, 2003.

[92] M. Padberg. The boolean quadric polytope: some characteristics, facets and relatives. *Math. Program.*, 45(1-3):139–172, 1989.

[93] M. Padberg and G. Rinaldi. A branch-and-cut algorithm for the resolution of large-scale symmetric traveling salesman problems. *SIAM Rev.*, 33(1):60–100, 1991.

[94] S. Pan, T. Tan, and Y. Jiang. A global continuation algorithm for solving binary quadratic programming problems. *Comput. Optim. Appl.*, 41(3):349–362, 2008.

[95] P. M. Pardalos and S. Jha. Graph separation techniques for quadratic zero-one programming. *Comput. Math. Appl.*, 21(6):107–113, 1991.

[96] P. M. Pardalos and S. Jha. Complexity of uniqueness and local search in quadratic 0–1 programming. *Oper. Res. Lett.*, 11(2):119–123, 1992.

[97] P. M. Pardalos and J. Xue. The maximum clique problem. *J. Global Optim.*, 4(3):301–328, 1994.

[98] E. Y. Pee and J. O. Royset. On solving large-scale finite minimax problems using exponential smoothing. *J. Optimiz. Theory App.*, 148(2):390–421, 2011.

[99] E. Polak. *Optimization: algorithms and consistent approximations*, volume 124. Springer Verlag, 1997.

[100] E. Polak, J. O. Royset, and R. S. Womersley. Algorithms with adaptive smoothing for finite minimax problems. *J. Optimiz. Theory App.*, 119(3):459–484, 2003.

[101] S. Poljak, F. Rendl, and H. Wolkowicz. A recipe for semidefinite relaxation for (0, 1)-quadratic programming. *J. Global Optim.*, 7(1):51–73, 1995.

[102] M. J. D. Powell. On trust region methods for unconstrained minimization without derivatives. *Math. Program.*, 97(3):605–623, 2003.

[103] F. Rendl, G. Rinaldi, and A. Wiegele. Solving max-cut to optimality by intersecting semidefinite and polyhedral relaxations. *Math. Program.*, 121(2):307–335, 2010.

[104] F. Rendl and H. Wolkowicz. A semidefinite framework for trust region subproblems with applications to large scale minimization. *Math. Program.*, 77(1):273–299, 1997.

[105] R. G. Roberts, H. G. Yu, and A. A. Maciejewski. Characterizing optimally fault-tolerant manipulators based on relative manipulability indices. In *Intelligent Robots and Systems, 2007 IEEE/RSJ International Conference on*, pages 3925–3930. IEEE, 2007.

[106] R. T. Rockafellar. *Convex analysis*. Princeton University Press, 1970.

[107] M. Rojas, S. A. Santos, and D. C. Sorensen. A new matrix-free algorithm for the large-scale trust-region subproblem. *SIAM J. Optimiz.*, 11(3):611–646, 2001.

[108] J. S. Rowlinson. Translation of J. D. van der Waals' 'The thermodynamik theory of capillarity under the hypothesis of a continuous variation of density'. *J. Stat. Phys.*, 20(2):197–200, 1979.

[109] J. O. Royset, E. Polak, and A. D. Kiureghian. Adaptive approximations and exact penalization for the solution of generalized semi-infinite min-max problems. *SIAM J. Optimiz.*, 14(1):1–34, 2004.

[110] N. Ruan. Complete solutions to mixed integer programming. *American Journal of Computational Mathematics*, 3(3B):27–30, 2013.

[111] N. Ruan. Solving facility location problem based on duality approach. In *Advances in Global Optimization*, pages 165–172. Springer, 2015.

[112] N. Ruan, Y. Chen, and D. Y. Gao. An efficient classification using support vector machines. In *Science and Information Conference (SAI) 2013*, pages 585–589. IEEE, 2013.

[113] N. Ruan and D. Gao. Global solutions to fractional programming problem with ratio of nonconvex functions. *Appl. Math. Comput.*, 255:66–72, 2014.

[114] N. Ruan and D. Y. Gao. Canonical duality theory and algorithm for solving challenging problems in network optimisation. In *Neural Information Processing*, pages 702–709. Springer, 2012.

[115] N. Ruan and D. Y. Gao. Global optimal solutions to nonconvex euclidean distance geometry problems. In *Proceedings of 20th International Symposium on Mathematical Theory of Networks and Systems (MTNS 2012)*, 2012.

[116] N. Ruan and D. Y. Gao. Canonical duality approach for non-linear dynamical systems. *IMA J. Appl. Math.*, 79(2):313–325, 2014.

[117] N. Ruan and D. Y. Gao. Global optimal solutions to general sensor network localization problem. *Perform. Evaluation*, 75-76:1–16, 2014.

[118] N. Ruan and D. Y. Gao. Application of canonical duality theory to fixed point problem. In *Advances in Global Optimization*, pages 157–163. Springer, 2015.

[119] N. Ruan, D. Y. Gao, and Y. Jia. Canonical dual least square method for solving general nonlinear systems of quadratic equations. *Comput. Optim. Appl.*, 47(2):335–347, 2010.

[120] D. C. Sorensen. Newton's method with a model trust region modification. *SIAM J. Numer. Anal.*, 19(2):409–426, 1982.

[121] D. C. Sorensen. Minimization of a large-scale quadratic functionsubject to a spherical constraint. *SIAM J. Optimiz.*, 7(1):141–161, 1997.

[122] R. J. Stern and H. Wolkowicz. Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations. *SIAM J. Optimiz.*, 5(2):286–313, 1995.

[123] G. Strang. A minimax problem in plasticity theory. In *Functional analysis methods in numerical analysis*, pages 319–333. Springer, 1979.

[124] P. D. Tao and L. T. H. An. A dc optimization algorithm for solving the trust-region subproblem. *SIAM J. Optimiz.*, 8(2):476–505, 1998.

[125] P. D. Tao and L. T. H. An. A dc optimization algorithm for solving the trust-region subproblem. *SIAM J. Optimiz.*, 8(2):476–505, 1998.

[126] D. Vandenbussche and G. L. Nemhauser. A branch-and-cut algorithm for nonconvex quadratic programs with box constraints. *Math. Program.*, 102(3):559–575, 2005.

[127] D. Vandenbussche and G. L. Nemhauser. A polyhedral study of nonconvex quadratic programs with box constraints. *Math. Program.*, 102(3):531–557, 2005.

[128] Z. B. Wang, S. C. Fang, D. Y. Gao, and W. X. Xing. Canonical dual approach to solving the maximum cut problem. *J. Global Optim.*, 54(2):341–351, 2012.

[129] K. Wilfred. *Advanced calculus*. Addison-Wesley Longman, Boston, 2002.

[130] Z. Wu, D. Li, L.-S. Zhang, and X.-M. Yang. Peeling off a nonconvex cover of an actual convex problem: hidden convexity. *SIAM J. Optimiz.*, 18(2):507–536, 2007.

[131] W. X. Xing, S. C. Fang, D. Y. Gao, R. L. Sheu, and L. Zhang. Canonical dual solutions to the quadratic programming over a quadratic constraint. *J. Global Optim.*, accepted for publication.

[132] Y. Yajima and T. Fujie. A polyhedral approach for nonconvex quadratic programming problems with box constraints. *J. Global Optim.*, 13(2):151–170, 1998.