STUDIES ON OPTIMIZATION METHODS FOR NONLINEAR SEMIDEFINITE PROGRAMMING PROBLEMS

YUYA YAMAKAWA

STUDIES ON
OPTIMIZATION METHODS FOR NONLINEAR SEMIDEFINITE PROGRAMMING PROBLEMS

by

YUYA YAMAKAWA

Submitted in partial fulfillment of
the requirement for the degree of
DOCTOR OF INFORMATICS
(Applied Mathematics and Physics)

KYOTO UNIVERSITY
KYOTO 606–8501, JAPAN
JANUARY 2015

Contents

Preface **v**

Acknowledgment **vii**

Notations **ix**

1 Introduction **1**
1.1 Nonlinear semidefinite programming problems and its applications 1
1.2 Solution methods for nonlinear semidefinite programming problems 4
1.2.1 Overview of solution methods 5
1.2.2 Primal-dual interior point methods 7
1.3 Motivations and contributions 10
1.4 Outline of the thesis 12

2 Preliminaries **13**
2.1 Notations and definitions 13
2.2 Fundamental mathematics 17
2.2.1 Linear algebra and analysis 17
2.2.2 Convex analysis 19
2.2.3 Symmetrized Kronecker product and its properties 21
2.2.4 Properties of a log-determinant function 24
2.3 Some optimality conditions for nonlinear SDP 26
2.4 Barrier KKT conditions for nonlinear SDP 28
2.5 Block coordinate descent method for nondifferentiable minimization 30

3 A differentiable merit function for shifted barrier Karush-Kuhn-Tucker conditions of nonlinear semidefinite programming problems **33**
3.1 Introduction 33
3.2 Primal-dual interior point method based on shifted barrier KKT conditions . . . 34
3.3 Finding a shifted barrier KKT point 35
3.3.1 Merit function and its properties 35
3.3.2 Newton-type method for minimization of the merit function 40
3.3.3 Global convergence of Algorithm 3.3.1 47
3.4 Numerical experiments 50

3.5 Concluding remarks 53

4 A two-step primal-dual interior point method for nonlinear semidefinite programming problems and its superlinear convergence 55

4.1 Introduction 55
4.2 Two-step primal-dual interior point method 56
4.2.1 Newton equation with scaling 56
4.2.2 Two-step primal-dual interior point method with the same coefficient matrix 58
4.3 Local and superlinear convergence of Algorithm 4.2.3 59
4.3.1 Assumptions and some resulting properties 60
4.3.2 Proof of superlinear convergence 64
4.4 Numerical experiments 72
4.5 Concluding remarks 74
A Proofs related to Assumptions 4.2.1 and 4.3.2 75

5 A block coordinate descent method for a maximum likelihood estimation problem of mixture distributions 79

5.1 Introduction 79
5.2 Maximum likelihood estimation for mixture distributions 80
5.3 Block coordinate descent method and its global convergence 83
5.4 Implementation issue for special cases 91
5.4.1 Maximum likelihood estimation with constraints on mixture coefficients . 92
5.4.2 Maximum likelihood estimation for Gaussian mixtures 95
5.4.3 Maximum likelihood estimation for Gaussian mixtures with constraints on precision matrices 96
5.4.4 Maximum likelihood estimation for Gaussian mixtures with sparse precision matrices 97
5.5 Numerical experiments 97
5.6 Concluding remarks 103

6 Conclusion 105

Preface

Nonlinear semidefinite programming (SDP) is a comparatively new problem which began to be studied from the 2000s. It is a natural extension of linear SDP, and includes a wide class of mathematical programming problems. In fact, nonlinear SDP represents not only linear SDP but also linear programming, second-order cone programming and nonlinear programming. There exist many applications that are formulated as nonlinear SDP, but cannot be represented as linear SDP. Thus, it is worth studying on optimization methods for nonlinear SDP in order to deal with such applications.

In this thesis, we focus on optimization methods for nonlinear SDP. Until now, some researchers have proposed solution methods for nonlinear SDP. Basically, these methods are derived from the existing methods for nonlinear programming, such as sequential quadratic programming methods, successive linearization methods, augmented Lagrangian methods and primal-dual interior point methods. Correa and Ramírez proposed a sequential semidefinite programming method which is an extension of a sequential quadratic programming method. Kanzow, Nagel, Kato and Fukushima presented a successive linearization method. Luo, Wu and Chen presented an augmented Lagrangian method. Yamashita, Yabe and Harada proposed a primal-dual interior point method. Although these methods can solve a certain nonlinear SDP, they still have theoretical and practical drawbacks. These methods have the global convergence property, which ensures to get a solution from an arbitrary initial point. However, these global convergence properties have been proven under some restrictive assumptions. To make matters worse, the assumptions include the boundedness of some generated sequences, which is not verified in advance.

The main purpose of this thesis is to propose efficient solution methods for nonlinear SDP and prove its convergence property under reasonable and clear assumptions. First, we propose a primal-dual interior point method with a Newton-type method. Moreover, we also propose a differentiable merit function, and we show some useful properties of the merit function. Especially, we prove that the level set of the merit function is bounded under some reasonable assumptions. The level boundedness of the merit function is not given in the literature related to nonlinear SDP. As the result, we show the global convergence of the proposed method with the merit function under some milder assumptions. Secondly, we present a two-step primal-dual interior point method for nonlinear SDP which is a modification of the first method proposed by Yamashita and Yabe. We prove its local and superlinear convergence. Note that two-step implies that two Newton equations are solved at each iteration. Yamashita and Yabe's two-step method has to solve two different Newton equations at each iteration. Although the proposed

method also has to solve two different Newton equations at each iteration, the coefficient matrix in the second Newton equation is equal to that in the first one. Thus, we can expect to reduce the computational cost to about half compared with that of Yamashita and Yabe's two-step method. In addition, we prove that the proposed method converges to a solution superlinearly under the same assumption as Yamashita and Yabe if we choose an initial point near the solution.

The second purpose of the thesis is to propose an efficient method for maximum likelihood estimation problems for mixture distributions. The estimation problems arise from various fields such as pattern recognition and machine learning. These problems are expressed as nonlinear SDP if mixture distributions are Gaussian mixtures. Recently, some researchers have considered the maximum likelihood estimation of a single Gaussian distribution with the L_1 regularization and/or some constraints on parameters. We present a general class of maximum likelihood estimation problems for mixture distributions that includes such regularized/constrained maximum likelihood estimation problems as a special case. Moreover, we propose a block coordinate descent (BCD) method for the general class. The BCD method sequentially solves small subproblems such that the objective function is minimized with respect to a few variables while all the other variables are fixed. In fact, this method is efficient if the subproblems are solved quickly. Thus, we propose some efficient methods for the subproblems when the problem has special structures.

The author hopes that the results of this thesis make some contributions to further studies on optimization methods for nonlinear semidefinite programming problems.

Yuya Yamakawa
January 2015

Acknowledgment

First of all, I would like to express my sincerest appreciation to Professor Nobuo Yamashita of Kyoto University, who supervised me throughout my doctoral course. Although I have often troubled him, he kindly supported me, gave me a lot of precise advice, and spared precious time for me in order to read my draft manuscript. Moreover, he taught me the importance of the eager and earnest attitude to the studies. I would like to bear such a teaching in my mind, and make effort to improve not only my technical skill of research in the future but also my mentality facing up to difficulties. I would also like to express my gratitude to Professor Masao Fukushima of Nanzan University. He also spared valuable time for me to read my manuscript, and suggested a lot of helpful comments to my research. I am grateful Professor Hiroshi Yabe of Tokyo University of Science. He supervised me throughout my master's course, and provided much useful advice for me even after I finished the master's course. I am indebted to Assistant Professor Ellen Hidemi Fukuda of Kyoto University. She always gave me helpful advice, encouraged me, and invited me to a meal.

I warmly thank all the past and present members of the Laboratory. It was very pleasant for me to discuss and study the optimization theory together with them. Especially, my heartfelt appreciation goes to Associate Professor Shunsuke Hayashi of Tohoku University, Assistant Professor Hiroshige Dan of Kansai University, and Assistant Professor Takayuki Okuno of Tokyo University of Science. They provided encouragement to me when I faced some difficulties. Moreover, they gave me much useful advice related to my job search.

Finally, I would like to dedicate this thesis to my family with my appreciation for their generous supports and encouragements.

Notations

$\mathbf{R}$	the set of real numbers
$\mathbf{R}^n$	the set of n-dimensional real vectors
$\mathbf{R}^{m \times n}$	the set of $m \times n$ real matrices
$\mathbf{S}^p$	the set of $p \times p$ real symmetric matrices
$\top$	the transposition of vectors or matrices
I	the identity matrix
v_i	the i-th element of a vector v
M_{ij}	the (i, j)-th element of a matrix M
$\mathrm{rank}(M)$	the rank of a matrix M
$\mathrm{tr}(M)$	the trace of a square matrix M
$\det(M)$	the determinant of a square matrix M
$\|\cdot\|$	the Euclidean norm
$\|\cdot\|_F$	the Frobenius norm
$M \succeq 0$	M is real symmetric positive semidefinite
$M \succ 0$	M is real symmetric positive definite
$A \succeq B$	$A - B$ is real symmetric positive semidefinite
$A \succ B$	$A - B$ is real symmetric positive definite
$\lambda_i(M)$	the eigenvalues of a real symmetric matrix M
$\lambda_{\min}(M)$	the minimum eigenvalue of a symmetric matrix M
$\lambda_{\max}(M)$	the maximum eigenvalue of a symmetric matrix M
$\nabla f(x)$	the gradient of a function f at x
$\nabla^2 f(x)$	the Hessian of a function f at x
$\log(x)$	the natural logarithm of a positive real number x
$\exp(x)$	e (Napier's constant) raised to the power of a real number x

Chapter 1

Introduction

1.1 Nonlinear semidefinite programming problems and its applications

In this thesis, we consider the following nonlinear semidefinite programming (SDP) problem:

$$\begin{array}{ll} \underset{x \in \mathbf{R}^n}{\text{minimize}} & f(x), \\ \text{subject to} & g(x) = 0,\ X(x) \succeq 0, \end{array} \tag{1.1.1}$$

where $f : \mathbf{R}^n \to \mathbf{R}$, $g : \mathbf{R}^n \to \mathbf{R}^m$ and $X : \mathbf{R}^n \to \mathbf{S}^p$ are twice continuously differentiable functions. Since nonlinear SDP (1.1.1) can be reduced to linear SDP if the functions f, g and X are all affine, we can say that nonlinear SDP (1.1.1) is a natural extension of linear SDP.

Nonlinear SDP is a comparatively new problem which began to be studied from the 2000s [4, 13, 19, 22, 29, 31, 32, 33, 34, 36, 49, 55, 56, 57, 58, 59, 65, 68, 69, 71, 72]. Moreover, it includes a wide class of mathematical programming problems, and has many applications. For example, linear programming [15], second-order cone programming [1], linear SDP [64] and nonlinear programming [6] can all be recast as nonlinear SDP.

Linear SDP has been studied extensively by many researchers [2, 17, 28, 60, 61, 64] because it arises from several fields such as statistics, finance, combinatorial optimization and control theory. Especially, primal-dual interior point methods are known as effective solution methods for linear SDP, and their theoretical and numerical analyses have been frequently done since the 1990s. However, there exist important formulations and applications that are expressed as nonlinear SDP, but cannot be reduced to linear SDP. In the following, we give some of such applications.

Problems with bilinear matrix inequality constraints

There exist optimization problems with bilinear (or biaffine) matrix inequality (BMI) constraints in many fields such as filtering problems [14] and structural optimization problems [27]. Optimization problems with BMI constraints are called BMI problems [21, 23, 50, 59, 63], which are

generally defined as

$$\begin{array}{ll} \text{minimize} & F(x,y), \\ \text{subject to} & M(x,y) \preceq 0, \end{array} \tag{1.1.2}$$

where $x \in \mathbf{R}^n$ and $y \in \mathbf{R}^m$ are decision variables, $F : \mathbf{R}^{n+m} \to \mathbf{R}$ is an objective function, and $M : \mathbf{R}^{n+m} \to \mathbf{S}^p$ is a quadratic function defined by

$$M(x,y) := A_0 + \sum_{i=1}^{n} x_i B_i + \sum_{j=1}^{m} y_j C_j + \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j D_{ij}$$

with constant matrices $A_0,\ B_i,\ C_j,\ D_{ij} \in \mathbf{S}^p$ $(i = 1,\ldots,n,\ j = 1,\ldots,m)$. Problem (1.1.2) is clearly nonlinear SDP (1.1.1).

Nearest correlation matrix problem

We present the following nearest correlation matrix problem with a rank constraint:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|X - A\|_F^2, \\ \text{subject to} & X \succeq 0, \\ & X_{ii} = 1,\ i = 1,\ldots,p, \\ & \text{rank}(X) \le r, \end{array} \tag{1.1.3}$$

where $X \in \mathbf{S}^p$ is a decision variable, $A \in \mathbf{S}^p$ is a constant matrix, and $r \in \mathbf{R}$ is a positive integer constant. The input matrix A is often a known correlation matrix but with rank larger than r. It is known that this problem has important applications in finance, etc. For further details, see [25, 74]. If $r = p$, problem (1.1.3) is equivalent to a standard nearest correlation matrix problem, that is,

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|X - A\|_F^2, \\ \text{subject to} & X \succeq 0, \\ & X_{ii} = 1,\ i = 1,\ldots,p. \end{array} \tag{1.1.4}$$

Note that problem (1.1.4) is convex, but problem (1.1.3) is nonconvex due to the constraint $\text{rank}(X) \le r$. In general, it is difficult to handle the constraint $\text{rank}(X) \le r$ directly. Thus, Li and Qi [37] showed that $X^* \in \mathbf{S}^p$ solves problem (1.1.3) if and only if there exists a matrix $U^* \in \mathbf{S}^p$ such that $(X^*, U^*) \in \mathbf{S}^p \times \mathbf{S}^p$ solves the following nonlinear SDP problem:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|X - A\|_F^2, \\ \text{subject to} & X \succeq 0, \\ & X_{ii} = 1,\ i = 1,\ldots,p, \\ & \text{tr}(XU) = p, \\ & \text{tr}(U) = r, \\ & I \succeq U \succeq 0, \end{array}$$

where $X,\ U \in \mathbf{S}^p$ are decision variables.

Maximum likelihood estimation problem

We provide applications that arise from the maximum likelihood estimation. In particular, we focus on the maximum likelihood estimation of parameters $\alpha_i \in \mathbf{R},\ \mu_i \in \mathbf{R}^d,\ \Lambda_i \in \mathbf{S}^d$ $(i = 1, \ldots, m)$ in Gaussian mixtures [8]:

$$p(x|\alpha, \mu, \Lambda) := \sum_{i=1}^{m} \alpha_i \mathcal{N}(x|\mu_i, \Lambda_i^{-1}),$$

where $\alpha := [\alpha_1, \ldots, \alpha_m],\ \mu := [\mu_1, \ldots, \mu_m],\ \Lambda := [\Lambda_1, \ldots, \Lambda_m]$ and

$$\mathcal{N}(x|\mu_i, \Lambda_i^{-1}) := \frac{\sqrt{\det \Lambda_i}}{(2\pi)^{d/2}} \exp\left[-\frac{1}{2}(x-\mu_i)^\top \Lambda_i (x-\mu_i)\right], \quad i = 1, \ldots, m.$$

In the maximum likelihood estimation, a log-likelihood function is maximized with respect to parameters, that is,

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_{k=1}^{n} \log\left(\sum_{i=1}^{m} \alpha_i \mathcal{N}(x_k|\mu_i, \Lambda_i^{-1})\right), \\ \text{subject to} & \alpha \in \Omega,\ \Lambda_i \succeq 0,\ i = 1, \ldots, m, \end{array} \tag{1.1.5}$$

where $\alpha_i \in \mathbf{R},\ \mu_i \in \mathbf{R}^d,\ \Lambda_i \in \mathbf{S}^d$ $(i = 1, \ldots, m)$ are decision variables, Ω is a certain set, and $x_k \in \mathbf{R}^d$ $(k = 1, \ldots, n)$ are observational data.

In addition, some researchers [38, 73] have recently investigated the maximum likelihood estimation of a single Gaussian distribution with the L_1 regularization and/or some constraints. In Chapter 5, we consider the following more general maximum likelihood estimation problem for Gaussian mixtures:

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_{k=1}^{n} \log\left(\sum_{i=1}^{m} \alpha_i \mathcal{N}(x_k|\mu_i, \Lambda_i^{-1})\right) - f_0(\alpha) - \sum_{i=1}^{m} \left[f_i^\mu(\mu_i) + f_i^\Lambda(\Lambda_i)\right], \\ \text{subject to} & \alpha \in \Omega,\ \Lambda_i \succeq 0,\ i = 1, \ldots, m, \end{array} \tag{1.1.6}$$

where $f_0,\ f_i^\mu$ and f_i^Λ are proper lower semicontinuous quasiconvex functions, such as indicator functions of sets which express constraints on $\alpha,\ \mu_i$ and Λ_i. If we choose appropriate functions $f_0,\ f_i^\mu$ and f_i^Λ according to additional constraints that we want to impose, we can obtain a maximum likelihood estimator that satisfies such constraints by solving problem (1.1.6). Note that problems (1.1.5) and (1.1.6) are nonlinear SDP.

Minimization of the maximal eigenvalue problem

The following minimization of the maximal eigenvalue problem arises mainly from the $\mathcal{H}_\infty$ controller design problem [10]:

$$\begin{array}{ll} \text{minimize} & \lambda_{\max}(M(v)), \\ \text{subject to} & v \in Q, \end{array} \tag{1.1.7}$$

where $v \in \mathbf{R}^n$ is a decision variable, M is a function from Q into $\mathbf{S}^p$, and $Q \subset \mathbf{R}^n$ is a constraint set. Note that M is not necessarily an affine function. Note also that $\lambda_{\max}(M(v)) \leq \eta$ if and

only if $\lambda_{\max}(M(v) - \eta I) \leq 0$, i.e., $\lambda_{\min}(\eta I - M(v)) \geq 0$. Thus, problem (1.1.7) is equivalent to the following nonlinear SDP:

$$\begin{array}{ll} \text{minimize} & \eta, \\ \text{subject to} & \eta I - M(v) \succeq 0, \\ & v \in Q, \end{array}$$

where $\eta \in \mathbf{R}$ and $v \in \mathbf{R}^n$ are decision variables.

Static output feedback control problem

In the static output feedback control, there exists the following SOF-$\mathcal{H}_\infty$ type problem:

$$\begin{array}{ll} \text{minimize} & \gamma, \\ \text{subject to} & Q \succeq 0, \\ & \gamma \geq 0, \\ & \left[\begin{array}{ccc} A(F)^\top Q + QA(F) & QB(F) & C(F)^\top \\ B(F)^\top Q & -\gamma I & D(F)^\top \\ C(F) & D(F) & -\gamma I \end{array} \right] \preceq 0, \end{array} \tag{1.1.8}$$

where $\gamma \in \mathbf{R}$, $F \in \mathbf{R}^{n_u \times n_y}$ and $Q \in \mathbf{S}^{n_x}$ are decision variables, and the functions A, B, C and D are defined by

$$\begin{aligned} A(F) &:= A + BFC, \\ B(F) &:= B_1 + BFD_{21}, \\ C(F) &:= C_1 + D_{12}FC, \\ D(F) &:= D_{11} + D_{12}FD_{21}, \end{aligned}$$

with given constant matrices $A \in \mathbf{R}^{n_x \times n_x}$, $B \in \mathbf{R}^{n_x \times n_u}$, $B_1 \in \mathbf{R}^{n_x \times n_w}$, $C \in \mathbf{R}^{n_y \times n_x}$, $C_1 \in \mathbf{R}^{n_z \times n_x}$, $D_{11} \in \mathbf{R}^{n_z \times n_w}$, $D_{12} \in \mathbf{R}^{n_z \times n_u}$ and $D_{21} \in \mathbf{R}^{n_y \times n_w}$. Furthermore, there also exists the following SOF-$\mathcal{H}_2$ type problem:

$$\begin{array}{ll} \text{minimize} & \mathrm{tr}(X), \\ \text{subject to} & Q \succeq 0, \\ & A(F)Q + QA(F)^\top + B_1 B_1^\top \preceq 0, \\ & \left[\begin{array}{cc} X & C(F)Q \\ QC(F)^\top & Q \end{array} \right] \succeq 0, \end{array} \tag{1.1.9}$$

where $X \in \mathbf{S}^{n_z}$, $F \in \mathbf{R}^{n_u \times n_y}$ and $Q \in \mathbf{S}^{n_x}$ are decision variables. Note that problems (1.1.8) and (1.1.9) have BMI constraints.

1.2 Solution methods for nonlinear semidefinite programming problems

The main goal of solution methods for nonlinear SDP (1.1.1) is to find a point that satisfies the first-order necessary optimality conditions for (1.1.1). The first-order necessary optimality

conditions are called the Karush-Kuhn-Tucker (KKT) conditions given by

$$\begin{bmatrix} \nabla_x L(x,y,Z) \\ g(x) \\ X(x)Z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad X(x) \succeq 0, \quad Z \succeq 0, \tag{1.2.1}$$

where $L : \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{S}^p \to \mathbf{R}$ is the Lagrangian function defined by $L(x,y,Z) = f(x) - g(x)^\top y - \mathrm{tr}(X(x)Z)$, and $y \in \mathbf{R}^m$ and $Z \in \mathbf{S}^p$ are Lagrange multipliers for $g(x) = 0$ and $X(x) \succeq 0$, respectively. Note that (x,y,Z) is called a KKT point of nonlinear SDP (1.1.1) if (x,y,Z) satisfies the KKT conditions (1.2.1). Note also that x is called a stationary point of nonlinear SDP (1.1.1) if there exist Lagrange multipliers y and Z such that (x,y,Z) is a KKT point. When a problem is convex, a stationary point is a global optimal solution. When a problem is nonconvex, it is difficult to find a global optimal solution, and hence we consider a method that finds a KKT point in general.

1.2.1 Overview of solution methods

Until now, some researchers have studied solution methods for nonlinear SDP since the 2000s. Basically, these methods are extensions of the existing methods for nonlinear programming.

Correa and Ramírez [13] proposed a sequential semidefinite programming method for nonlinear SDP. This method is an extension of a sequential quadratic programming method. It solves the following subproblem at the k-th iteration to get a search direction:

$$\begin{array}{ll} \underset{d \in \mathbf{R}^n}{\text{minimize}} & \nabla f(x_k)^\top d + \dfrac{1}{2} d^\top M_k d, \\ \text{subject to} & g(x_k) + J_g(x_k) d = 0, \\ & X(x_k) + \displaystyle\sum_{i=1}^{n} d_i A_i(x_k) \succeq 0, \end{array} \tag{1.2.2}$$

where x_k is the k-th iteration point, M_k is a certain symmetric positive definite matrix containing the second-order information of (1.1.1), $J_g(x_k)$ is a Jacobian of g at x_k, and $A_i(x_k)$ is a partial derivative of X at x_k with respect to its i-th component. Since M_k is symmetric positive definite, subproblem (1.2.2) has a unique global minimizer d_k. Using d_k, we get the next iteration point x_{k+1}. We often exploit the line search strategy in order to guarantee the global convergence, that is, we set $x_{k+1} := x_k + t_k d_k$, where t_k is a step size. In fact, Correa and Ramírez [13] used the line search strategy, and gave some conditions under which the proposed method is globally convergent. One of the conditions is the boundedness of the sequence $\{x_k\}$. However, they did not provide concrete sufficient conditions under which the sequence $\{x_k\}$ is bounded.

Kanzow, Nagel, Kato and Fukushima [33] presented a successive linearization method for nonlinear SDP. Although this method is essentially the same as the above sequential semidefinite programming method, it solves subproblem (1.2.2) with $M_k = c_k I$, where c_k is a positive

parameter. Such a subproblem is equivalent to the following linear SDP:

$$\begin{array}{ll} \underset{(d,t)\in\mathbf{R}^n\times\mathbf{R}}{\text{minimize}} & t, \\ \text{subject to} & g(x_k) + J_g(x_k)d = 0, \\ & X(x_k) + \displaystyle\sum_{i=1}^{n} d_i A_i(x_k) \succeq 0, \\ & \begin{bmatrix} I & \sqrt{c_k}d \\ \sqrt{c_k}d^\top & t - \nabla f(x_k)^\top d \end{bmatrix} \succeq 0. \end{array}$$

Moreover, this method has no line search strategy. Instead, it adjusts the length of the search direction d_k by selecting the parameter c_k appropriately. However, they showed the global convergence of the proposed method under rather strong assumptions on generated sequences. Meanwhile, it is generally known that such a method has a slow convergence rate because subproblem (1.2.2) with $M_k = c_k I$ does not contain the second-order information of (1.1.1).

Luo, Wu and Chen [41] presented augmented Lagrangian methods for nonlinear SDP. First, these methods obtain a new primal variable x_k by solving the following unconstrained minimization subproblem at each iteration:

$$\underset{x\in\mathbf{R}^n}{\text{minimize}} \quad \mathcal{L}_{c_k}(x, y_k, Z_k),$$

where c_k is a positive parameter, and $\mathcal{L}_c : \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{S}^p \to \mathbf{R}$ is an augmented Lagrangian function defined by

$$\mathcal{L}_c(x, y, Z) := f(x) + \frac{1}{2c}\left(\mathrm{tr}([Z + cX(x)]_+^2) - \mathrm{tr}(Z^2)\right) + g(x)^\top y + \frac{c}{2}\|g(x)\|^2,$$

where $[\,\cdot\,]_+ : \mathbf{S}^p \to \mathbf{S}^p$ is a operator defined by

$$[A]_+ := P \begin{bmatrix} \max\{0, \lambda_1(A)\} & & 0 \\ & \ddots & \\ 0 & & \max\{0, \lambda_p(A)\} \end{bmatrix} P^\top,$$

and P is an orthogonal matrix in an orthogonal decomposition of A. Secondly, the methods update the positive parameter c_k and the Lagrange multipliers y_k and Z_k appropriately. The augmented Lagrangian methods get a solution by repeating such two procedures.

Luo, Wu and Chen [41] gave various types of updating methods associated with the positive parameter c_k and the Lagrange multipliers y_k and Z_k for the global convergence of the augmented Lagrangian methods. Furthermore, they proved the global convergence of the proposed methods under some assumptions which include that the sequence $\{x_k\}$ is bounded and the sequence $\{c_k\}$ diverges to ∞. However, we are anxious about becoming numerically unstable by the second assumption. On the other hand, it is generally known that augmented Lagrangian methods have a slow convergence rate.

Recently, several researchers have proposed primal-dual interior point methods for nonlinear SDP [34, 71, 72]. We give details of these methods in the next subsection.

In nonlinear programming, block coordinate descent (BCD) methods are often used for solving large-scale problems. They sequentially solve small subproblems such that the objective

function is minimized with respect to a few variables while all the other variables are fixed. Thus, BCD methods are efficient for large-scale problems if the subproblems are solved quickly. Although several types of BCD methods [3, 66] have recently been proposed for linear SDP, such methods have not yet been studied for nonlinear SDP. We will propose a BCD method for nonlinear SDP derived from maximum likelihood estimation problems for mixture distributions.

1.2.2 Primal-dual interior point methods

Although there exist several solution methods described above, we mainly focus on primal-dual interior point methods. In particular, there exist roughly two primal-dual interior point methods. One is a method based on the following barrier KKT conditions:

$$r_0(x, y, Z, \mu) := \begin{bmatrix} \nabla_x L(x, y, Z) \\ g(x) \\ X(x)Z - \mu I \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad X(x) \succ 0, \quad Z \succ 0, \tag{1.2.3}$$

where $\mu > 0$ is called a barrier parameter. The primal-dual interior point methods proposed by Yamashita and Yabe [71] and Yamashita, Yabe and Harada [72] are based on the barrier KKT conditions (1.2.3). Another is a method based on the following shifted barrier KKT conditions:

$$r_1(x, y, Z, \mu) := \begin{bmatrix} \nabla_x L(x, y, Z) \\ g(x) + \mu y \\ X(x)Z - \mu I \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad X(x) \succ 0, \quad Z \succ 0. \tag{1.2.4}$$

The primal-dual interior point method proposed by Kato, Yabe and Yamashita [34] is based on the shifted barrier KKT conditions (1.2.4). Note that a point (x, y, Z) satisfying $X(x) \succ 0$ and $Z \succ 0$ is called an interior point. Note also that a point (x, y, Z) satisfying (1.2.3) or (1.2.4) is an interior point.

When $\mu \to 0$ in (1.2.3) and (1.2.4), a point which satisfies the (shifted) barrier KKT conditions comes close to a KKT point which satisfies the KKT conditions (1.2.1). Therefore, the primal-dual interior point methods described above generate an interior point which satisfies the (shifted) barrier KKT conditions approximately for a given barrier parameter μ_k, and update the barrier parameter so as to satisfy $0 < \mu_{k+1} < \mu_k$ at each iteration. Summing up the above discussion, we give a framework of a primal-dual interior point method. To this end, we use the following notations:

$$\rho(x, y, Z) := \sqrt{\left\| \begin{bmatrix} \nabla_x L(x, y, Z) \\ g(x) \end{bmatrix} \right\|^2 + \|X(x)Z\|_F^2},$$

$$\rho_0(x, y, Z, \mu) := \sqrt{\left\| \begin{bmatrix} \nabla_x L(x, y, Z) \\ g(x) \end{bmatrix} \right\|^2 + \|X(x)Z - \mu I\|_F^2},$$

$$\rho_1(x, y, Z, \mu) := \sqrt{\left\| \begin{bmatrix} \nabla_x L(x, y, Z) \\ g(x) + \mu y \end{bmatrix} \right\|^2 + \|X(x)Z - \mu I\|_F^2}.$$

Primal-dual interior point method

Step 0. Choose positive constants ε and σ. Select a positive sequence $\{\mu_k\}$ converging to 0. Set $k := 0$.

Step 1. Find an interior point $(x_{k+1}, y_{k+1}, Z_{k+1})$ which satisfies the (shifted) barrier KKT conditions approximately for μ_k, i.e.,

$$X(x_{k+1}) \succ 0,\ Z_{k+1} \succ 0,\ \rho_0(x_{k+1}, y_{k+1}, Z_{k+1}, \mu_k) \leq \sigma\mu_k\ (\rho_1(x_{k+1}, y_{k+1}, Z_{k+1}, \mu_k) \leq \sigma\mu_k).$$

Step 2. If $\rho(x_{k+1}, y_{k+1}, Z_{k+1}) \leq \varepsilon$ is satisfied, then stop.

Step 3. Set $k := k + 1$ and go to Step 1. □

In the above method, we have to find an interior point which satisfies the (shifted) barrier KKT conditions approximately for μ_k. In order to find such a point, a Newton-type method is used in [34, 71, 72].

Newton equations in the Newton-type method are generated from $r_0(x, y, Z, \mu) = 0$ or $r_1(x, y, Z, \mu) = 0$. Before we present the concrete Newton equations, we introduce scaling. Scaling is frequently exploited in order to solve Newton equations efficiently as seen later. Instead of $X(x)$ and Z, we deal with matrices $\widetilde{X}(x) := TX(x)T^\top$ and $\widetilde{Z} := T^{-\top}ZT^{-1}$, where a nonsingular scaling matrix T satisfies that

$$TX(x)T^\top T^{-\top}ZT^{-1} = T^{-\top}ZT^{-1}TX(x)T^\top. \tag{1.2.5}$$

Then, $\widetilde{X}(x)\widetilde{Z} = \widetilde{Z}\widetilde{X}(x)$ from (1.2.5). Moreover, we replace the matrices $X(x)$ and Z in (1.2.3) and (1.2.4) with $\widetilde{X}(x)$ and $\widetilde{Z}$, respectively. Then, we define the scaled barrier KKT conditions as

$$\widetilde{r}_0(x, y, Z, \mu) := \begin{bmatrix} \nabla_x L(x, y, Z) \\ g(x) \\ \widetilde{X}(x)\widetilde{Z} - \mu I \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \widetilde{X}(x) \succ 0, \quad \widetilde{Z} \succ 0.$$

Similarly, we define the scaled shifted barrier KKT conditions as

$$\widetilde{r}_1(x, y, Z, \mu) := \begin{bmatrix} \nabla_x L(x, y, Z) \\ g(x) + \mu y \\ \widetilde{X}(x)\widetilde{Z} - \mu I \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \widetilde{X}(x) \succ 0, \quad \widetilde{Z} \succ 0.$$

Note that the scaled (shifted) barrier KKT conditions are equivalent to the (shifted) barrier KKT conditions.

Next, we present Newton equations. Newton equations are generated by $\widetilde{r}_0(x, y, Z, \mu) = 0$ or $\widetilde{r}_1(x, y, Z, \mu) = 0$. When we generate Newton equations from $\widetilde{r}_0(x, y, Z, \mu) = 0$, they are expressed as

$$\begin{bmatrix} G + H & -J_g(x)^\top \\ J_g(x) & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \begin{bmatrix} -\nabla f(x) + J_g(x)^\top y + \mu\mathcal{A}^*(x)X(x)^{-1} \\ -g(x) \end{bmatrix}, \tag{1.2.6}$$

$$(T^\top \odot T^\top)(\widetilde{X}(x) \odot I)^{-1}(\widetilde{Z} \odot I)(T \odot T)\mathcal{A}(x)\Delta x + \Delta Z = \mu X(x)^{-1} - Z, \tag{1.2.7}$$

where

$(\Delta x, \Delta y, \Delta Z) \in \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{S}^p$ is the Newton direction,
$G \in \mathbf{S}^n$ is $\nabla^2_{xx} L(x, y, Z)$ or its approximation,
$J_g(x) \in \mathbf{R}^{m \times n}$ is a Jacobian of g at x,
$\mathcal{A}(x) : \mathbf{R}^n \to \mathbf{S}^p$ is an operator such that $v \mapsto \sum_{i=1}^n v_i A_i(x)$, where $A_i(x) := \frac{\partial}{\partial x_i} X(x)$,
$\mathcal{A}^*(x) : \mathbf{S}^p \to \mathbf{R}^n$ is an adjoint operator of $\mathcal{A}(x)$ such that $U \mapsto [\mathrm{tr}(A_1(x)U), \ldots, \mathrm{tr}(A_n(x)U)]^\top$,
$(P \odot Q) : \mathbf{S}^p \to \mathbf{S}^p$ is an operator such that $U \mapsto \frac{1}{2}(PUQ^\top + QUP^\top)$, where $P,\ Q \in \mathbf{R}^{p \times p}$,

and $H \in \mathbf{R}^{n \times n}$ is a matrix whose (i, j)-th element is given by

$$H_{ij} := \mathrm{tr}\left[A_i(x)(T^\top \odot T^\top)(\widetilde{X}(x) \odot I)^{-1}(\widetilde{Z} \odot I)(T \odot T) A_j(x) \right]. \tag{1.2.8}$$

Note that these Newton equations are used in [71] and [72]. Similarly, when we generate Newton equations from $\widetilde{r}_1(x, y, Z, \mu) = 0$, they are expressed as

$$\begin{bmatrix} G + H & -J_g(x)^\top \\ J_g(x) & \mu I \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \begin{bmatrix} -\nabla f(x) + J_g(x)^\top y + \mu \mathcal{A}^*(x) X(x)^{-1} \\ -g(x) - \mu y \end{bmatrix},$$
$$(T^\top \odot T^\top)(\widetilde{X}(x) \odot I)^{-1}(\widetilde{Z} \odot I)(T \odot T)\mathcal{A}(x)\Delta x + \Delta Z = \mu X(x)^{-1} - Z.$$

Note that these Newton equations are exploited in [34].

The operator $(\widetilde{X}(x) \odot I)^{-1}$ in the Newton equations are usually difficult to handle. This is because when we calculate $(\widetilde{X}(x) \odot I)^{-1} U = V$, we have to solve a Lyapunov equation $X(x)V + VX(x) = 2U$ with respect to V. However, note that the operator $(\widetilde{X}(x) \odot I)^{-1}$ appears as $(\widetilde{X}(x) \odot I)^{-1}(\widetilde{Z} \odot I)$. Thus, when $\widetilde{X}(x) = I$, it is clear that $(\widetilde{X}(x) \odot I)^{-1}(\widetilde{Z} \odot I) = (\widetilde{Z} \odot I)$ and $\widetilde{X}(x)\widetilde{Z} = \widetilde{Z}\widetilde{X}(x)$. On the other hand, when $\widetilde{X}(x) = \widetilde{Z}$, we see that $(\widetilde{X}(x) \odot I)^{-1}(\widetilde{Z} \odot I)$ is the identity mapping, and $\widetilde{X}(x)\widetilde{Z} = \widetilde{Z}\widetilde{X}(x)$. Therefore, if we choose the scaling matrix T such that $\widetilde{X}(x) = I$ or $\widetilde{X}(x) = \widetilde{Z}$, we do not have to solve the Lyapunov equation. If we do not use scaling, that is, $T = I$, then we have to solve the Lyapunov equation. This is one of reasons why we exploit scaling.

Yamashita, Yabe and Harada [72] presented a nondifferentiable L_1 merit function, and showed the global convergence of the proposed method with their merit function under some unclear assumptions regarding a generated sequence. Kato, Yabe and Yamashita [34] proposed a differentiable merit function, and proved the global convergence of the proposed method with their merit function under some weaker assumptions compared with those of [72]. However, since the proposed merit function is rather complicated, the convergence analysis is also complicated. Furthermore, it might not be easy to implement the proposed method with their merit function. Note that [34] and [72] do not investigate the rate of convergence of their methods.

Yamashita and Yabe [71] investigated the superlinear convergence of the primal-dual interior point method. They presented two methods. One is a method with scaling. Another is a method without scaling. In general, since scaling is frequently exploited in order to solve the Newton equation effectively as mentioned above, the method with scaling is more important than the method without scaling.

However, although the method without scaling may only solve one Newton equation in a single iteration, the method with scaling has to solve two different Newton equations in a single

iteration. Thus, the method with scaling is called a two-step primal-dual interior point method. At the first part of the k-th iteration in the two-step primal-dual interior point method, we obtain the Newton direction $(\Delta x_k, \Delta y_k, \Delta Z_k)$ from the Newton equations (1.2.6) and (1.2.7) as $(x, y, Z) = (x_k, y_k, Z_k)$, and update $(x_{k+\frac{1}{2}}, y_{k+\frac{1}{2}}, Z_{k+\frac{1}{2}}) = (x_k + \Delta x_k, y_k + \Delta y_k, Z_k + \Delta Z_k)$, where (x_k, y_k, Z_k) denotes the k-th iteration point. Furthermore, at the second part of the k-th iteration, we obtain the Newton direction $(\Delta x_{k+\frac{1}{2}}, \Delta y_{k+\frac{1}{2}}, \Delta Z_{k+\frac{1}{2}})$ from the Newton equations (1.2.6) and (1.2.7) as $(x, y, Z) = (x_{k+\frac{1}{2}}, y_{k+\frac{1}{2}}, Z_{k+\frac{1}{2}})$, update $(x_{k+1}, y_{k+1}, Z_{k+1}) = (x_{k+\frac{1}{2}} + \Delta x_{k+\frac{1}{2}}, y_{k+\frac{1}{2}} + \Delta y_{k+\frac{1}{2}}, Z_{k+\frac{1}{2}} + \Delta Z_{k+\frac{1}{2}})$, and go to the next iteration. Then, the main calculation is to obtain the Newton directions $(\Delta x_k, \Delta y_k, \Delta Z_k)$ and $(\Delta x_{k+\frac{1}{2}}, \Delta y_{k+\frac{1}{2}}, \Delta Z_{k+\frac{1}{2}})$. In particular, it is known that a construction of the matrix H is the biggest burden, and its calculation time is $O(np^3 + n^2p^2)$ from (1.2.8). In summary, the two-step primal-dual interior point method has to construct the matrix H twice in a single iteration.

1.3 Motivations and contributions

As mentioned in Section 1.1, there exist many applications of nonlinear SDP. Moreover, although some researchers have studied primal-dual interior point methods for nonlinear SDP, there still exist a lot of issues which should be studied. On the other hand, since primal-dual interior point methods are based on Newton-type methods, they may not be suitable to some large-scale nonlinear SDP.

Therefore, such many applications and issues motivate us to study solution methods for nonlinear SDP. In the following, we describe concrete aims and contributions of this study.

(1) To propose a primal-dual interior point method that is convergent globally under milder conditions

One of aims on this study is to propose a primal-dual interior point method for nonlinear SDP (1.1.1) that is convergent globally under milder conditions compared with the existing methods described in Section 1.2. In particular, there exist some unclear assumptions on a generated sequence in [34] and [72]. We specify conditions for the global convergence related to the problem data, i.e., f, g and X of (1.1.1).

Moreover, we also present a differentiable merit function F which has some nice properties compared with those of [34]. This function is an extension of a merit function proposed by Forsgren and Gill [18] developed for nonlinear programming, and it consists of simple functions, such as log-determinant and trace. Thus, it is easy to implement the proposed method with the merit function F. We show the following important properties of the merit function F:

(i) The merit function F is differentiable;

(ii) Any stationary point of the merit function F is a shifted barrier KKT point;

(iii) The level set of the merit function F is bounded under some reasonable assumptions.

Kato, Yabe and Yamashita [34] also showed that their merit function satisfies the properties (i) and (ii), but they did not show the property (iii). These properties mean that we can find a point that satisfies the shifted barrier KKT conditions by minimizing the merit function F.

(2) To propose a superlinear convergent two-step primal-dual interior point method that exploits scaling but its computational cost at each iteration is almost same as that of a one-step primal-dual interior point method

Next, we present a primal-dual interior point method for nonlinear SDP (1.1.1) which has the local and superlinear convergence property. As already mentioned in Section 1.2, Yamashita and Yabe [71] proposed a two-step primal-dual interior point method with scaling, which has to solve two different Newton equations in a single iteration. In this thesis, we also present a two-step primal-dual interior point method with scaling. However, in order to reduce calculations, we replace the coefficient matrix in the second equation with that in the first one. Thus, we can solve the second equation more rapidly using some computational results obtained by solving the first equation. Recall that the great portion of the computational time is to construct the matrix H defined by (1.2.8), and its computational time is $O(np^3 + n^2p^2)$ as described in Section 1.2. Although the method proposed by [71] has to construct the matrix H twice in a single iteration, the method proposed by this thesis calculates the matrix H only once in a single iteration. In other words, its computational cost at each iteration is almost same as that of a one-step primal-dual interior point method. As the result, we can expect to reduce the computational cost to about half compared with that of Yamashita and Yabe's two-step method [71]. In addition, we show the superlinear convergence under the same assumptions as [71] despite this change.

(3) To model a general maximum likelihood estimation problem, and give a block coordinate descent method for the problem

Finally, we consider an efficient solution method for a concrete application of nonlinear SDP. Then, we focus on maximum likelihood estimation problems for mixture distributions. Recently, some researchers have studied the maximum likelihood estimation of a single Gaussian distribution with the L_1 regularization and/or some constraints. We present a general class of maximum likelihood estimation problems for mixture distributions that includes such regularized/constrained maximum likelihood estimation problems as a special case. Such a general class is reduced to nonlinear SDP when the mixture distribution is the Gaussian mixtures. However, it may not be suitable to solve the problem by the primal-dual interior point method when the problem is large-scale. As described in Subsection 1.2.1, BCD methods are efficient for large-scale problems, and hence we propose a BCD method for the general class of maximum likelihood estimation problems for mixture distributions. Since the proposed BCD method has to solve simple subproblems at each iteration, we also propose efficient methods for such subproblems by exploiting their special structure.

1.4 Outline of the thesis

This thesis is organized as follows.

In Chapter 2, we first introduce some notations and definitions. Secondly, we provide several basic properties of mathematics. Moreover, we present some optimality conditions and barrier KKT conditions for nonlinear SDP (1.1.1). Finally, we give some concepts related to BCD methods for general nonlinear programming.

In Chapter 3, we first give a framework of a primal-dual interior point method based on the shifted barrier KKT conditions. Next, we propose a differentiable merit function F for the shifted barrier KKT conditions, and prove some nice properties of the merit function F. Moreover, we construct a Newton-type method for minimizing the merit function F, and show its global convergence under milder conditions. Finally, we report some numerical experiments for the proposed method.

In Chapter 4, we present a two-step primal-dual interior point method with scaling which solves two different Newton equations in a single iteration. Then, we argue that the proposed method is expected to find the next point faster than Yamashita and Yabe's two-step method [71] at each iteration because the two equations have the same coefficient matrices. Moreover, we prove the superlinear convergence of the proposed method under the same assumptions as those of Yamashita and Yabe [71]. Finally, we report some numerical experiments for the proposed method.

In Chapter 5, we consider maximum likelihood estimation problems for mixture distributions. Then, we mention that maximum likelihood estimation problems are written as nonlinear SDP when the mixture distribution is Gaussian mixtures. Moreover, we propose a general class of maximum likelihood estimation problems for mixture distributions that includes maximum likelihood estimation problems with the L_1 regularization and/or some constraints as a special case, and we present a BCD method for the general class. Then, since we must solve some subproblems generated in the proposed BCD method, we give efficient solution methods for such subproblems. Finally, we report some numerical experiments related to maximum likelihood estimation problems for Gaussian mixtures.

In Chapter 6, we give some concluding remarks, and state future works.

Chapter 2

Preliminaries

In this chapter, we introduce some mathematical notations, definitions and concepts. Note that propositions and theorems with proof are new results of this thesis.

2.1 Notations and definitions

We introduce some sets in the following. Let $m,\ n,\ p$ be positive integers.

$\mathbf{R}$	the set of real numbers
$\mathbf{R}^n$	the set of n-dimensional real vectors
$\mathbf{R}^{m\times n}$	the set of $m \times n$ real matrices
$\mathbf{S}^p$	the set of $p \times p$ real symmetric matrices

We use the following notations.

$\top$	the transposition of vectors or matrices
I	the identity matrix
I_n	the $n \times n$ identity matrix
v_i	the i-th element of a vector v
M_{ij}	the (i,j)-th element of a matrix M
$\mathrm{rank}(M)$	the rank of a matrix M
$\mathrm{tr}(M)$	the trace of a square matrix M
$\det(M)$	the determinant of a square matrix M

Moreover, we define some subsets of $\mathbf{R}^n$ and $\mathbf{R}^{m\times n}$.

$$\begin{aligned}
\mathbf{R}^n_+ &:= \{\ v \in \mathbf{R}^n \mid v_i \geq 0,\ i = 1, \ldots, n\ \},\\
\mathbf{R}^n_{++} &:= \{\ v \in \mathbf{R}^n \mid v_i > 0,\ i = 1, \ldots, n\ \},\\
\mathbf{R}^{m\times n}_+ &:= \{\ M \in \mathbf{R}^{m\times n} \mid M_{ij} \geq 0,\ i = 1, \ldots, m,\ j = 1, \ldots, n\ \},\\
\mathbf{R}^{m\times n}_{++} &:= \{\ M \in \mathbf{R}^{m\times n} \mid M_{ij} > 0,\ i = 1, \ldots, m,\ j = 1, \ldots, n\ \}.
\end{aligned}$$

We introduce some definitions in basic mathematics. Vectors $v_1, \ldots, v_n$ are called linearly independent if there exists no set of scalars $t_1, \ldots, t_n$, at least one of which is nonzero, such that

$t_1v_1 + \cdots + t_nv_n = 0$. Matrices $M_1, \ldots, M_n$ are called linearly independent if there exists no set of scalars $t_1, \ldots, t_n$, at least one of which is nonzero, such that $t_1M_1 + \cdots + t_nM_n = 0$. A square matrix M is called singular if $\det(M) = 0$. A square matrix M is called nonsingular or invertible if $\det(M) \neq 0$. A square matrix M is called the inverse of a nonsingular matrix N, and it is denoted by N^{-1} if $MN = NM = I$. A square matrix U is called an orthogonal matrix if $U^\top U = UU^\top = I$.

We define an inner product $\langle \cdot, \cdot \rangle$ and a norm $\| \cdot \|$ on $\mathbf{R}^n$ as follows: For any vectors $a,\ b \in \mathbf{R}^n$,

$$\langle a, b \rangle := a^\top b, \qquad \|a\| := \sqrt{\langle a, a \rangle},$$

respectively, where the norm is called the Euclidean norm. We define an inner product $\langle \cdot, \cdot \rangle$ and norms $\| \cdot \|_F,\ \| \cdot \|_1$ and $\| \cdot \|_2$ on $\mathbf{R}^{m \times n}$ as follows: For any matrices $A,\ B \in \mathbf{R}^{m \times n}$,

$$\langle A, B \rangle := \mathrm{tr}(A^\top B), \quad \|A\|_F := \sqrt{\langle A, A \rangle}, \quad \|A\|_1 = \sum_{i=1}^{m} \sum_{j=1}^{n} |A_{ij}|, \quad \|A\|_2 := \sup_{v \in \mathbf{R}^n \backslash \{0\}} \frac{\|Av\|}{\|v\|}, \tag{2.1.1}$$

respectively, where the first norm is called the Frobenius norm, the second norm is called the L_1 norm, and the last norm is called the operator norm. We define an inner product $\langle \cdot, \cdot \rangle$ and norms $\| \cdot \|_F,\ \| \cdot \|_1$ and $\| \cdot \|_2$ on $\mathbf{S}^p$ as (2.1.1). In the following, we call a set with an inner product $\langle \cdot, \cdot \rangle$, such as $\mathbf{R}^n,\ \mathbf{R}^{m \times n}$ and $\mathbf{S}^p$, an inner product space. Unless otherwise noted, we define a norm $\| \cdot \|$ on an inner product space as $\| \cdot \| := \sqrt{\langle \cdot, \cdot \rangle}$.

Let $S_1, \ldots, S_n$ be sets. We define the Cartesian product of $S_1, \ldots, S_n$ as

$$S_1 \times \cdots \times S_n := \{\ [s_1, \ldots, s_n] \mid s_1 \in S_1, \ldots, s_n \in S_n\ \}.$$

For any element $s \in S_1 \times \cdots \times S_n$, we use the following notations by using certain elements $s_1 \in S_1, \ldots, s_n \in S_n$:

$$s = \begin{bmatrix} s_1 \\ \vdots \\ s_n \end{bmatrix}, \quad s = [s_1, \ldots, s_n].$$

Moreover, let $[s_1, \ldots, s_n],\ [t_1, \ldots, t_n] \in S_1 \times \cdots \times S_n$. We say that $[s_1, \ldots, s_n]$ and $[t_1, \ldots, t_n]$ are equal if $s_1 = t_1, \ldots, s_n = t_n$.

Let $\mathcal{V}_1, \ldots, \mathcal{V}_n$ be inner product spaces. We define an inner product $\langle \cdot, \cdot \rangle$ on $\mathcal{V}_1 \times \cdots \times \mathcal{V}_n$ as follows: For any elements $v = [v_1, \ldots, v_n],\ w = [w_1, \ldots, w_n] \in \mathcal{V}_1 \times \cdots \times \mathcal{V}_n$,

$$\langle v, w \rangle := \langle v_1, w_1 \rangle + \cdots + \langle v_n, w_n \rangle.$$

We define the positive semidefiniteness and definiteness of real symmetric matrices. A real symmetric matrix $M \in \mathbf{S}^p$ is called positive semidefinite if

$$\langle Mv, v \rangle \geq 0 \quad \text{for all } v \in \mathbf{R}^n.$$

A real symmetric matrix $M \in \mathbf{S}^p$ is called positive definite if

$$\langle Mv, v \rangle > 0 \quad \text{for all } v \in \mathbf{R}^n \backslash \{0\}.$$

Then, we define the following notations related to the positive semidefiniteness and definiteness of real symmetric matrices.

$\mathbf{S}^p_+$	the set of $p \times p$ real symmetric positive semidefinite matrices
$\mathbf{S}^p_{++}$	the set of $p \times p$ real symmetric positive definite matrices
$M \succeq 0$	M is real symmetric positive semidefinite
$M \succ 0$	M is real symmetric positive definite
$A \succeq B$	$A - B$ is real symmetric positive semidefinite
$A \succ B$	$A - B$ is real symmetric positive definite

Moreover, we define the following notations related to eigenvalues of real symmetric matrices.

$\lambda_i(M)$	the eigenvalue of a real symmetric matrix M
$\lambda_{\min}(M)$	the minimum eigenvalue of a real symmetric matrix M
$\lambda_{\max}(M)$	the maximum eigenvalue of a real symmetric matrix M

For $d_1, \ldots, d_p \in \mathbf{R}$, we define

$$\mathrm{diag}(d_1, \ldots, d_p) := \begin{bmatrix} d_1 & & O \\ & \ddots & \\ O & & d_p \end{bmatrix}.$$

For a matrix $V \in \mathbf{S}^p_+$ $(\in \mathbf{S}^p_{++})$, $V^{\frac{1}{2}}$ denotes a real symmetric positive semidefinite (definite) matrix such that $V = V^{\frac{1}{2}} V^{\frac{1}{2}}$, that is,

$$V^{\frac{1}{2}} := U \Lambda U^\top, \quad \Lambda := \mathrm{diag}\left[(\lambda_1(V))^{\frac{1}{2}}, \ldots, (\lambda_p(V))^{\frac{1}{2}} \right],$$

where U is a certain orthogonal matrix such that $V = U \Lambda^2 U^\top$.

Let $\mathcal{V}$, $\mathcal{W}$ and $\mathcal{X}$ be inner product spaces, such as $\mathbf{R}^n$, $\mathbf{R}^{m \times n}$ and $\mathbf{S}^p$. For $x \in \mathcal{V}$ and $r > 0$, we define

$$B(x, r) := \{\ v \in \mathcal{V} \mid \|v - x\| < r\ \} \subset \mathcal{V}.$$

We say that a set $S \subset \mathcal{V}$ is bounded if

$$\exists x \in \mathcal{V}, \quad \exists r \in (0, \infty) \quad \text{such that} \quad S \subset B(x, r);$$

a set $S \subset \mathcal{V}$ is open if

$$\forall v \in S, \quad \exists r > 0 \quad \text{such that} \quad B(v, r) \subset S;$$

a set $S \subset \mathcal{V}$ is closed if $\mathcal{V} \backslash S$ is open; a set $S \subset \mathcal{V}$ is compact if S is bounded and closed. Let $\varphi : S \to \mathcal{W}$ be a function, where $S \subset \mathcal{V}$ is a set. We say that the function φ is continuous at $x \in S$ if

$$\forall \varepsilon > 0, \quad \exists \delta > 0 \quad \text{such that} \quad \|\varphi(x) - \varphi(y)\| < \varepsilon, \quad \forall y \in B(x, \delta) \cap S;$$

the function φ is continuous on S if φ is continuous at all $x \in S$. When $S = \mathcal{V}$, we say that the function φ is continuous if φ is continuous at all $x \in \mathcal{V}$. When $\mathcal{W} = \mathbf{R}$, we say that the function φ is lower semicontinuous at $x \in S$ if

$$\forall \varepsilon > 0, \quad \exists \delta > 0 \quad \text{such that} \quad \varphi(x) < \varphi(y) + \varepsilon, \quad \forall y \in B(x, \delta) \cap S;$$

the function φ is lower semicontinuous on S if φ is lower semicontinuous at all $x \in S$. When $S = \mathcal{V}$ and $\mathcal{W} = \mathbf{R}$, we say that the function φ is lower semicontinuous if φ is lower semicontinuous at all $x \in \mathcal{V}$.

Let $\phi : B(v, r) \to \mathcal{W}$ and $\psi : B(v, r) \to \mathcal{X}$ be functions, where $r > 0$ and $v \in \mathcal{V}$. If the functions ϕ and ψ satisfy that

$$\lim_{h \to v} \frac{\|\phi(h)\|}{\|\psi(h)\|} = 0, \tag{2.1.2}$$

we express (2.1.2) as $\phi(h) = o(\psi(h))$ $(h \to v)$. If the functions ϕ and ψ satisfy that there exists a positive constant $c \in \mathbf{R}$ such that

$$\lim_{h \to v} \frac{\|\phi(h)\|}{\|\psi(h)\|} = c, \tag{2.1.3}$$

we express (2.1.3) as $\phi(h) = O(\psi(h))$ $(h \to v)$.

For sets S and T, we denote a set of linear bounded operators from S into T by $L(S, T)$. Let $\Phi : D \to \mathcal{W}$ be a function, where $D \subset \mathcal{V}$ is an open set. The function Φ is called Fréchet differentiable at $x \in D$ if there exists $A_x \in L(\mathcal{V}, \mathcal{W})$ such that, for any $\Delta x \in \mathcal{V}$ with $x + \Delta x \in D$,

$$\Phi(x + \Delta x) = \Phi(x) + A_x(\Delta x) + o(\|\Delta x\|) \qquad (\|\Delta x\| \to 0).$$

The function Φ is called Fréchet differentiable on D if Φ is Fréchet differentiable at all $x \in D$. When $D = \mathcal{V}$, the function Φ is called Fréchet differentiable if Φ is Fréchet differentiable at all $x \in \mathcal{V}$. Note that if Φ is Fréchet differentiable on D, the linear operator A_x is unique for each $x \in D$. Thus, let $\mathcal{D}\Phi : D \to L(\mathcal{V}, \mathcal{W})$ be a function such that $\mathcal{D}\Phi(x) = A_x$. The function Φ is called continuously Fréchet differentiable on D if Φ is Fréchet differentiable on D and $\mathcal{D}\Phi$ is continuous on D. When $D = \mathcal{V}$, the function Φ is called continuously Fréchet differentiable if Φ is continuously Fréchet differentiable on $\mathcal{V}$. If $\Phi : D \subset \mathcal{V} \to \mathbf{R}$ is Fréchet differentiable at $x \in D$, then $\mathcal{D}\Phi(x)$ is a bounded linear operator such that $\mathcal{D}\Phi(x) : \Delta x \mapsto \langle \nabla\Phi(x), \Delta x \rangle$, where $\nabla\Phi(x) \in \mathcal{V}$. Then, we call $\nabla\Phi(x)$ a gradient of Φ at x. In particular, when $\mathcal{V} = \mathbf{R}$, $\nabla\Phi(x) = \Phi'(x)$, where $\Phi'(x)$ denotes a derivative of Φ at x; when $\mathcal{V} = \mathbf{R}^n$,

$$\nabla\Phi(x) = \begin{bmatrix} \frac{\partial}{\partial x_1}\Phi(x) \\ \vdots \\ \frac{\partial}{\partial x_n}\Phi(x) \end{bmatrix},$$

where $\frac{\partial}{\partial x_i}\Phi(x)$ denotes a partial derivative of Φ at x with respect to its i-th component. In addition, if $\Phi : D \subset \mathbf{R}^n \to \mathbf{R}^m$ is a function such that $\Phi(x) := [\Phi_1(x), \ldots, \Phi_m(x)]^\top$ and it is Fréchet differentiable at $x \in D$, then $\mathcal{D}\Phi(x)$ is a bounded linear operator such that $\mathcal{D}\Phi(x)$:

$\Delta x \mapsto J_\Phi(x)\Delta x$, where

$$J_\Phi(x) = \begin{bmatrix} \frac{\partial}{\partial x_1}\Phi_1(x) & \frac{\partial}{\partial x_2}\Phi_1(x) & \cdots & \frac{\partial}{\partial x_n}\Phi_1(x) \\ \frac{\partial}{\partial x_1}\Phi_2(x) & \frac{\partial}{\partial x_2}\Phi_2(x) & \cdots & \frac{\partial}{\partial x_n}\Phi_2(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1}\Phi_m(x) & \frac{\partial}{\partial x_2}\Phi_m(x) & \cdots & \frac{\partial}{\partial x_n}\Phi_m(x) \end{bmatrix}.$$

Then, we call $J_\Phi(x)$ a Jacobian of Φ at x. If $\Phi : D \subset \mathbf{R}^n \to \mathbf{R}$ is twice Fréchet differentiable at $x \in D$, $\nabla^2\Phi(x)$ denotes a Jacobian of $\nabla\Phi$ at x, that is,

$$\nabla^2\Phi(x) = \begin{bmatrix} \frac{\partial^2}{\partial x_1^2}\Phi(x) & \frac{\partial^2}{\partial x_2 \partial x_1}\Phi(x) & \cdots & \frac{\partial}{\partial x_n \partial x_1}\Phi(x) \\ \frac{\partial^2}{\partial x_1 \partial x_2}\Phi(x) & \frac{\partial^2}{\partial x_2^2}\Phi(x) & \cdots & \frac{\partial^2}{\partial x_n \partial x_2}\Phi(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_1 \partial x_n}\Phi(x) & \frac{\partial^2}{\partial x_2 \partial x_n}\Phi(x) & \cdots & \frac{\partial^2}{\partial x_n^2}\Phi(x) \end{bmatrix}.$$

Then, we call $\nabla^2\Phi(x)$ a Hessian of Φ at x. Let $\Psi : D \times E \to \mathcal{X}$ be a function, where $D \subset \mathcal{V}$ and $E \subset \mathcal{W}$ are sets, and let $y \in E$. If $\mathcal{X} = \mathbf{R}$ and $\Psi(\cdot, y) : D \to \mathbf{R}$ is Fréchet differentiable at $x \in D$, $\nabla_x\Psi(x, y)$ denotes a gradient of Ψ at (x, y) with respect to x. If $\mathcal{X} = \mathbf{R}^m$ and $\Psi(\cdot, y) : D \to \mathbf{R}^m$ is Fréchet differentiable at $x \in D$, $\partial_x\Psi(x, y)$ denotes a Jacobian of Ψ at (x, y) with respect to x. If $\mathcal{X} = \mathbf{R}$ and $\Psi(\cdot, y) : D \to \mathbf{R}$ is twice Fréchet differentiable at $x \in D$, $\nabla^2_{xx}\Psi(x, y)$ denotes a Hessian of Ψ at (x, y) with respect to x. Unless otherwise noted, differentiable means Fréchet differentiable.

Finally, in what follows, we list other notations that appear in the thesis.

- $\emptyset$: the empty set
- cl$S$: the closure of a set S (the smallest closed set containing S)
- int$S$: the interior of a set S (the largest open set contained in S)
- $\log(x)$: the natural logarithm of a positive real number x
- $\exp(x)$: e (Napier's constant) raised to the power of a real number x
- $\operatorname{argmin}\{h(x)|x \in D\}$: the set of minimizers of a function h over a nonempty set D
- $\operatorname{argmax}\{h(x)|x \in D\}$: the set of maximizers of a function h over a nonempty set D

2.2 Fundamental mathematics

2.2.1 Linear algebra and analysis

We present some well-known facts which are exploited in the thesis. First, we give some results related to linear algebra.

Proposition 2.2.1. *[5, 24, 30, 54] The following statements hold.*

(a) *Let $A \in \mathbf{R}^{m\times n}$. Then, $\|A\|_2 \le \|A\|_F \le \sqrt{n}\|A\|_2$.*

(b) *Let* $A, B \in \mathbf{S}^n$ *be matrices such that* $0 \prec B \preceq A$. *Then,* $\det B \leq \det A$.

(c) *Let* $A, B \in \mathbf{S}^n$. *Then,* $\lambda_{\min}(A)\mathrm{tr}(B) \leq \mathrm{tr}(AB)$.

(d) *Let* $A \in \mathbf{S}^n$ *be a matrix such that* $\|A\|_F < 1$. *Then,* $I - A$ *is nonsingular. Moreover,* $\|(I-A)^{-1}\|_F \leq \frac{n}{1-\|A\|_F}$.

(e) *Let* $A_1, \ldots A_m \in \mathbf{S}^n$ *be matrices such that they commute mutually. Then, there exists an orthogonal matrix* $U \in \mathbf{R}^{n\times n}$ *such that* $U^\top A_i U = \mathrm{diag}[\lambda_1(A_i), \ldots, \lambda_n(A_i)]$ *for all* $i = 1, \ldots, m$. □

Secondly, we give some results associated with analysis.

Proposition 2.2.2. *[46] The following statements hold.*

(a) *Let* $\Phi : D \subset \mathbf{R}^n \to \mathbf{R}^m$ *be continuously differentiable on a convex set* $D_0 \subset D$. *Suppose that there exists* $L > 0$ *such that*

$$\|J_\Phi(u) - J_\Phi(v)\|_F \leq L\|u - v\| \quad \text{for all } u,\ v \in D_0.$$

Then, we have

$$\|\Phi(y) - \Phi(x) - J_\Phi(x)(y-x)\| \leq \frac{L}{2}\|x - y\|^2 \quad \text{for all } x,\ y \in D_0.$$

(b) *Let* $\Psi : D \subset \mathbf{R} \to \mathbf{R}$ *be twice continuously differentiable on a bounded convex set* $D_0 \subset D$. *Then, we have*

$$|\Psi(u) - \Psi(v) - \Psi'(v)(u-v)| \leq C|u-v|^2 \quad \text{for all } u,\ v \in D_0,$$

where $C := \sup\{\ |\Psi'(x)|\ |\ x \in D_0\ \}$. □

Finally, we give the mean value theorem and the implicit function theorem.

Theorem 2.2.1. *[46] Let* $\Phi : D \subset \mathbf{R}^n \to \mathbf{R}$ *be differentiable on a convex set* $D_0 \subset D$. *Then, for any* $x, y \in D_0$, *there exists* $t \in (0,1)$ *such that* $\Phi(y) - \Phi(x) = \langle \nabla\Phi(tx + (1-t)y), y - x\rangle$. □

Theorem 2.2.2. *[46] Suppose that* $\Phi : D \subset \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n$ *is continuous on an open neighborhood* $D_0 \subset D$ *of a point* (x_0, y_0) *such that* $\Phi(x_0, y_0) = 0$. *Suppose also that* $\partial_x\Phi$ *exists in a neighborhood of* (x_0, y_0) *and is continuous at* (x_0, y_0) *and* $\partial_x\Phi(x_0, y_0)$ *is nonsingular. Then, there exist open neighborhoods* $P \subset \mathbf{R}^n$ *and* $Q \subset \mathbf{R}^m$ *of* x_0 *and* y_0, *respectively, such that, for any* $y \in \mathrm{cl}Q$, *the equation* $\Phi(x, y) = 0$ *has a unique solution* $x = \Psi(y) \in \mathrm{cl}P$, *and the function* $\Psi : Q \to \mathbf{R}^n$ *is continuous on* Q. *Moreover, if* $\partial_y\Phi(x_0, y_0)$ *exists, then* Ψ *is differentiable at* y_0 *and* $J_\Psi(y_0) = -\left[\partial_x\Phi(x_0, y_0)\right]^{-1}\partial_y\Phi(x_0, y_0)$. □

2.2.2 Convex analysis

In this section, we provide some properties related to convex analysis. To begin with, we define the convexity of sets and functions. Let $\mathcal{V}$ be an inner product space, such as $\mathbf{R}^n$, $\mathbf{R}^{m\times n}$ and $\mathbf{S}^p$. A set $C \subset \mathcal{V}$ is called convex if

$$\lambda x + (1-\lambda)y \in C \quad \text{for all } \lambda \in [0,1] \text{ and } x,\ y \in C.$$

Let $C \subset \mathcal{V}$ be a convex set, and let $f : C \to \mathbf{R}$ be a function. The function f is called convex on C if

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) \quad \text{for all } \lambda \in [0,1] \text{ and } x,\ y \in C.$$

The function f is called strictly convex on C if

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y) \quad \text{for all } \lambda \in (0,1) \text{ and } x,\ y \in C \text{ such that } x \ne y.$$

When $C = \mathcal{V}$, the function f is called (strictly) convex if f is (strictly) convex on $\mathcal{V}$. The function f is called (strictly) concave on C if $-f$ is (strictly) convex on C. When $C = \mathcal{V}$, the function f is called (strictly) concave if $-f$ is (strictly) convex on $\mathcal{V}$.

The following proposition gives a necessary and sufficient condition for a differentiable function to be convex.

Proposition 2.2.3. *[7, 52] Let $C \subset \mathcal{V}$ be a nonempty open convex set and let $f : C \to (-\infty, \infty]$ be a differentiable function on C. Then, the function f is convex on C if and only if*

$$\langle \nabla f(x), y - x \rangle \le f(y) - f(x) \quad \text{for all } x,\ y \in C.$$

Moreover, the function f is strictly convex on C if and only if the above inequality is strict whenever $x \ne y$. □

Next, we define the effective domain, the properness and the directional differentiability of functions. Let $S \subset \mathcal{V}$ be a set, and let $f : S \to \mathbf{R}$ be a function. We define the effective domain of f by

$$\mathrm{dom} f := \{\ x \in \mathcal{V} \mid f(x) < \infty\ \}.$$

We say that the function f is proper if $\mathrm{dom} f \ne \emptyset$. For any $x \in \mathrm{dom} f$ and $d \in \mathcal{V}$, we define the (lower) directional derivative of f at x in the direction d by

$$f'(x;d) := \liminf_{\lambda \to +0} \frac{f(x+\lambda d) - f(x)}{\lambda}.$$

The next proposition provides several properties of the directional derivative for convex functions.

Proposition 2.2.4. *[7, 52] Let $f : \mathcal{V} \to (-\infty, \infty]$ be a proper convex function. Moreover, let $x \in \mathrm{dom} f$ and $d \in \mathcal{V}$ with $d \ne 0$. Then, the difference quotient in the definition of $f'(x;d)$ is a nondecreasing function of $\lambda > 0$, so that*

$$f'(x;d) = \lim_{\lambda \to +0} \frac{f(x+\lambda d) - f(x)}{\lambda} = \inf_{\lambda > 0} \frac{f(x+\lambda d) - f(x)}{\lambda}.$$

Furthermore, if there exists $\varepsilon > 0$ such that $x + td \in \mathrm{dom} f$ for all $t \in (0, \varepsilon]$, then $f'(x;d)$ is finite. □

We consider the following optimization problem:

$$\begin{array}{ll} \text{minimize} & f(x), \\ \text{subject to} & x \in S, \end{array} \tag{2.2.1}$$

where $f : \mathcal{V} \to (-\infty, \infty]$ is a proper function and $S \subset \mathcal{V}$ is a nonempty closed set. By using the directional derivative of f, we provide a necessary condition for optimality in problem (2.2.1).

Proposition 2.2.5. *Let f and S be the function and the set in problem (2.2.1), respectively. Suppose that S is a convex set. Then, if $\overline{x} \in S$ is a local minimum of problem (2.2.1),*

$$f'(\overline{x}; x - \overline{x}) \geq 0 \quad \text{for all } x \in S. \tag{2.2.2}$$

Proof. Let $x \in S$ be arbitrary. It follows from the convexity of S that $\overline{x} + \lambda(x - \overline{x}) \in S$ for any $\lambda \in [0, 1]$. Then, since $\overline{x} \in S$ is a local minimum of problem (2.2.1), there exists $\varepsilon \in (0, 1]$ such that $f(\overline{x} + t(x - \overline{x})) - f(\overline{x}) \geq 0$ for all $t \in (0, \varepsilon]$. Dividing both sides by $t \in (0, \varepsilon]$, we obtain

$$\frac{f(\overline{x} + t(x - \overline{x})) - f(\overline{x})}{t} \geq 0 \quad \text{for all } t \in (0, \varepsilon].$$

Therefore, we have from $t \to +0$ that $f'(\overline{x}; x - \overline{x}) \geq 0$ for all $x \in S$. □

From Proposition 2.2.5, we see that (2.2.2) is a necessary condition for optimality in problem (2.2.1). In the remainder of this thesis, we say that $\overline{x} \in S$ is a stationary point of problem (2.2.1) if $\overline{x}$ satisfies the condition (2.2.2). Note that if f is differentiable at $\overline{x}$ and $S = \mathcal{V}$, (2.2.2) is equivalent to $\nabla f(\overline{x}) = 0$. Moreover, if problem (2.2.1) is convex, then (2.2.2) is a necessary and sufficient condition for optimality in problem (2.2.1). We show this fact in the next proposition.

Proposition 2.2.6. *Let f and S be the function and the set in problem (2.2.1), respectively. Suppose that S is a convex set, and f is a proper convex function on S. Suppose also that (2.2.1) has a nonempty optimal solution set. Then, a stationary of (2.2.1) is also a global minimum of (2.2.1). In addition, if the function f is strictly convex on S, then the global minimum of (2.2.1) is unique.*

Proof. We show the first part of this proposition. By Proposition 2.2.5, it suffices to show that (2.2.2) is a sufficient condition for optimality in problem (2.2.1). Suppose that $\overline{x}$ is a stationary point of (2.2.1), that is, $\overline{x}$ satisfies (2.2.2). Let $x \in S$ be arbitrary. It follows from Proposition 2.2.4 that

$$0 \leq f'(\overline{x}; x - \overline{x}) = \inf_{\lambda > 0} \frac{f(\overline{x} + \lambda(x - \overline{x})) - f(\overline{x})}{\lambda} \leq \frac{f(\overline{x} + \lambda(x - \overline{x})) - f(\overline{x})}{\lambda} \quad \text{for all } \lambda > 0.$$

Then, we have from $\lambda = 1$ that $f(\overline{x}) \leq f(x)$ for any $x \in S$. Therefore, $\overline{x}$ is a global minimum of problem (2.2.1).

We show the second part of this proposition by contradiction. Suppose that f is strictly convex on S. Suppose also that $\overline{x}$ and $\widetilde{x}$ are two distinct global minima of problem (2.2.1). Let α be an optimal value of (2.2.1), and let $\lambda \in (0, 1)$. Since the function f is strictly convex on S, we have

$$f(\lambda\overline{x} + (1 - \lambda)\widetilde{x}) < \lambda f(\overline{x}) + (1 - \lambda) f(\widetilde{x}) = \alpha. \tag{2.2.3}$$

It follows from the convexity of S that $\lambda\overline{x} + (1-\lambda)\widetilde{x} \in S$, and hence we have by (2.2.3) that $\lambda\overline{x} + (1-\lambda)\widetilde{x}$ is a global minimum of (2.2.1). Therefore, we obtain a contradiction. □

Finally, we consider the following unconstrained optimization problem by using f and S in problem (2.2.1):

$$\begin{array}{ll} \text{minimize} & f(x) + \delta_S(x), \\ \text{subject to} & x \in \mathcal{V}, \end{array} \tag{2.2.4}$$

where the function $\delta_S : \mathcal{V} \to (-\infty, \infty]$ is an indicator function of S, that is,

$$\delta_S(x) := \begin{cases} 0 & \text{if } x \in S, \\ +\infty & \text{otherwise.} \end{cases}$$

We give a property associated with a stationary point of problems (2.2.1) and (2.2.4).

Proposition 2.2.7. *Let f and S be the function and the set in problem (2.2.1), respectively. Suppose that S is a convex set. If $\overline{x} \in S$ is a stationary point of problem (2.2.4), then $\overline{x}$ is that of problem (2.2.1).*

Proof. Since $\overline{x} \in S$ is a stationary point of problem (2.2.4),

$$\liminf_{\lambda \to +0} \frac{f(\overline{x} + \lambda(x - \overline{x})) - f(\overline{x}) + \delta_S(\overline{x} + \lambda(x - \overline{x})) - \delta_S(\overline{x})}{\lambda} \geq 0 \quad \text{for all } x \in \mathcal{V}.$$

It then follows from $S \subset \mathcal{V}$ that

$$\liminf_{\lambda \to +0} \frac{f(\overline{x} + \lambda(x - \overline{x})) - f(\overline{x}) + \delta_S(\overline{x} + \lambda(x - \overline{x})) - \delta_S(\overline{x})}{\lambda} \geq 0 \quad \text{for all } x \in S. \tag{2.2.5}$$

On the other hand, we obtain $\delta_S(y + t(z - y)) = 0$ for all $y,\ z \in S$ and $t \in [0, 1]$ by the convexity of S. Thus, we have from (2.2.5) that

$$\liminf_{\lambda \to +0} \frac{f(\overline{x} + \lambda(x - \overline{x})) - f(\overline{x})}{\lambda} \geq 0 \quad \text{for all } x \in S.$$

Therefore, $\overline{x}$ is a stationary point of problem (2.2.1). □

2.2.3 Symmetrized Kronecker product and its properties

In this section, we define the following notations.

(i) We define a partial derivative of the function $X : \mathbf{R}^n \to \mathbf{S}^p$ at x with respect to its i-th component as $A_i(x) := \frac{\partial}{\partial x_i} X(x)$.

(ii) Let $x \in \mathbf{R}^n$. We define an operator $\mathcal{A}(x) : \mathbf{R}^n \to \mathbf{S}^p$ as

$$\mathcal{A}(x)w := w_1 A_1(x) + \ldots + w_n A_n(x) \quad \text{for all } w \in \mathbf{R}^n.$$

(iii) Let $x \in \mathbf{R}^n$. We define an adjoint operator of $\mathcal{A}(x) : \mathbf{R}^n \to \mathbf{S}^p$ as $\mathcal{A}^*(x) : \mathbf{S}^p \to \mathbf{R}^n$, i.e.,

$$\mathcal{A}^*(x)U = [\langle A_1(x), U\rangle, \ldots, \langle A_n(x), U\rangle]^\top \quad \text{for all } U \in \mathbf{S}^p.$$

(iv) Let $P,\ Q \in \mathbf{R}^{p\times p}$. We define an operator $P \odot Q : \mathbf{S}^p \to \mathbf{S}^p$ as

$$(P \odot Q)U := \frac{1}{2}(PUQ^\top + QUP^\top) \quad \text{for all } U \in \mathbf{S}^p.$$

(v) We define an operator $\mathrm{svec} : \mathbf{S}^p \to \mathbf{R}^{\frac{p(p+1)}{2}}$ as

$$\mathrm{svec}(U) = [U_{11}, \sqrt{2}U_{21}, \ldots, \sqrt{2}U_{p1}, U_{22}, \sqrt{2}U_{32}, \ldots, \sqrt{2}U_{p2}, U_{33}, \ldots, U_{pp}]^\top \text{ for all } U \in \mathbf{S}^p.$$

(vi) Let $P,\ Q \in \mathbf{R}^{p\times p}$. We denote the symmetrized Kronecker product as $P \otimes_S Q : \mathbf{R}^{\frac{p(p+1)}{2}} \to \mathbf{R}^{\frac{p(p+1)}{2}}$ which satisfies that

$$(P \otimes_S Q)\mathrm{svec}(U) = \mathrm{svec}((P \odot Q)U) \quad \text{for all } U \in \mathbf{S}^p.$$

(vii) We define an operator $A : \mathbf{R}^n \to \mathbf{R}^{\frac{p(p+1)}{2}\times n}$ as

$$A(x) := [\mathrm{svec}(A_1(x)), \ldots, \mathrm{svec}(A_n(x))] \quad \text{for all } x \in \mathbf{R}^n.$$

(viii) We define

$$U \circ V := \frac{UV + VU}{2} \quad \text{for all } U,\ V \in \mathbf{S}^p.$$

In the following, we give some propositions related to the above definitions.

Proposition 2.2.8. *[61, 72] The following statements hold.*

(a) *For any matrices $U,\ V \in \mathbf{S}^p$,*

$$\langle U, V\rangle = \mathrm{tr}(UV) = \mathrm{svec}(U)^\top \mathrm{svec}(V), \quad \|U\|_F = \|\mathrm{svec}(U)\|.$$

(b) *For any matrices $U,\ V \in \mathbf{S}^p_+$ and $\mu \in \mathbf{R}$, $U \circ V = \mu I$ is equivalent to $UV = \mu I$.* □

Proposition 2.2.9. *[61, 72] Let P and Q be arbitrary nonsingular matrices in $\mathbf{R}^{p\times p}$. Then the following statements hold.*

(a) *The operator $P \odot Q$ is invertible.*

(b) *For all $U,\ V \in \mathbf{S}^p$,*

$$\langle U, (P \odot Q)V\rangle = \langle (P^\top \odot Q^\top)U, V\rangle, \quad \left\langle U, (P \odot Q)^{-1}V\right\rangle = \langle (P^\top \odot Q^\top)^{-1}U, V\rangle.$$

(c) $(P \odot P)^{-1} = (P^{-1} \odot P^{-1})$. □

Proposition 2.2.10. *Let P and Q be arbitrary matrices in $\mathbf{R}^{p\times p}$, and let $C_1 := \sqrt{\frac{p(p+1)}{2}}$. Then, $\|P \otimes_S Q\|_F \le C_1\|P\|_F\|Q\|_F$.*

Proof. It follows from Proposition 2.2.1 (a) that

$$\|P \otimes_S Q\|_F \leq C_1 \|P \otimes_S Q\|_2. \tag{2.2.6}$$

Let $U \in \mathbf{S}^p$. The definition of the symmetrized Kronecker product and Proposition 2.2.8 (a) yield that

$$\begin{aligned} \|(P \otimes_S Q)\mathrm{svec}(U)\| &= \|\mathrm{svec}((P \odot Q)U)\| \\ &= \|(P \odot Q)U\|_F \\ &= \frac{1}{2}\|PUQ^\top + QUP^\top\|_F \\ &\leq \|P\|_F \|Q\|_F \|U\|_F, \end{aligned}$$

and hence

$$\|P \otimes_S Q\|_2 = \sup_{\mathrm{svec}(U)\neq 0} \frac{\|(P \otimes_S Q)\mathrm{svec}(U)\|}{\|\mathrm{svec}(U)\|} \leq \sup_{U\neq 0} \frac{\|P\|_F \|Q\|_F \|U\|_F}{\|U\|_F} = \|P\|_F \|Q\|_F. \tag{2.2.7}$$

We have by (2.2.6) and (2.2.7) that $\|P \otimes_S Q\|_F \leq C_1 \|P\|_F \|Q\|_F$. □

Several interior point methods for SDP employ scaling of $X(x)$ and Z, where $Z \in \mathbf{S}^p$ corresponds to a Lagrange multiplier matrix for $X(x) \succeq 0$ in (1.1.1). The details of Z are given in Section 2.3. Let T be a nonsingular matrix in $\mathbf{R}^{p\times p}$. We consider the scaled matrices $\widetilde{X}(x)$ and $\widetilde{Z}$ defined by

$$\widetilde{X}(x) := (T \odot T)X(x) \quad \text{and} \quad \widetilde{Z} := (T^{-\top} \odot T^{-\top})Z.$$

The details of scaling are given in Section 2.4. In the following, we show some useful properties of $\widetilde{X}(x)$ and $\widetilde{Z}$.

Proposition 2.2.11. *The following statements hold.*

(a) *Suppose that $X(x)$ and Z are symmetric positive definite. Suppose also that $\widetilde{X}(x)$ and $\widetilde{Z}$ commute. Then we have*

$$\left\langle (\widetilde{Z} \odot I)(\widetilde{X}(x) \odot I)U, U \right\rangle \geq 0 \quad \text{for all } U \in \mathbf{S}^p.$$

Furthermore, the strict inequality holds in the above if and only if $U \neq 0$.

(b) *Suppose that $\widetilde{X}(x)$ and $\widetilde{Z}$ commute. Then we have*

$$(\widetilde{X}(x) \odot I)(\widetilde{Z} \odot I) = (\widetilde{Z} \odot I)(\widetilde{X}(x) \odot I).$$

Proof. (a) Since the matrices $X(x)$ and Z are symmetric positive definite, $\widetilde{X}(x)$ and $\widetilde{Z}$ are also symmetric positive definite. It then follows from the commutativity of $\widetilde{X}(x)$ and $\widetilde{Z}$ that $\widetilde{X}(x)\widetilde{Z}$ is symmetric positive definite. Thus, there exists $(\widetilde{X}(x)\widetilde{Z})^{\frac{1}{2}} \in \mathbf{S}^p_{++}$ such that $\widetilde{X}(x)\widetilde{Z} =$

$(\widetilde{X}(x)\widetilde{Z})^{\frac{1}{2}}(\widetilde{X}(x)\widetilde{Z})^{\frac{1}{2}}$. Let $U \in \mathbf{S}^p$. Then we have

$$\begin{aligned}
\left\langle (\widetilde{Z}\odot I)(\widetilde{X}(x)\odot I)U, U\right\rangle &= \frac{1}{4}\mathrm{tr}((\widetilde{Z}\widetilde{X}(x)U + \widetilde{X}(x)U\widetilde{Z} + \widetilde{Z}U\widetilde{X}(x) + U\widetilde{X}(x)\widetilde{Z})U) \\
&= \frac{1}{4}\mathrm{tr}(\widetilde{X}(x)U\widetilde{Z}U) + \frac{1}{4}\mathrm{tr}(\widetilde{Z}U\widetilde{X}(x)U) \\
&\qquad + \frac{1}{4}\mathrm{tr}(U\widetilde{X}(x)\widetilde{Z}U) + \frac{1}{4}\mathrm{tr}(U\widetilde{Z}\widetilde{X}(x)U) \\
&= \frac{1}{4}\mathrm{tr}(\widetilde{X}(x)^{\frac{1}{2}}U\widetilde{Z}^{\frac{1}{2}}\widetilde{Z}^{\frac{1}{2}}U\widetilde{X}(x)^{\frac{1}{2}}) + \frac{1}{4}\mathrm{tr}(\widetilde{Z}^{\frac{1}{2}}U\widetilde{X}(x)^{\frac{1}{2}}\widetilde{X}(x)^{\frac{1}{2}}U\widetilde{Z}^{\frac{1}{2}}) \\
&\qquad + \frac{1}{4}\mathrm{tr}(U\widetilde{X}(x)\widetilde{Z}U) + \frac{1}{4}\mathrm{tr}(U\widetilde{X}(x)\widetilde{Z}U) \\
&= \frac{1}{2}\mathrm{tr}(\widetilde{X}(x)^{\frac{1}{2}}U\widetilde{Z}^{\frac{1}{2}}\widetilde{Z}^{\frac{1}{2}}U\widetilde{X}(x)^{\frac{1}{2}}) + \frac{1}{2}\mathrm{tr}(U(\widetilde{X}(x)\widetilde{Z})^{\frac{1}{2}}(\widetilde{X}(x)\widetilde{Z})^{\frac{1}{2}}U) \\
&= \frac{1}{2}\|\widetilde{X}(x)^{\frac{1}{2}}U\widetilde{Z}^{\frac{1}{2}}\|_F^2 + \frac{1}{2}\|(\widetilde{X}(x)\widetilde{Z})^{\frac{1}{2}}U\|_F^2 \\
&\geq 0,
\end{aligned}$$

where the third equality follows from the commutativity of $\widetilde{X}(x)$ and $\widetilde{Z}$. Note that, since $\widetilde{X}(x)^{\frac{1}{2}}$, $\widetilde{Z}^{\frac{1}{2}}$ and $(\widetilde{X}(x)\widetilde{Z})^{\frac{1}{2}}$ are positive definite, the strict inequality holds in the above if and only if $U \neq 0$.

(b) For any $U \in \mathbf{S}^p$, we have

$$\begin{aligned}
(\widetilde{X}(x)\odot I)(\widetilde{Z}\odot I)U &= \frac{1}{4}(\widetilde{X}(x)\widetilde{Z}U + \widetilde{Z}U\widetilde{X}(x) + \widetilde{X}(x)U\widetilde{Z} + U\widetilde{Z}\widetilde{X}(x)) \\
&= \frac{1}{4}(\widetilde{Z}\widetilde{X}(x)U + \widetilde{X}(x)U\widetilde{Z} + \widetilde{Z}U\widetilde{X}(x) + U\widetilde{X}(x)\widetilde{Z}) \\
&= (\widetilde{Z}\odot I)(\widetilde{X}(x)\odot I)U,
\end{aligned}$$

where the second equality follows from the commutativity of $\widetilde{X}(x)$ and $\widetilde{Z}$. Hence, we obtain $(\widetilde{X}(x)\odot I)(\widetilde{Z}\odot I) = (\widetilde{Z}\odot I)(\widetilde{X}(x)\odot I)$. □

2.2.4 Properties of a log-determinant function

Let Ω be a set defined by $\Omega := \{\ x \in \mathbf{R}^n \mid X(x) \succ 0\ \}$. Furthermore, let $\phi : \mathbf{S}^p_{++} \to \mathbf{R}$ and $\varphi : \Omega \to \mathbf{R}$ be functions defined by $\phi(M) := -\log\det M$ and $\varphi(x) := \phi(X(x))$, respectively. We first give the differentiability and convexity of ϕ and φ.

Proposition 2.2.12. *[60] The following statements hold.*

(a) *The function ϕ is differentiable on $\mathbf{S}^p_{++}$, and $\nabla\phi(M) = -M^{-1}$ for all $M \in \mathbf{S}^p_{++}$.*

(b) *The function ϕ is strictly convex on $\mathbf{S}^p_{++}$.* □

Proposition 2.2.13. *The following statements hold.*

(a) *The function φ is differentiable on Ω, and $\nabla\varphi(x) = -\mathcal{A}^*(x)X(x)^{-1}$ for all $x \in \Omega$.*

(b) *Suppose that X is nondifferentiable on Ω, and satisfies that*

$$X(\lambda u + (1-\lambda)v) - \lambda X(u) - (1-\lambda)X(v) \succeq 0 \quad \text{for all } \lambda \in [0,1] \text{ and } u,\ v \in \Omega. \quad (2.2.8)$$

Then φ is convex on Ω. Moreover, if X is injective on Ω, then φ is strictly convex.

(c) *Suppose that X is differentiable on Ω, and satisfies (2.2.8). Suppose also that $A_1(x), \ldots, A_n(x)$ are linearly independent for all $x \in \Omega$. Then φ is strictly convex.*

Proof. We have from Proposition 2.2.12 (a) and the chain rule that

$$\nabla\varphi(x) = -\mathcal{A}^*(x)X(x)^{-1}. \quad (2.2.9)$$

(b) It follows from $\lambda X(u) + (1-\lambda)X(v) \succ 0$, (2.2.8) and Proposition 2.2.1 (b) that

$$\det[\lambda X(u) + (1-\lambda)X(v)] \leq \det[X(\lambda u + (1-\lambda)v)].$$

Since $-\log$ is a decreasing function on $(0,\infty)$ and ϕ is strictly convex from Proposition 2.2.12 (b), we have

$$\begin{aligned}
\varphi(\lambda u + (1-\lambda)v) &= -\log\det[X(\lambda u + (1-\lambda)v)] \\
&\leq -\log\det[\lambda X(u) + (1-\lambda)X(v)] \\
&= \phi(\lambda X(u) + (1-\lambda)X(v)) \\
&\leq \lambda\phi(X(u)) + (1-\lambda)\phi(X(v)) \\
&= \lambda\varphi(u) + (1-\lambda)\varphi(v),
\end{aligned}$$

and hence φ is convex on Ω.

Suppose that $u \neq v$. Then, since X is injective on Ω, $X(u) \neq X(v)$. Moreover, since ϕ is strictly convex,

$$\begin{aligned}
\varphi(\lambda u + (1-\lambda)v) &\leq \phi(\lambda X(u) + (1-\lambda)X(v)) \\
&< \lambda\phi(X(u)) + (1-\lambda)\phi(X(v)) \\
&= \lambda\varphi(u) + (1-\lambda)\varphi(v)
\end{aligned}$$

for $\lambda \in (0,1)$. Thus, φ is strictly convex.
(c) Since X is differentiable, $X(v + \lambda(u-v)) - X(v) = \lambda\mathcal{A}(v)(u-v) + o(\lambda)$ for $u, v \in \Omega$ and $\lambda \in (0,1)$. Then (2.2.8) can be written as $\lambda\mathcal{A}(v)(u-v) - \lambda(X(u) - X(v)) + o(\lambda) \succeq 0$. Dividing both sides by λ, we have $\mathcal{A}(v)(u-v) - X(u) + X(v) + \frac{o(\lambda)}{\lambda} \succeq 0$. Letting $\lambda \to 0$ yields

$$\mathcal{A}(v)(u-v) - X(u) + X(v) \succeq 0.$$

Let $M := \mathcal{A}(v)(u-v) - X(u) + X(v)$. Since $M \in \mathbf{S}^p_+$ and $X(v)^{-1} \in \mathbf{S}^p_{++}$, there exist $M^{\frac{1}{2}}$ and $X(v)^{-\frac{1}{2}}$. Then we have

$$\left\langle X(v)^{-1}, M \right\rangle = \operatorname{tr}(X(v)^{-1}M) = \operatorname{tr}(X(v)^{-\frac{1}{2}}M^{\frac{1}{2}}M^{\frac{1}{2}}X(v)^{-\frac{1}{2}}) = \|M^{\frac{1}{2}}X(v)^{-\frac{1}{2}}\|_F^2.$$

It then follows from the definition of φ, Proposition 2.2.12 (a) and (b) that

$$\begin{aligned}\varphi(u) - \varphi(v) &= \phi(X(u)) - \phi(X(v)) \\ &\geq \left\langle -X(v)^{-1}, X(u) - X(v) \right\rangle \\ &= \left\langle X(v)^{-1}, M \right\rangle + \left\langle X(v)^{-1}, -\mathcal{A}(v)(u-v) \right\rangle \\ &= \|M^{\frac{1}{2}} X(v)^{-\frac{1}{2}}\|_F^2 + \left\langle -\mathcal{A}^*(v)X(v)^{-1}, u-v \right\rangle \\ &\geq \left\langle \nabla\varphi(v), u-v \right\rangle, \end{aligned} \tag{2.2.10}$$

where the last inequality follows from (2.2.9).

Since φ is convex by (b), it suffices to show that $\varphi(u) - \varphi(v) = \langle \nabla\varphi(v), u-v \rangle$ if and only if $u = v$. If $u = v$, it is clear that $\varphi(u) - \varphi(v) = \langle \nabla\varphi(v), u-v \rangle$. Conversely, suppose that $\varphi(u) - \varphi(v) = \langle \nabla\varphi(v), u-v \rangle$. Since the equality holds in (2.2.10), we see that

$$\phi(X(u)) - \phi(X(v)) = \left\langle -X(v)^{-1}, X(u) - X(v) \right\rangle, \quad \|M^{\frac{1}{2}} X(v)^{-\frac{1}{2}}\|_F = 0. \tag{2.2.11}$$

We have from Proposition 2.2.12 (b) and the first equality of (2.2.11) that $X(u) = X(v)$, that is, $M = \mathcal{A}(v)(u-v)$ by the definition of M. Then, the regularity of $X(v)^{-\frac{1}{2}}$ and the second equality of (2.2.11) yield that $0 = M = \mathcal{A}(v)(u-v)$. Since $A_1(x), \ldots, A_n(x)$ are linearly independent for all $x \in \Omega$, we have $u = v$. □

Note that Proposition 2.2.13 (b) does not assume the differentiability of X.

We next show that matrices in a level set of ϕ are uniformly positive definite.

Proposition 2.2.14. *For a given $\gamma \in \mathbf{R}$, let $\mathcal{L}_\phi(\gamma) = \{U \in \mathbf{S}^p_{++} | \phi(U) \leq \gamma\}$. Let Γ be a bounded subset of $\mathbf{S}^p$. Then, there exists $\underline{\lambda} > 0$ such that $\lambda_{\min}(U) \geq \underline{\lambda}$ for all $U \in \mathcal{L}_\phi(\gamma) \cap \Gamma$.*

Proof. Suppose the contrary, that is, there exists a sequence $\{U_j\} \subset \mathcal{L}_\phi(\gamma) \cap \Gamma$ such that $\lambda_{\min}(U_j) \to 0$ as $j \to \infty$. Then

$$-\log \lambda_{\min}(U_j) \to \infty \quad (j \to \infty). \tag{2.2.12}$$

Since $U_j \in \mathcal{L}_\phi(\gamma)$, we have $\gamma \geq \phi(U_j) = -\log\det U_j = -\sum_{i=1}^p \log\lambda_i(U_j)$. It then follows from (2.2.12) that there exist an index k and an infinite subset $\mathcal{J}$ such that $\lim_{j\to\infty, j\in\mathcal{J}} -\log\lambda_k(U_j) = -\infty$, that is, $\lim_{j\to\infty, j\in\mathcal{J}} \lambda_k(U_j) = \infty$. However, this is contrary to the boundedness of $\{U_j\}$. Therefore, there exists $\underline{\lambda} > 0$ such that $\lambda_{\min}(U) \geq \underline{\lambda}$ for all $U \in \mathcal{L}_\phi(\gamma) \cap \Gamma$. □

2.3 Some optimality conditions for nonlinear SDP

We first introduce the first-order optimality conditions for nonlinear SDP (1.1.1). The Lagrangian function L of (1.1.1) is given by

$$L(x, y, Z) := f(x) - g(x)^\top y - \langle X(x), Z \rangle,$$

where $y \in \mathbf{R}^m$ and $Z \in \mathbf{S}^p$ are the Lagrange multiplier vector and matrix for $g(x) = 0$ and $X(x) \succeq 0$, respectively. A gradient of the Lagrangian function L with respect to x is given by

$$\nabla_x L(x, y, Z) = \nabla f(x) - J_g(x)^\top y - \mathcal{A}^*(x)Z.$$

The Karush-Kuhn-Tucker (KKT) conditions of (1.1.1) are written as

$$\begin{bmatrix} \nabla_x L(x,y,Z) \\ g(x) \\ X(x)Z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \qquad X(x) \succeq 0, \qquad Z \succeq 0. \tag{2.3.1}$$

Next, we introduce definitions of the stationary point, the Mangasarian-Fromovitz constraint qualification condition, the second-order sufficient condition, the strict complementarity condition and the nondegeneracy condition.

We say that a point $x^* \in \mathbf{R}^n$ is a stationary point of (1.1.1) if there exist Lagrange multipliers $y^* \in \mathbf{R}^m$ and $Z^* \in \mathbf{S}^p$ such that (x^*, y^*, Z^*) satisfies the KKT conditions (2.3.1). We say that the Mangasarian-Fromovitz constraint qualification condition holds at x^* if $\mathrm{rank}(J_g(x^*)) = m$ and there exists a vector $v \in \mathbf{R}^n$ such that

$$v \neq 0, \quad J_g(x^*)v = 0, \quad X(x^*) + \sum_{i=1}^{n} v_i A_i(x^*) \succ 0.$$

Then, we present a theorem associated with the Mangasarian-Fromovitz constraint qualification condition.

Theorem 2.3.1. *[13] Let x^* be a local optimal solution of nonlinear SDP (1.1.1). If the Mangasarian-Fromovitz constraint qualification condition holds at x^*, there exist Lagrange multipliers $y^* \in \mathbf{R}^m$ and $Z^* \in \mathbf{S}^p$ such that (x^*, y^*, Z^*) satisfies the KKT conditions (2.3.1).* □

Let x^* be a stationary point of nonlinear SDP (1.1.1), and let $\Lambda(x^*)$ be a set defined by

$$\Lambda(x^*) := \{\ (y, Z) \in \mathbf{R}^m \times \mathbf{S}^p \mid (x^*, y, Z) \text{ satisfies (2.3.1)}\ \}.$$

First, we describe the second-order sufficient condition for nonlinear SDP (1.1.1). Let $C(x^*)$ be the critical cone of (1.1.1) at x^*, that is,

$$C(x^*) := \left\{ \ h \in \mathbf{R}^n \ \middle| \ J_g(x^*)h = 0,\ \sum_{i=1}^{n} h_i A_i(x^*) \in \mathcal{T}_{\mathbf{S}^p_+}(X(x^*)),\ \nabla f(x^*)^\top h = 0 \ \right\},$$

where

$$\begin{aligned} &\mathcal{T}_{\mathbf{S}^p_+}(X(x^*)) := \left\{\ D \in \mathbf{S}^p \mid \mathrm{dist}\left(X(x^*) + tD, \mathbf{S}^p_+\right) = o(t),\ t \geq 0\ \right\}, \\ &\mathrm{dist}\left(P,\ \mathbf{S}^p_+\right) := \inf\{\ \|P - Q\|_F \mid Q \in \mathbf{S}^p_+\ \}. \end{aligned}$$

Then, we say that the second-order sufficient condition holds at x^* if

$$\sup_{(y,Z)\in\Lambda(x^*)} h^\top \left(\nabla^2_{xx} L(x^*, y, Z) + \hat{H}(x^*, Z)\right) h > 0 \quad \text{for all } h \in C(x^*)\backslash\{0\},$$

where the (i,j)-th element of $\hat{H}(x^*, Z)$ is $2\mathrm{tr}(A_i(x^*)X(x^*)^\dagger A_j(x^*)Z)$, and $X(x^*)^\dagger$ denotes the Moore-Penrose generalized inverse of $X(x^*)$. In the following, we propose a theorem related to the second-order sufficient condition.

Theorem 2.3.2. *[71] Suppose that the Mangasarian-Fromovitz constraint qualification condition holds at x^*. Then, the second-order sufficient condition holds at x^* if and only if x^* is a strict local optimal solution of nonlinear SDP (1.1.1).* □

Next, we describe the strict complementarity condition and the nondegeneracy condition. We say that the strict complementarity condition holds at x^* if there exists $(y^*, Z^*) \in \Lambda(x^*)$ such that $\mathrm{rank}(X(x^*)) + \mathrm{rank}(Z^*) = p$. Then, without loss of generality, we may assume that $X(x^*)$ and Z^* are written as

$$X(x^*) = \begin{bmatrix} \overline{X}^* & 0 \\ 0 & 0 \end{bmatrix}, \quad Z^* = \begin{bmatrix} 0 & 0 \\ 0 & \underline{Z}^* \end{bmatrix},$$

where $\overline{X}^* \in \mathbf{S}^q_{++}$ and $\underline{Z}^* \in \mathbf{S}^r_{++}$, and q and r are positive integers such that $q + r = p$. Then, for each $i \in \{1, \ldots, n\}$, let $\underline{A}_i(x) \in \mathbf{S}^r$ be a submatrix of $A_i(x)$ such that

$$A_i(x) = \begin{bmatrix} \overline{A}_i(x) & \widehat{A}_i(x) \\ \widehat{A}_i(x)^\top & \underline{A}_i(x) \end{bmatrix},$$

where $\overline{A}_i(x)$ and $\widehat{A}_i(x)$ are appropriate submatrices of $A_i(x)$. We define

$$B(x) := [\mathrm{svec}(\underline{A}_1(x)), \ldots, \mathrm{svec}(\underline{A}_n(x))] \in \mathbf{R}^{\frac{r(r+1)}{2} \times n}, \quad K(x) := \begin{bmatrix} J_g(x) \\ B(x) \end{bmatrix} \in \mathbf{R}^{(m + \frac{r(r+1)}{2}) \times n}.$$

We say that the nondegeneracy condition holds at x^* if $\mathrm{rank}(K(x^*)) = m + \frac{r(r+1)}{2}$. Finally, we give a theorem related to the Lagrange multipliers corresponding to a stationary point $x^* \in \mathbf{R}^n$.

Theorem 2.3.3. *[71] Let $x^* \in \mathbf{R}^n$ be a stationary point of nonlinear SDP (1.1.1). If the strict complementarity condition holds at x^*, then $\Lambda(x^*)$ is a singleton if and only if the nondegeneracy condition is satisfied at x^*.* □

2.4 Barrier KKT conditions for nonlinear SDP

Most of solution methods for nonlinear SDP are developed to find a point $w := (x, y, Z)$ that satisfies the KKT conditions. However, it is difficult to get such a point directly due to the complementarity condition $X(x)Z = 0$ with $X(x) \succeq 0$ and $Z \succeq 0$. To overcome this difficulty, the primal-dual interior point methods proposed by a few researchers exploited the following two conditions with a barrier parameter $\mu > 0$:

Barrier KKT conditions

$$\begin{bmatrix} \nabla_x L(w) \\ g(x) \\ X(x)Z - \mu I \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad X(x) \succ 0, \quad Z \succ 0, \tag{2.4.1}$$

Shifted barrier KKT conditions

$$\begin{bmatrix} \nabla_x L(w) \\ g(x) + \mu y \\ X(x)Z - \mu I \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad X(x) \succ 0, \quad Z \succ 0. \tag{2.4.2}$$

The conditions (2.4.1) are called the barrier KKT conditions, and they were proposed by Yamashita, Yabe and Harada [72]. Moreover, the barrier KKT conditions come from Yamashita [70] for nonlinear programming. On the other hand, the conditions (2.4.2) are called the shifted barrier KKT conditions, and they were proposed by Kato, Yabe and Yamashita [34]. Moreover, the shifted barrier KKT conditions are derived from Forsgren and Gill [18] for nonlinear programming. In what follows, we call a point w satisfying the (shifted) barrier KKT conditions a *(shifted) barrier KKT point*.

Furthermore, we define the following generalized shifted barrier KKT conditions which are a new concept proposed in this thesis:

Generalized shifted barrier KKT conditions

$$r_\kappa(w,\mu) := \begin{bmatrix} \nabla_x L(w) \\ g(x) + \kappa\mu y \\ \mathrm{svec}(X(x) \circ Z - \mu I) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad X(x) \succeq 0, \quad Z \succeq 0, \tag{2.4.3}$$

where $\kappa \in [0,\infty)$ and $\mu \geq 0$. In the conditions (2.4.3), we generalize $g(x) = 0$ and $g(x) + \mu y = 0$ in (2.4.1) and (2.4.2) as $g(x) + \kappa\mu y = 0$ by using $\kappa \in [0,\infty)$. Moreover, we also replace $X(x)Z - \mu I = 0$ with $\mathrm{svec}(X(x) \circ Z - \mu I) = 0$. Note that since $X(x), Z \in \mathbf{S}^p_+$, it follows from Proposition 2.2.8 (b) that $X(x)Z - \mu I = 0$ is equivalent to $X(x) \circ Z - \mu I = 0$. In the remaining thesis, we call (2.4.3) the generalized shifted barrier KKT conditions. If $\mu = 0$, the generalized shifted barrier KKT conditions (2.4.3) are reduced to the KKT conditions (2.3.1). Note that when $\mu > 0$, the conditions $X(x) \succeq 0$ and $Z \succeq 0$ in (2.4.3) are equivalent to $X(x) \succ 0$ and $Z \succ 0$. Moreover, if $\kappa = 0$ and $\mu > 0$, then (2.4.3) are reduced to the barrier KKT conditions (2.4.1). Similarly, if $\kappa = 1$ and $\mu > 0$, then (2.4.3) are equal to the shifted barrier KKT conditions (2.4.2). For a given $\xi > 0$, a point $w \in \mathbf{R}^l$ such that $\|r_\kappa(w,0)\| \leq \xi$, $X(x) \succeq 0$ and $Z \succeq 0$ is called an *approximate KKT point*. Similarly, if $w \in \mathbf{R}^l$ satisfies that $\|r_\kappa(w,\mu)\| \leq \xi$ with $\mu > 0$, $X(x) \succeq 0$ and $Z \succeq 0$, we call w an *approximate generalized shifted barrier KKT point*. In particular, when $\kappa = 0$ ($\kappa = 1$), we call w an *approximate (shifted) barrier KKT point*. Finally, we define a set $\mathcal{W}$ by

$$\mathcal{W} := \{\ w \mid X(x) \succ 0,\ Z \succ 0\ \}.$$

We call a point $w \in \mathcal{W}$ an *interior point*.

As described in Subsection 1.2.2, scaling is frequently exploited in the existing primal-dual interior point methods. Scaling means that we generate matrices

$$\widetilde{X}(x) := TX(x)T^\top \quad \text{and} \quad \widetilde{Z} := T^{-\top} Z T^{-1}$$

by using a nonsingular matrix T such that $\widetilde{X}(x)\widetilde{Z} = \widetilde{Z}\widetilde{X}(x)$. We call T a scaling matrix. Moreover, we also use the following scaled generalized shifted barrier KKT conditions:

Scaled generalized shifted barrier KKT conditions

$$\widetilde{r}_\kappa(w,\mu) := \begin{bmatrix} \nabla_x L(w) \\ g(x) + \kappa\mu y \\ \mathrm{svec}(\widetilde{X}(x) \circ \widetilde{Z} - \mu I) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \widetilde{X}(x) \succeq 0, \quad \widetilde{Z} \succeq 0. \tag{2.4.4}$$

Since $\widetilde{X}(x), \widetilde{Z} \in \mathbf{S}^p_+$, it follows from Proposition 2.2.8 (b) that $\widetilde{X}(x) \circ \widetilde{Z} - \mu I = 0$ is equivalent to $\widetilde{X}(x)\widetilde{Z} - \mu I = 0$. It is clear that $\widetilde{X}(x)\widetilde{Z} - \mu I = 0$ is equivalent to $X(x)Z - \mu I = 0$. Therefore, since $X(x)Z - \mu I = 0$ is equivalent to $X(x) \circ Z - \mu I = 0$ by Proposition 2.2.8 (b), $\widetilde{X}(x) \circ \widetilde{Z} - \mu I = 0$ is equivalent to $X(x) \circ Z - \mu I = 0$, i.e., (2.4.3) and (2.4.4) are equivalent. Note that we call (2.4.4) the scaled barrier KKT conditions when $\kappa = 0$ and $\mu > 0$. Similarly, we call (2.4.4) the scaled shifted barrier KKT conditions when $\kappa = 1$ and $\mu > 0$.

Finally, we present the well-known scaling matrix T.

Choice of scaling matrix

(i) If we consider $T = X(x)^{-\frac{1}{2}}$, then $\widetilde{X}(x) = I$ and $\widetilde{Z} = X^{\frac{1}{2}} Z X^{\frac{1}{2}}$. This choice corresponds to the HRVW/KSH/M direction for linear SDP [28, 35, 42]. Clearly, the matrices $\widetilde{X}(x)$ and $\widetilde{Z}$ satisfy $\widetilde{X}(x)\widetilde{Z} = \widetilde{Z}\widetilde{X}(x)$.

(ii) If we consider $T = W^{-\frac{1}{2}}$ with $W := X^{\frac{1}{2}}(X^{\frac{1}{2}} Z X^{\frac{1}{2}})^{-\frac{1}{2}} X^{\frac{1}{2}}$, then $\widetilde{X}(x) = W^{-\frac{1}{2}} X W^{-\frac{1}{2}} = W^{\frac{1}{2}} Z W^{\frac{1}{2}} = \widetilde{Z}$. This choice corresponds to the NT direction for linear SDP [44, 45]. Clearly, the matrices $\widetilde{X}(x)$ and $\widetilde{Z}$ satisfy $\widetilde{X}(x)\widetilde{Z} = \widetilde{Z}\widetilde{X}(x)$.

2.5 Block coordinate descent method for nondifferentiable minimization

In this section, we introduce a block coordinate descent (BCD) method for nondifferentiable minimization, and we present some results related to the nondifferentiable minimization and the BCD method. Note that these results are derived from Tseng [62]. First, we consider the following unconstrained optimization problem:

$$\underset{x \in \mathbf{R}^n}{\text{minimize}} \quad f(x) := f_0(x) + \sum_{k=1}^{N} f_k(x_k), \tag{2.5.1}$$

where $f : \mathbf{R}^{n_1+\cdots+n_N} \to \mathbf{R} \cup \{\infty\}$ is proper, that is, there exists $x \in \mathbf{R}^{n_1+\cdots+n_N}$ such that $f(x) < \infty$, and $f_0 : \mathbf{R}^{n_1+\cdots+n_N} \to \mathbf{R} \cup \{\infty\}$ and $f_k : \mathbf{R}^{n_k} \to \mathbf{R} \cup \{\infty\}$ for $k = 1, \ldots, N$. Note that $N, n_1, \ldots, n_N$ are positive integers, and $x_1, \ldots, x_N$ denote coordinate blocks of $x = [x_1, \ldots, x_N]$. In the following, we introduce some concepts. We say that x is a coordinatewise minimum point of f if $x \in \mathrm{dom} f$ and

$$f(x + (0, \ldots, d_k, \ldots, 0)) \geq f(x) \quad \text{for all } d_k \in \mathbf{R}^{n_k} \text{ and } k = 1, \ldots, N,$$

where we denote by $(0, \ldots, d_k, \ldots, 0)$ the vector in $\mathbf{R}^{n_1+\cdots+n_N}$ whose k-th coordinate block is d_k and whose other coordinates are zero. We say that f is quasiconvex on a convex set C if

$$f(\lambda x + (1-\lambda)y) \leq \max\{f(x), f(y)\} \quad \text{for all } \lambda \in [0,1] \text{ and } x,\ y \in C.$$

We say that f is hemivariate on a set $D \subset \mathrm{dom} f$ if f is not constant on any line segment of D, that is, if there exist no distinct points $x, y \in \mathrm{dom} f$ such that

$$tx + (1-t)y \in D, \quad f(tx + (1-t)y) = f(x) \quad \text{for all } t \in [0,1].$$

In what follows, we present a proposition related to a stationary point and a coordinate minimum point.

Proposition 2.5.1. *[62] Let f be the function in (2.5.1). Suppose that x is a coordinatewise minimum point of f. If f_0 is differentiable at x, then x is a stationary point of (2.5.1).* □

Next, we present a BCD method based on the cyclic rule.

Block coordinate descent method

Step 0. Choose any $x^0 := [x_1^0, \ldots, x_N^0] \in \mathrm{dom} f$. Set $r := 0$.

Step 1. Calculate $x^{r+1} := [x_1^{r+1}, \ldots, x_N^{r+1}]$ by solving the following problems:

$$\begin{aligned} x_1^{r+1} &\in \operatorname*{argmin}_{x \in \mathbf{R}^{n_1}} f(x, x_2^r, \ldots, x_N^r), \\ x_2^{r+1} &\in \operatorname*{argmin}_{x \in \mathbf{R}^{n_2}} f(x_1^{r+1}, x, x_3^r \ldots, x_N^r), \\ &\vdots \\ x_N^{r+1} &\in \operatorname*{argmin}_{x \in \mathbf{R}^{n_N}} f(x_1^{r+1}, \ldots, x_{N-1}^{r+1}, x). \end{aligned}$$

Step 2. If a termination criterion is satisfied, then stop.

Step 3. Set $r := r + 1$, and go to Step 1. □

Finally, we provide a proposition associated with a convergence analysis for the BCD method.

Proposition 2.5.2. *[62] Let $\{x^r\}$ be a sequence generated by the BCD method. Suppose that $f, f_0, f_1, \ldots, f_N$ satisfy the following assumptions:*

(i) *f_0 is continuous on* $\mathrm{dom} f_0$*;*

(ii) *For each $k \in \{1, \ldots, N\}$ and $x_j \in \mathbf{R}^{n_j}$ $(j = 1, \ldots, N,\ j \neq k)$, the function $x_k \mapsto f(x_1, \ldots, x_N)$ is quasiconvex and hemivariate;*

(iii) *$f_0, f_1, \ldots, f_N$ are lower semicontinuous;*

(iv) *There exist $Y_k \subset \mathbf{R}^{n_k}$ $(k = 1, \ldots, N)$ such that* $\mathrm{dom} f_0 = Y_1 \times \cdots \times Y_N$.

Then, either $\{f(x^r)\} \downarrow -\infty$ or else every accumulation point x^* is a coordinatewise minimum point of f. □

Chapter 3

A differentiable merit function for shifted barrier Karush-Kuhn-Tucker conditions of nonlinear semidefinite programming problems

3.1 Introduction

In this chapter, we consider the following nonlinear semidefinite programming (SDP) problem:

$$\begin{array}{ll} \underset{x \in \mathbf{R}^n}{\text{minimize}} & f(x), \\ \text{subject to} & g(x) = 0, \quad X(x) \succeq 0, \end{array} \tag{3.1.1}$$

where $f : \mathbf{R}^n \to \mathbf{R}$, $g : \mathbf{R}^n \to \mathbf{R}^m$ and $X : \mathbf{R}^n \to \mathbf{S}^p$ are twice continuously differentiable functions.

For nonlinear SDP, there exist several solution methods which have the global convergence such as the methods described in Chapter 1. However, as mentioned in Chapter 1, there exist some issues associated with assumptions for the global convergence. The aim of this chapter is to propose a primal-dual interior point method for (3.1.1) that is convergent globally under milder conditions compared with the existing methods. In particular, we specify the conditions related to the problem data, i.e., f, g and X. We also show that these conditions hold for linear SDP.

In this chapter, we propose a new merit function F whose stationary points satisfy the shifted barrier KKT conditions. This function is an extension of a merit function proposed by Forsgren and Gill [18] developed for nonlinear programming, and it consists of simple functions, such as log-determinant and trace. Thus, it is easy to implement the proposed method with the merit function F. We show the following important properties of the merit function F:

(i) The merit function F is differentiable;

(ii) Any stationary point of the merit function F is a shifted barrier KKT point;

(iii) The level set of the merit function F is bounded under some reasonable assumptions.

These properties mean that we can find a shifted barrier KKT point by minimizing the merit function F. To minimize F, we also propose a Newton-type method based on nonlinear equations in the shifted barrier KKT conditions. We show that the Newton direction is sufficiently descent for the merit function F. As a result, we prove the global convergence of the proposed Newton-type method. These details are provided in Section 3.3.

This chapter is organized as follows. In Section 3.2, we introduce some important concepts, which are used in the subsequent section, and we present a primal-dual interior point method based on the shifted barrier KKT conditions. In Section 3.3, we first propose a merit function F for a shifted barrier KKT point and present its properties. Secondly, we propose a Newton-type method that minimizes the merit function F. Moreover, we prove the global convergence of the proposed Newton-type method. In Section 3.4, we report some numerical results for the proposed method. Finally, we make some concluding remarks in Section 3.5.

3.2 Primal-dual interior point method based on shifted barrier KKT conditions

As described in Chapter 1, the main goal of solution methods for nonlinear SDP (3.1.1) is to find a KKT point which satisfies the following KKT conditions:

$$\begin{bmatrix} \nabla_x L(v) \\ g(x) \\ \mathrm{svec}(X(x) \circ Z) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad X(x) \succeq 0, \quad Z \succeq 0, \tag{3.2.1}$$

where $v := (x, y, Z) \in \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{S}^p$ and L is the Lagrangian function, that is, $L(v) = f(x) - g(x)^\top y - \langle X(x), Z \rangle$. In what follows, we introduce a prototype of a primal-dual interior point method based on the shifted barrier KKT conditions (2.4.2). To this end, we use the generalized shifted barrier KKT conditions (2.4.3) with $\kappa = 1$, that is,

$$r_1(v, \mu) = \begin{bmatrix} \nabla_x L(v) \\ g(x) + \mu y \\ \mathrm{svec}(X(x) \circ Z - \mu I) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad X(x) \succeq 0, \quad Z \succeq 0, \tag{3.2.2}$$

where $\mu \geq 0$. Note that if $\mu = 0$, the conditions (3.2.2) are equivalent to the KKT conditions (3.2.1). Note also that if $\mu > 0$, the conditions (3.2.2) hold if and only if $r_1(v, \mu) = 0$ and $v \in \mathcal{W}$, that is, the shifted barrier KKT conditions hold. Note that $\mathcal{W} = \{(x, y, Z) | X(x) \succ 0, Z \succ 0\}$.

Now, we give a framework of a primal-dual interior point method.

Algorithm 3.2.1.

Step 0. *Let $\{\mu_k\}$ be a positive sequence such that $\mu_k \to 0$ as $k \to \infty$. Choose positive constants σ and ϵ. Set $k := 0$.*

Step 1. *Find an approximate shifted barrier KKT point v_{k+1} such that $\|r_1(v_{k+1}, \mu_k)\| \leq \sigma \mu_k$ and $v_{k+1} \in \mathcal{W}$.*

Step 2. *If* $\|r_1(v_{k+1}, 0)\| \leq \epsilon$, *then stop.*

Step 3. *Set* $k := k + 1$ *and go to Step 1.* □

The following theorem gives conditions for the global convergence of Algorithm 3.2.1. It can be proven in a way similar to [72, Theorem 1]. Thus, we omit the proof.

Theorem 3.2.1. *Suppose that an approximate shifted barrier KKT point* v_{k+1} *is found in Step 1 at every iteration. Moreover, suppose that the sequence* $\{x_k\}$ *is bounded and that the Mangasarian-Fromovitz constraint qualification condition holds at any accumulation point of* $\{x_k\}$, *i.e., for any accumulation point* x^* *of* $\{x_k\}$, *the matrix* $J_g(x^*)$ *is of full rank and there exists a nonzero vector* $w \in \mathbf{R}^n$ *such that*

$$J_g(x^*)w = 0 \quad \text{and} \quad X(x^*) + \sum_{i=1}^{n} w_i A_i(x^*) \succ 0.$$

Then, the sequences $\{y_k\}$ *and* $\{Z_k\}$ *are bounded, and any accumulation point of* $\{v_k\}$ *satisfies the KKT conditions (3.2.1).* □

The theorem guarantees the global convergence if an approximate shifted barrier KKT point v_{k+1} is found at each iteration. Thus it is important to present a method that finds such a point. In the next section, we will propose a merit function for the shifted barrier KKT point and a Newton-type method for solving an unconstrained minimization problem of the merit function.

3.3 Finding a shifted barrier KKT point

In order to find the approximate shifted barrier KKT point v_{k+1} in Step 1 of Algorithm 3.2.1, we may solve the following unconstrained minimization problem:

$$\begin{array}{ll} \text{minimize} & \|r_1(w, \mu)\|^2, \\ \text{subject to} & w := (x, y, Z) \in \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{S}^p, \end{array}$$

Unfortunately, a stationary point of the problem is not necessarily a shifted barrier KKT point unless a Jacobian of r_1 with respect to w at (w, μ) is invertible. In this section, we first construct a differentiable merit function F whose stationary point is always a shifted barrier KKT point. Moreover, we show that a Newton direction for the nonlinear equations $r_1(w, \mu) = 0$ is a descent direction of the merit function F. Next, we propose a Newton-type method for solving an unconstrained minimization of the merit function F. Finally, we show that the proposed algorithm finds a shifted barrier KKT point under some mild assumptions.

3.3.1 Merit function and its properties

We propose the following merit function $F : \mathcal{W} \to \mathbf{R}$ for the shifted barrier KKT point:

$$F(x, y, Z) := F_{BP}(x) + \nu F_{PD}(x, y, Z),$$

where ν is a positive constant, and the functions $F_{BP} : \Omega \to \mathbf{R}$ and $F_{PD} : \mathcal{W} \to \mathbf{R}$ are defined by

$$F_{BP}(x) := f(x) + \frac{1}{2\mu}\|g(x)\|^2 - \mu \log\det X(x),$$

and

$$F_{PD}(x, y, Z) := \frac{1}{2\mu}\|g(x) + \mu y\|^2 + \langle X(x), Z\rangle - \mu \log\det X(x) \det Z,$$

respectively. Note that $\Omega = \{\ x \in \mathbf{R}^n \mid X(x) \succ 0\ \}$. The functions F_{BP} and F_{PD} are called the primal barrier penalty function and the primal-dual barrier penalty function, respectively. Note that F is convex with respect to x when f is convex and g, X are affine. The merit function F is an extension of the one proposed by Forsgren and Gill [18] for nonlinear programming.

Remark 3.3.1. *For the shifted barrier KKT conditions, Kato, Yabe and Yamashita [34] also proposed the merit function $\widetilde{F} : \mathcal{W} \to \mathbf{R}$ as*

$$\widetilde{F}(x, y, Z) := F_{BP}(x) + \nu \widetilde{F}_{PD}(x, y, Z),$$

where $\widetilde{F}_{PD} : \mathcal{W} \to \mathbf{R}$ is defined by

$$\widetilde{F}_{PD}(x, y, Z) := \frac{1}{2}\|g(x) + \mu y\|^2 + \log \frac{\frac{1}{p}\langle X(x), Z\rangle + \|Z^{\frac{1}{2}} X(x) Z^{\frac{1}{2}} - \mu I\|_F^2}{(\det(X(x)Z))^{\frac{1}{p}}}.$$

They showed that $\widetilde{F}$ has nice properties like the merit function F. However, $\widetilde{F}$ is more complicated than F, and hence it might not be easy to implement the Newton-type method based on $\widetilde{F}$ in [34]. Furthermore, even if f is convex and g, X are affine, $\widetilde{F}$ is not necessarily convex with respect to x.

In the rest of this subsection, we present some useful properties of the merit function F such as the differentiability, the equivalence between a stationary point of F and a shifted barrier KKT point, and the level boundedness.

First of all, we present a concrete formula of the derivatives of the merit function F.

Theorem 3.3.1. *The merit function F is differentiable on $\mathcal{W}$. Moreover, its derivative is given by*

$$\nabla F(w) = \begin{bmatrix} \nabla F_{BP}(x) + \nu \nabla_x F_{PD}(w) \\ \nu \nabla_y F_{PD}(w) \\ \nu \nabla_Z F_{PD}(w) \end{bmatrix},$$

where $\nabla F_{BP}(x) = \nabla f(x) + \frac{1}{\mu} J_g(x)^\top g(x) - \mu \mathcal{A}^*(x) X(x)^{-1}$, $\nabla_x F_{PD}(w) = \frac{1}{\mu} J_g(x)^\top (g(x) + \mu y) + \mathcal{A}^*(x)(Z - \mu X(x)^{-1})$, $\nabla_y F_{PD}(w) = g(x) + \mu y$ and $\nabla_Z F_{PD}(w) = X(x) - \mu Z^{-1}$. □

Next, we show the equivalence between a stationary point of the merit function F and a shifted barrier KKT point.

Theorem 3.3.2. *A point $w^* \in \mathcal{W}$ is a stationary point of the merit function F if and only if w^* is a shifted barrier KKT point.*

Proof. First, let $w^* = (x^*, y^*, Z^*) \in \mathcal{W}$ be a stationary point of the merit function F. Then, Theorem 3.3.1 yields that

$$\nabla f(x^*) + \frac{1}{\mu} J_g(x^*)^\top \{(1+\nu)g(x^*) + \nu\mu y^*\} + \mathcal{A}^*(x^*)\left\{\nu Z^* - \mu(1+\nu)X(x^*)^{-1}\right\} = 0, \quad (3.3.1)$$

$$g(x^*) + \mu y^* = 0, \quad X(x^*) - \mu(Z^*)^{-1} = 0. \quad (3.3.2)$$

Thus we have

$$\begin{aligned}
\nabla_x L(w^*) &= \nabla f(x^*) - J_g(x^*)^\top y^* - \mathcal{A}^*(x^*)Z^* \\
&= \nabla f(x^*) + \frac{1}{\mu} J_g(x^*)^\top g(x^*) - \mu\mathcal{A}^*(x^*)X(x^*)^{-1} \\
&= -\frac{\nu}{\mu} J_g(x^*)^\top \{g(x^*) + \mu y^*\} - \nu\mathcal{A}^*(x^*)X(x^*)^{-1}\left\{X(x^*) - \mu(Z^*)^{-1}\right\} Z^* \\
&= 0,
\end{aligned}$$

where the second and third equalities follow from (3.3.2) and (3.3.1), respectively. Therefore, w^* is a shifted barrier KKT point.

Conversely, let $w^* = (x^*, y^*, Z^*)$ be a shifted barrier KKT point. Then, we obtain that

$$\nabla_x L(w^*) = 0, \quad g(x^*) + \mu y^* = 0, \quad X(x^*)Z^* - \mu I = 0.$$

From Theorem 3.3.1, it is clear that $\nabla_y F(w^*) = \nu\{g(x^*) + \mu y^*\} = 0$ and $\nabla_Z F(w^*) = \nu\{X(x^*) - \mu(Z^*)^{-1}\} = \nu\{X(x^*)Z^* - \mu I\}(Z^*)^{-1} = 0$. Moreover,

$$\begin{aligned}
\nabla_x F(w^*) &= \nabla f(x^*) + \frac{1}{\mu} J_g(x^*)^\top \{(1+\nu)g(x^*) + \nu\mu y^*\} + \mathcal{A}^*(x^*)\left\{\nu Z^* - \mu(1+\nu)X(x^*)^{-1}\right\} \\
&= \nabla f(x^*) + \frac{1}{\mu} J_g(x^*)^\top g(x^*) - \mu\mathcal{A}^*(x^*)X(x^*)^{-1} \\
&\qquad + \frac{\nu}{\mu} J_g(x^*)^\top \{g(x^*) + \mu y^*\} + \nu\mathcal{A}^*(x^*)\left\{Z^* - \mu X(x^*)^{-1}\right\} \\
&= \nabla_x L(x^*) + \frac{\nu}{\mu} J_g(x^*)^\top \{g(x^*) + \mu y^*\} + \nu\mathcal{A}^*(x^*)X(x^*)^{-1}\{X(x^*)Z^* - \mu I\} \\
&= 0.
\end{aligned}$$

Therefore, w^* is a stationary point of F. □

This theorem is an extension of [18, Lemma 3.1] for nonlinear programming.

From this theorem, we can find an approximate shifted barrier KKT point by solving the following unconstrained minimization problem:

$$\begin{aligned}
&\text{minimize} && F(w), \\
&\text{subject to} && w \in \mathcal{W}.
\end{aligned} \quad (3.3.3)$$

One of the sufficient conditions under which descent methods find a stationary point is that a level set of the objective function is bounded. Thus, it is worth providing sufficient conditions for the level boundedness of the merit function F. For a given $\alpha \in \mathbf{R}$, we define a level set $\mathcal{L}(\alpha)$ of F by

$$\mathcal{L}(\alpha) = \{w \in \mathcal{W} \mid F(w) \le \alpha\}.$$

We first give two lemmas. The following lemma follows directly from [72, Lemma 1].

Lemma 3.3.1. *Let $w = (x, y, Z) \in \mathcal{W}$ and $\mu > 0$. Then the following properties hold.*

(a) $\langle X(x), Z\rangle - \mu \log\det X(x)Z \geq p\mu(1 - \log\mu)$.

(b) $F_{PD}(w) \geq p\mu(1-\log\mu)$. *The equality holds if and only if* $g(x)+\mu y = 0$ *and* $X(x)Z-\mu I = 0$.

(c) $\lim_{\langle X(x),Z\rangle\downarrow 0} F_{PD}(w) = \infty$ and $\lim_{\langle X(x),Z\rangle\uparrow\infty} F_{PD}(w) = \infty$. □

Lemma 3.3.2. *Suppose that an infinite sequence $\{w_j = (x_j, y_j, Z_j)\}$ is included in $\mathcal{L}(\alpha)$. Suppose also that the sequence $\{x_j\}$ is bounded. Then, the sequences $\{y_j\}$ and $\{Z_j\}$ are also bounded. In addition, the sequences $\{X(x_j)\}$ and $\{Z_j\}$ are uniformly positive definite.*

Proof. Since $\{x_j\}$ is bounded, $\{-\log\det X(x_j)\}$ is bounded below. Thus, there exists a real number M_1 such that $M_1 \leq F_{BP}(x_j)$ for all j. Then, the definition of F and $w_j \in \mathcal{L}(\alpha)$ imply that $F_{PD}(w_j) \leq \frac{1}{\nu}(\alpha - M_1)$ for all j, which can be rewritten as

$$\frac{1}{2\mu}\|g(x_j) + \mu y_j\|^2 \leq \frac{\alpha - M_1}{\nu} - \langle X(x_j), Z_j\rangle + \mu\log\det X(x_j)Z_j \leq \frac{\alpha - M_1}{\nu} - p\mu(1 - \log\mu),$$

where the last inequality follows from Lemma 3.3.1 (a). Hence, $\{y_j\}$ is bounded.

Next, we show that $\{X(x_j)\}$ is uniformly positive definite. From Lemma 3.3.1 (b), we have

$$M_1 \leq F_{BP}(x_j) = F(w_j) - \nu F_{PD}(w_j) \leq \alpha - \nu F_{PD}(w_j) \leq \alpha - \nu p\mu(1 - \log\mu) \quad \text{for all } j,$$

and hence, $\{F_{BP}(x_j)\}$ is bounded. It then follows from the boundedness of $\{x_j\}$ and the definition of F_{BP} that $\{-\log\det X(x_j)\}$ is also bounded. From Proposition 2.2.14, the boundedness of $\{-\log\det X(x_j)\}$ and $\{X(x_j)\}$ implies that $\{X(x_j)\}$ is uniformly positive definite, that is, there exists $\underline{\lambda}$ such that $\lambda_{\min}(X(x_j)) \geq \underline{\lambda} > 0$ for all j.

Next we show that $\{Z_j\}$ is bounded. From Lemma 3.3.1 (b), we have

$$p\mu(1 - \log\mu) \leq F_{PD}(w_j) \leq \frac{1}{\nu}(\alpha - M_1) \quad \text{for all } j,$$

and hence $\{F_{PD}(w_j)\}$ is bounded. Then, Lemma 3.3.1 (c) yields that $\{\langle X(x_j), Z_j\rangle\}$ is bounded. Thus, there exists a real number M_2 such that for all j,

$$M_2 \geq \mathrm{tr}(X(x_j)Z_j) \geq \lambda_{\min}(X(x_j))\mathrm{tr}(Z_j) \geq \underline{\lambda}\mathrm{tr}(Z_j) = \underline{\lambda}\sum_{k=1}^{p}\lambda_k(Z_j) \tag{3.3.4}$$

where the second inequality follows from Proposition 2.2.1 (c). Since $\{Z_j\}$ is positive definite, $\lambda_k(Z_j) > 0$ for $k = 1, \ldots, p$. Then, (3.3.4) implies that $\{\lambda_k(Z_j)\}$ is bounded for $k = 1, \ldots, p$, and hence $\{Z_j\}$ is bounded.

Finally, we show that $\{Z_j\}$ is uniformly positive definite. Recall that

$$F_{PD}(w_j) = \frac{1}{2\mu}\|g(x_j) + \mu y_j\|^2 + \langle X(x_j), Z_j\rangle - \mu\log\det X(x_j) - \mu\log\det Z_j,$$

and that $\{x_j\}, \{y_j\}, \{\langle X(x_j), Z_j\rangle\}, \{-\log\det X(x_j)\}$ and $\{F_{PD}(w_j)\}$ are bounded. Therefore, $\{-\log\det Z_j\}$ is also bounded. It then follows from Proposition 2.2.14 and the boundedness of $\{Z_j\}$ that $\{Z_j\}$ is uniformly positive definite. □

We now give sufficient conditions under which any level set of the merit function F is bounded.

Theorem 3.3.3. *Suppose that the following five assumptions hold.*

(i) *The function f is convex.*

(ii) *The functions $g_1, \ldots, g_m$ are affine.*

(iii) *The function X satisfies $X(\lambda u + (1-\lambda)v) - \lambda X(u) - (1-\lambda)X(v) \succeq 0$ for all $\lambda \in [0,1]$ and $u, v \in \Omega$.*

(iv) *The matrices $A_1(x), \ldots, A_n(x)$ are linearly independent for all $x \in \Omega$;*

(v) *There exists a shifted barrier KKT point w^*.*

Then, the level set $\mathcal{L}(\alpha)$ of F is bounded for all $\alpha \in \mathbf{R}$.

Proof. Let $\{(x_k, y_k, Z_k)\}$ be an infinite sequence in $\mathcal{L}(\alpha)$. We first show that $\{x_k\}$ is bounded. In order to prove this by contradiction, we suppose that there exists a subset $\mathcal{I} \subset \{0, 1, \ldots\}$ such that $\lim_{k\to\infty, k\in\mathcal{I}} \|x_k\| = \infty$. Since $F(w_k) \leq \alpha$ and $F_{PD}(w_k) \geq p\mu(1 - \log\mu)$ from Lemma 3.3.1 (b), $F_{BP}(x_k) = F(w_k) - \nu F_{PD}(w_k) \leq \alpha - \nu p\mu(1 - \log\mu)$.

On the other hand, since w^* is a shifted barrier KKT point, Theorem 3.3.1 implies that

$$0 = \nabla_x L(w^*) = \nabla f(x^*) + \frac{1}{\mu} J_g(x^*)^\top g(x^*) - \mu \mathcal{A}^*(x^*) X(x^*)^{-1} = \nabla F_{BP}(x^*). \tag{3.3.5}$$

Note that F_{BP} is strictly convex from Proposition 2.2.13 (c) and the assumptions (i)–(iv). Thus, (3.3.5) implies that x^* is a unique global minimizer of $\min\{F_{BP}(x) \mid x \in \Omega\}$. Note that $x^* \in \Omega = \{\, x \in \mathbf{R}^n \mid X(x) \succ 0 \,\}$. Then, there exists $\varepsilon > 0$ such that $\{x^* + \varepsilon u \mid \|u\| = 1\} \subset \Omega$ and $\min\{F_{BP}(x^* + \varepsilon u) \mid \|u\| = 1\} > F_{BP}(x^*)$. Let $d_k := \frac{1}{\varepsilon}(x_k - x^*)$ $(k \in \mathcal{I})$ and $F_{BP}^{\varepsilon} := \min\{F_{BP}(x^* + \varepsilon u) \mid \|u\| = 1\}$. Note that $\|d_k\| \to \infty$ $(k \to \infty, k \in \mathcal{I})$. Without loss of generality, we suppose that $\|d_k\| > 1$ for all $k \in \mathcal{I}$. From the convexity of F_{BP}, we have

$$\frac{\|d_k\| - 1}{\|d_k\|} F_{BP}(x^*) + \frac{1}{\|d_k\|} F_{BP}(x^* + \varepsilon d_k) \geq F_{BP}\left(x^* + \varepsilon \frac{d_k}{\|d_k\|}\right) \geq F_{BP}^{\varepsilon},$$

and hence $F_{BP}(x_k) = F_{BP}(x^* + \varepsilon d_k) \geq \|d_k\|(F_{BP}^{\varepsilon} - F_{BP}(x^*)) + F_{BP}(x^*)$. Thus, since $F_{BP}^{\varepsilon} - F_{BP}(x^*) > 0$, we have $F_{BP}(x_k) \to \infty$ $(k \to \infty, k \in \mathcal{I})$. However, this result contradicts $F_{BP}(x_k) \leq \alpha - p\mu(1 - \log\mu)$. Hence, for arbitrary $\{x_k, y_k, Z_k\} \subset \mathcal{L}(\alpha)$, $\{x_k\}$ is bounded. It then follows from Lemma 3.3.2 and the boundedness of $\{F(w_j)\}$ that $\{y_k\}$ and $\{Z_k\}$ are also bounded. □

Remark 3.3.2. *The level boundedness of the merit function for nonlinear programming is not given in [18]. Applying Theorem 3.3.3, it is easy to show that the merit function M in [18] is level bounded if the objective function f is convex, the constraint functions c_i $(i \in \mathcal{E})$ are affine, and* $\mathrm{rank}(J_c) = n$.

Remark 3.3.3. *Kato, Yabe and Yamashita [34] showed that their merit function $\widetilde{F}$ is differentiable and its stationary point is a shifted barrier KKT point. However, they did not discuss the level boundedness of their merit function.*

Remark 3.3.4. *Theorem 3.3.3 assumes that f is convex and g is affine. These assumptions are rather restrictive for some applications. We can replace these assumptions with the following coerciveness condition:*

$$\lim_{\|x\|\to\infty, x\in\Omega} \frac{1}{\|x\|}\left(f(x) + \frac{1}{2\mu}\|g(x)\|^2\right) = \infty.$$

Due to Theorems 3.3.1–3.3.3, we can solve unconstrained minimization problem (3.3.3) by any descent method, such as quasi-Newton methods and steepest descent methods, and hence we can get an approximate shifted barrier KKT point v_{k+1} in Step 1 of Algorithm 3.2.1.

3.3.2 Newton-type method for minimization of the merit function

In this subsection, we propose a Newton-type method for unconstrained minimization problem (3.3.3) of the merit function F.

We exploit scaling which enables us to calculate a Newton direction easily. Note that scaling have already been described in Section 2.4. Let $T \in \mathbf{R}^{p\times p}$ be a nonsingular scaling matrix such that

$$TX(x)T^\top T^{-\top} ZT^{-1} = T^{-\top} ZT^{-1} TX(x)T^\top. \tag{3.3.6}$$

As described in Section 2.4, $\widetilde{X}(x)$ and $\widetilde{Z}$ denote the following matrices:

$$\widetilde{X}(x) = TX(x)T^\top = (T \odot T)X(x), \quad \widetilde{Z} = T^{-\top} ZT^{-1} = (T^{-\top} \odot T^{-\top})Z.$$

Note that $\widetilde{X}(x)\widetilde{Z} = \widetilde{Z}\widetilde{X}(x)$ by (3.3.6). In the subsequent discussions, for simplicity, we denote $X(x)$ and $\widetilde{X}(x)$ by X and $\widetilde{X}$, respectively.

Next, we give a Newton direction and show that it is a descent direction for the merit function F. The Newton direction is derived from the nonlinear equations $r_1(w,\mu) = 0$ in (3.2.2). However, as seen later, a pure Newton direction $(\Delta x, \Delta y, \Delta Z)$ for $r_1(w,\mu) = 0$ is not necessarily a descent direction for the merit function F. Thus, we consider the following scaled shifted barrier KKT conditions:

$$\widetilde{r}_1(w,\mu) := \begin{bmatrix} \nabla_x L(w) \\ g(x) + \mu y \\ \mathrm{svec}(\widetilde{X}\circ\widetilde{Z} - \mu I) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \widetilde{X} \succ 0, \quad \widetilde{Z} \succ 0. \tag{3.3.7}$$

Note that the conditions (3.3.7) are equivalent to the scaled generalized shifted barrier KKT conditions with $\kappa = 1$ and $\mu > 0$ as mentioned in Section 2.4. We apply the Newton method to the equation $\widetilde{r}_1(w,\mu) = 0$ in (3.3.7). Then, the Newton equations derived from $\widetilde{r}_1(w,\mu) = 0$ are written as

$$G\Delta x - J_g(x)^\top \Delta y - \mathcal{A}^*(x)\Delta Z = -\nabla_x L(w), \tag{3.3.8}$$

$$J_g(x)\Delta x + \mu\Delta y = -g(x) - \mu y, \tag{3.3.9}$$

$$\widetilde{Z}\Delta\widetilde{X} + \Delta\widetilde{X}\widetilde{Z} + \widetilde{X}\Delta\widetilde{Z} + \Delta\widetilde{Z}\widetilde{X} = 2\mu I - \widetilde{X}\widetilde{Z} - \widetilde{Z}\widetilde{X}, \tag{3.3.10}$$

where G denotes a Hessian of the Lagrangian function L with respect to x or its approximation. In what follows, we call a solution $\Delta w := (\Delta x, \Delta y, \Delta Z)$ of the Newton equations (3.3.8)–(3.3.10) a *Newton direction*.

Next, we give an explicit form of the Newton direction Δw. From (3.3.9), we have

$$\Delta y = -\frac{1}{\mu}(g(x) + \mu y + J_g(x)\Delta x). \tag{3.3.11}$$

Moreover, since $(\widetilde{X} \odot I)\Delta\widetilde{Z} = \frac{1}{2}(\widetilde{X}\Delta\widetilde{Z} + \Delta\widetilde{Z}\widetilde{X})$, $(\widetilde{X} \odot I)(\mu\widetilde{X}^{-1} - \widetilde{Z}) = \frac{1}{2}(2\mu I - \widetilde{Z}\widetilde{X} - \widetilde{X}\widetilde{Z})$ and $(\widetilde{Z} \odot I)\Delta\widetilde{X} = \frac{1}{2}(\widetilde{Z}\Delta\widetilde{X} + \Delta\widetilde{X}\widetilde{Z})$, the equation (3.3.10) can be rewritten as

$$(\widetilde{X} \odot I)\Delta\widetilde{Z} + (\widetilde{Z} \odot I)\Delta\widetilde{X} = (\widetilde{X} \odot I)(\mu\widetilde{X}^{-1} - \widetilde{Z}). \tag{3.3.12}$$

Since X and T are nonsingular, $\widetilde{X} = TXT^\top$ is also nonsingular. Thus, the operator $(\widetilde{X} \odot I)$ is invertible from Proposition 2.2.9 (a). Moreover, $\widetilde{X}^{-1} = T^{-\top}X^{-1}T^{-1} = (T^{-\top} \odot T^{-\top})X^{-1}$. It then follows from (3.3.12) that

$$\begin{aligned}(T^{-\top} \odot T^{-\top})\Delta Z &= \mu\widetilde{X}^{-1} - \widetilde{Z} - (\widetilde{X} \odot I)^{-1}(\widetilde{Z} \odot I)\Delta\widetilde{X} \\ &= (T^{-\top} \odot T^{-\top})(\mu X^{-1} - Z) - (\widetilde{X} \odot I)^{-1}(\widetilde{Z} \odot I)(T \odot T)\mathcal{A}(x)\Delta x, \end{aligned} \tag{3.3.13}$$

where the last equality follows from the definitions of $\widetilde{Z}$ and $\Delta\widetilde{X}$. Since $(T^{-\top} \odot T^{-\top})^{-1} = (T^\top \odot T^\top)$ from Proposition 2.2.9 (c), multiplying both sides of (3.3.13) by $(T^{-\top} \odot T^{-\top})^{-1}$ yields that

$$\Delta Z = \mu X^{-1} - Z - (T^\top \odot T^\top)(\widetilde{X} \odot I)^{-1}(\widetilde{Z} \odot I)(T \odot T)\mathcal{A}(x)\Delta x. \tag{3.3.14}$$

Finally, we give a concrete form of Δx. Substituting (3.3.11) and (3.3.14) into (3.3.8), we obtain

$$\begin{aligned}\left(G + H + \frac{1}{\mu}J_g(x)^\top J_g(x)\right)\Delta x &= -\nabla_x L(w) - \frac{1}{\mu}J_g(x)^\top(g(x) + \mu y) + \mathcal{A}^*(x)(\mu X^{-1} - Z) \\ &= -\left(\nabla f(x) + \frac{1}{\mu}J_g(x)^\top g(x) - \mu\mathcal{A}^*(x)X^{-1}\right), \end{aligned} \tag{3.3.15}$$

where

$$H := \mathcal{A}^*(x)(T^\top \odot T^\top)(\widetilde{X} \odot I)^{-1}(\widetilde{Z} \odot I)(T \odot T)\mathcal{A}(x).$$

Note that $H : \mathbf{R}^n \to \mathbf{R}^n$ is a linear operator such that

$$Hu = \mathcal{A}^*(x)(T^\top \odot T^\top)(\widetilde{X} \odot I)^{-1}(\widetilde{Z} \odot I)(T \odot T)\mathcal{A}(x)u \quad \text{for all } u \in \mathbf{R}^n.$$

From the definitions of $\mathcal{A}(x)$ and $\mathcal{A}^*(x)$, the linear operator H is regarded as a matrix whose (i, j)-th element is written as

$$H_{ij} = \left\langle A_i(x), (T^\top \odot T^\top)(\widetilde{X} \odot I)^{-1}(\widetilde{Z} \odot I)(T \odot T)A_j(x)\right\rangle. \tag{3.3.16}$$

Since $J_g(x)^\top J_g(x)$ is positive semidefinite, we can solve the linear equation (3.3.15) with respect to Δx if $G + H$ is positive definite. Fortunately, H is positive semidefinite as shown below.

Lemma 3.3.3. *Suppose that X and Z are symmetric positive definite. Then, H is symmetric positive semidefinite. Furthermore, if $A_1(x), \ldots, A_n(x)$ are linearly independent for all $x \in \mathbf{R}^n$, then H is symmetric positive definite.*

Proof. Since $\widetilde{X}$ is positive definite, the operator $\widetilde{X} \odot I$ is invertible from Proposition 2.2.9 (a). Let $u \in \mathbf{R}^n$ and $V := (\widetilde{X} \odot I)^{-1}(T \odot T)\mathcal{A}(x)u$. Then, we have

$$\begin{aligned}
\langle Hu, u\rangle &= \left\langle \mathcal{A}^*(x)(T^\top \odot T^\top)(\widetilde{X} \odot I)^{-1}(\widetilde{Z} \odot I)(T \odot T)\mathcal{A}(x)u, u\right\rangle \\
&= \left\langle (\widetilde{Z} \odot I)(T \odot T)\mathcal{A}(x)u, (\widetilde{X} \odot I)^{-1}(T \odot T)\mathcal{A}(x)u\right\rangle \\
&= \left\langle (\widetilde{Z} \odot I)(\widetilde{X} \odot I)(\widetilde{X} \odot I)^{-1}(T \odot T)\mathcal{A}(x)u, (\widetilde{X} \odot I)^{-1}(T \odot T)\mathcal{A}(x)u\right\rangle \\
&= \left\langle (\widetilde{Z} \odot I)(\widetilde{X} \odot I)V, V\right\rangle \\
&\geq 0, \qquad (3.3.17)
\end{aligned}$$

where the second equality follows from Proposition 2.2.9 (b) and the last inequality follows from Proposition 2.2.11 (a). Therefore, H is positive semidefinite.

Next we show that H is symmetric. From (3.3.16), we have

$$\begin{aligned}
H_{ij} &= \left\langle A_i(x), (T^\top \odot T^\top)(\widetilde{X} \odot I)^{-1}(\widetilde{Z} \odot I)(T \odot T)A_j(x)\right\rangle \\
&= \left\langle (T^\top \odot T^\top)(\widetilde{X} \odot I)^{-1}(\widetilde{Z} \odot I)(\widetilde{X} \odot I)(\widetilde{X} \odot I)^{-1}(T \odot T)A_j(x), A_i(x)\right\rangle \\
&= \left\langle (T^\top \odot T^\top)(\widetilde{X} \odot I)^{-1}(\widetilde{X} \odot I)(\widetilde{Z} \odot I)(\widetilde{X} \odot I)^{-1}(T \odot T)A_j(x), A_i(x)\right\rangle \\
&= \left\langle (T^\top \odot T^\top)(\widetilde{Z} \odot I)(\widetilde{X} \odot I)^{-1}(T \odot T)A_j(x), A_i(x)\right\rangle \\
&= \left\langle A_j(x), (T^\top \odot T^\top)(\widetilde{X} \odot I)^{-1}(\widetilde{Z} \odot I)(T \odot T)A_i(x)\right\rangle \\
&= H_{ji},
\end{aligned}$$

where the third equality follows from Proposition 2.2.11 (b) and the fifth equality follows from Proposition 2.2.9 (b).

Furthermore, suppose that $A_1(x), \ldots, A_n(x)$ are linearly independent for all $x \in \mathbf{R}^n$ and $u \neq 0$. Then, we have $V = (\widetilde{X} \odot I)^{-1}(T \odot T)\mathcal{A}(x)u \neq 0$. It follows from Proposition 2.2.11 (a) and (3.3.17) that $\langle Hu, u\rangle > 0$, i.e., H is positive definite. □

Remark 3.3.5. *In the case of linear SDP, $A_1(x), \ldots, A_n(x)$ are usually supposed to be linearly independent for $x \in \mathbf{R}^n$. Then, H is positive definite from Lemma 3.3.3.*

To summarize the discussion above, we give concrete formulae of the Newton direction Δw in the following theorem.

Theorem 3.3.4. *Let $\mu > 0$ and $w = (x, y, Z) \in \mathcal{W}$. Suppose that $G + H$ is positive definite. Then, the Newton equations (3.3.8)–(3.3.10) have a unique solution $\Delta w = (\Delta x, \Delta y, \Delta Z)$ such that*

$$\begin{aligned}
\Delta x &= -\left(G + H + \frac{1}{\mu}J_g(x)^\top J_g(x)\right)^{-1}\left(\nabla f(x) + \frac{1}{\mu}J_g(x)^\top g(x) - \mu\mathcal{A}^*(x)X^{-1}\right), \\
\Delta y &= -\frac{1}{\mu}(g(x) + \mu y + J_g(x)\Delta x), \\
\Delta Z &= \mu X^{-1} - Z - (T^\top \odot T^\top)(\widetilde{X} \odot I)^{-1}(\widetilde{Z} \odot I)(T \odot T)\mathcal{A}(x)\Delta x.
\end{aligned}$$

Proof. It is clear that $\frac{1}{\mu}J_g(x)^\top J_g(x)$ is positive semidefinite. Thus, the positive definiteness of $G+H$ and (3.3.15) yield that

$$\Delta x = -\left(G+H+\frac{1}{\mu}J_g(x)^\top J_g(x)\right)^{-1}\left(\nabla f(x)+\frac{1}{\mu}J_g(x)^\top g(x)-\mu\mathcal{A}^*(x)X^{-1}\right).$$

Furthermore, Δy and ΔZ directly follow from (3.3.11) and (3.3.14), respectively. □

Next, we show that the Newton direction Δw is a descent direction for the merit function F. For this purpose, we first show the following two lemmas.

Lemma 3.3.4. *Let $\mu > 0$ and $w = (x, y, Z) \in \mathcal{W}$. Suppose that $G + H$ is positive definite. Let Δx be given in Theorem 3.3.4. Then we have*

$$\nabla F_{BP}(x)^\top \Delta x = -\Delta x^\top \left(G+H+\frac{1}{\mu}J_g(x)^\top J_g(x)\right)\Delta x \le 0.$$

Furthermore, $\nabla F_{BP}(x)^\top \Delta x = 0$ if and only if $\Delta x = 0$.

Proof. We easily see that $G+H+\frac{1}{\mu}J_g(x)^\top J_g(x)$ is positive definite from the positive definiteness of $G+H$. Since $\nabla F_{BP}(x) = \nabla f(x) + \frac{1}{\mu}J_g(x)^\top g(x) - \mu\mathcal{A}^*(x)X^{-1}$ from Theorem 3.3.1, it then follows from (3.3.15) that

$$\begin{aligned}\nabla F_{BP}(x)^\top \Delta x &= \Delta x^\top\left(\nabla f(x)+\frac{1}{\mu}J_g(x)^\top g(x)-\mu\mathcal{A}^*(x)X^{-1}\right)\\ &= -\Delta x^\top\left(G+H+\frac{1}{\mu}J_g(x)^\top J_g(x)\right)\Delta x\\ &\le 0.\end{aligned}$$

Furthermore, since $G+H+\frac{1}{\mu}J_g(x)^\top J_g(x)$ is positive definite, $\nabla F_{BP}(x)^\top \Delta x = 0$ if and only if $\Delta x = 0$. □

Lemma 3.3.5. *Let $\mu > 0$ and $w = (x, y, Z) \in \mathcal{W}$. Let $\Delta w = (\Delta x, \Delta y, \Delta Z)$ be given in Theorem 3.3.4. Then we have*

$$\langle \nabla F_{PD}(w), \Delta w\rangle = -\frac{1}{\mu}\|g(x)+\mu y\|^2 - \left\|(\widetilde{X}\widetilde{Z})^{-\frac{1}{2}}(\mu I - \widetilde{X}\widetilde{Z})\right\|_F^2 \le 0.$$

Furthermore, $\langle \nabla F_{PD}(w), \Delta w\rangle = 0$ if and only if $g(x) + \mu y = 0$ and $XZ - \mu I = 0$.

Proof. From Theorem 3.3.1, we obtain

$$\begin{aligned}\langle \nabla F_{PD}(w), \Delta w\rangle &= \langle \nabla_x F_{PD}(w), \Delta x\rangle + \langle \nabla_y F_{PD}(w), \Delta y\rangle + \langle \nabla_Z F_{PD}(w), \Delta Z\rangle\\ &= \frac{1}{\mu}\Delta x^\top J_g(x)^\top (g(x)+\mu y) + \Delta x^\top \mathcal{A}^*(x)(Z-\mu X^{-1})\\ &\qquad +(g(x)+\mu y)^\top \Delta y + \left\langle X-\mu Z^{-1}, \Delta Z\right\rangle. \qquad (3.3.18)\end{aligned}$$

On the other hand, we have from the definitions of $\mathcal{A}^*(x)$ and ΔX that

$$\begin{aligned}\Delta x^\top \mathcal{A}^*(x)(Z-\mu X^{-1}) &= \sum_{i=1}^n \Delta x_i \left\langle A_i(x), Z-\mu X^{-1}\right\rangle\\ &= \left\langle \sum_{i=1}^n \Delta x_i A_i(x), Z-\mu X^{-1}\right\rangle\\ &= \left\langle \mathcal{A}(x)\Delta x, Z-\mu X^{-1}\right\rangle\\ &= \left\langle \Delta X, Z-\mu X^{-1}\right\rangle. \qquad (3.3.19)\end{aligned}$$

From Proposition 2.2.9 (b) and (c), we have

$$\begin{aligned}\left\langle \Delta X, Z-\mu X^{-1}\right\rangle &= \left\langle (T\odot T)^{-1}(T\odot T)\Delta X, Z-\mu X^{-1}\right\rangle \\ &= \left\langle (T^{-1}\odot T^{-1})(T\odot T)\Delta X, Z-\mu X^{-1}\right\rangle \\ &= \left\langle (T\odot T)\Delta X, (T^{-\top}\odot T^{-\top})(Z-\mu X^{-1})\right\rangle.\end{aligned}$$

Moreover, since $\widetilde{X}^{-1} = ((T\odot T)X)^{-1} = (TXT^{\top})^{-1} = T^{-\top}X^{-1}T^{-1} = (T^{-\top}\odot T^{-\top})X^{-1}$, we obtain

$$\begin{aligned}\left\langle \Delta X, Z-\mu X^{-1}\right\rangle &= \left\langle \Delta\widetilde{X}, \widetilde{Z}-\mu\widetilde{X}^{-1}\right\rangle \\ &= \left\langle \Delta\widetilde{X}, (I-\mu\widetilde{X}^{-1}\widetilde{Z}^{-1})\widetilde{Z}\right\rangle \\ &= \mathrm{tr}\left[\Delta\widetilde{X}(I-\mu\widetilde{X}^{-1}\widetilde{Z}^{-1})\widetilde{Z}\right] \\ &= \mathrm{tr}\left[(I-\mu\widetilde{X}^{-1}\widetilde{Z}^{-1})\widetilde{Z}\Delta\widetilde{X}\right] \\ &= \left\langle I-\mu\widetilde{X}^{-1}\widetilde{Z}^{-1}, \widetilde{Z}\Delta\widetilde{X}\right\rangle. \end{aligned} \tag{3.3.20}$$

Since $\widetilde{X}$ and $\widetilde{Z}$ commute, $\widetilde{X}^{-1}$ and $\widetilde{Z}^{-1}$ also commute. Then we also get

$$\begin{aligned}\left\langle \Delta X, Z-\mu X^{-1}\right\rangle &= \left\langle \widetilde{Z}-\mu\widetilde{X}^{-1}, \Delta\widetilde{X}\right\rangle \\ &= \mathrm{tr}(\widetilde{Z}\left[I-\mu\widetilde{Z}^{-1}\widetilde{X}^{-1})\Delta\widetilde{X}\right] \\ &= \mathrm{tr}(\widetilde{Z}\left[I-\mu\widetilde{X}^{-1}\widetilde{Z}^{-1})\Delta\widetilde{X}\right] \\ &= \mathrm{tr}\left[(I-\mu\widetilde{X}^{-1}\widetilde{Z}^{-1})\Delta\widetilde{X}\widetilde{Z}\right] \\ &= \left\langle I-\mu\widetilde{X}^{-1}\widetilde{Z}^{-1}, \Delta\widetilde{X}\widetilde{Z}\right\rangle. \end{aligned} \tag{3.3.21}$$

From (3.3.20) and (3.3.21), we obtain

$$\begin{aligned}\left\langle \Delta X, Z-\mu X^{-1}\right\rangle &= \frac{1}{2}\left\langle \Delta X, Z-\mu X^{-1}\right\rangle + \frac{1}{2}\left\langle \Delta X, Z-\mu X^{-1}\right\rangle \\ &= \frac{1}{2}\left\langle I-\mu\widetilde{X}^{-1}\widetilde{Z}^{-1}, \widetilde{Z}\Delta\widetilde{X}\right\rangle + \frac{1}{2}\left\langle I-\mu\widetilde{X}^{-1}\widetilde{Z}^{-1}, \Delta\widetilde{X}\widetilde{Z}\right\rangle. \end{aligned} \tag{3.3.22}$$

In a way similar to prove (3.3.22), we also have

$$\left\langle X-\mu Z^{-1}, \Delta Z\right\rangle = \frac{1}{2}\left\langle I-\mu\widetilde{X}^{-1}\widetilde{Z}^{-1}, \widetilde{X}\Delta\widetilde{Z}\right\rangle + \frac{1}{2}\left\langle I-\mu\widetilde{X}^{-1}\widetilde{Z}^{-1}, \Delta\widetilde{Z}\widetilde{X}\right\rangle. \tag{3.3.23}$$

From (3.3.18), (3.3.19), (3.3.22) and (3.3.23), we obtain

$$\begin{aligned}\left\langle \nabla F_{PD}(w), \Delta w\right\rangle &= \frac{1}{\mu}(g(x)+\mu y)^{\top}(J_g(x)\Delta x+\mu\Delta y) \\ &\quad +\frac{1}{2}\left\langle I-\mu\widetilde{X}^{-1}\widetilde{Z}^{-1}, \widetilde{Z}\Delta\widetilde{X}+\Delta\widetilde{X}\widetilde{Z}+\widetilde{X}\Delta\widetilde{Z}+\Delta\widetilde{Z}\widetilde{X}\right\rangle. \end{aligned} \tag{3.3.24}$$

Then, by substituting (3.3.9) and (3.3.10) into (3.3.24), we have

$$\left\langle \nabla F_{PD}(w), \Delta w\right\rangle = -\frac{1}{\mu}\|g(x)+\mu y\|^2 + \left\langle I-\mu\widetilde{X}^{-1}\widetilde{Z}^{-1}, \mu I-\widetilde{X}\widetilde{Z}\right\rangle \tag{3.3.25}$$

Note that since $\widetilde{X}$ and $\widetilde{Z}$ are symmetric positive definite and commute, $\widetilde{X}\widetilde{Z}$ is symmetric positive definite, and hence there exists $(\widetilde{X}\widetilde{Z})^{-\frac{1}{2}}$. Let $Y := \widetilde{X}\widetilde{Z}$. We get

$$\begin{aligned}\left\langle I - \mu\widetilde{X}^{-1}\widetilde{Z}^{-1}, \mu I - \widetilde{X}\widetilde{Z}\right\rangle &= -\mathrm{tr}\left[Y^{-1}(\mu I - Y)(\mu I - Y)\right] \\ &= -\mathrm{tr}\left[Y^{-\frac{1}{2}}(\mu I - Y)(\mu I - Y)Y^{-\frac{1}{2}}\right] \\ &= -\left\|Y^{-\frac{1}{2}}(\mu I - Y)\right\|_F^2 \\ &= -\left\|(\widetilde{X}\widetilde{Z})^{-\frac{1}{2}}(\mu I - \widetilde{X}\widetilde{Z})\right\|_F^2. \end{aligned} \tag{3.3.26}$$

Thus, we have from (3.3.25) and (3.3.26) that

$$\langle \nabla F_{PD}(w), \Delta w\rangle = -\frac{1}{\mu}\|g(x) + \mu y\|^2 - \left\|(\widetilde{X}\widetilde{Z})^{-\frac{1}{2}}(\mu I - \widetilde{X}\widetilde{Z})\right\|_F^2 \leq 0.$$

Furthermore, we can easily see that $\langle \nabla F_{PD}(w), \Delta w\rangle = 0$ if and only if $g(x) + \mu y = 0$ and $XZ - \mu I = 0$. □

Now, we show that the Newton direction Δw is a descent direction for the merit function F.

Theorem 3.3.5. *Let $\mu > 0$ and $w = (x, y, Z) \in \mathcal{W}$. Assume that $G + H$ is positive definite. Then, $\Delta w = (\Delta x, \Delta y, \Delta Z)$ given in Theorem 3.3.4 is a descent direction for the merit function F, i.e.,*

$$\begin{aligned}\langle \nabla F(w), \Delta w\rangle &= -\Delta x^\top \left(G + H + \frac{1}{\mu}J_g(x)^\top J_g(x)\right)\Delta x \\ &\qquad - \frac{\nu}{\mu}\|g(x) + \mu y\|^2 - \nu\left\|(\widetilde{X}\widetilde{Z})^{-\frac{1}{2}}(\mu I - \widetilde{X}\widetilde{Z})\right\|_F^2 \\ &\leq 0.\end{aligned}$$

Furthermore, $\langle \nabla F(w), \Delta w\rangle = 0$ if and only if w is a shifted barrier KKT point.

Proof. Note that

$$\langle \nabla F(w), \Delta w\rangle = \nabla F_{BP}(x)^\top \Delta x + \nu \langle \nabla F_{PD}(w), \Delta w\rangle. \tag{3.3.27}$$

Then, we have from Lemmas 3.3.4 and 3.3.5 that

$$\begin{aligned}\langle \nabla F(w), \Delta w\rangle &= -\Delta x^\top \left(G + H + \frac{1}{\mu}J_g(x)^\top J_g(x)\right)\Delta x \\ &\qquad - \frac{\nu}{\mu}\|g(x) + \mu y\|^2 - \nu\left\|(\widetilde{X}\widetilde{Z})^{-\frac{1}{2}}(\mu I - \widetilde{X}\widetilde{Z})\right\|_F^2 \\ &\leq 0.\end{aligned}$$

Now, we show the second part of this theorem. Suppose that w is a shifted barrier KKT point, i.e., $\nabla f(x) - J_g(x)^\top y - \mathcal{A}^*(x)Z = 0, g(x) + \mu y = 0$ and $XZ - \mu I = 0$. Then we have

$$\nabla f(x) + \frac{1}{\mu}J_g(x)^\top g(x) - \mu\mathcal{A}^*(x)X^{-1} = \nabla f(x) - J_g(x)^\top y - \mathcal{A}^*(x)Z = 0.$$

It then follows from (3.3.15) and the regularity of $G+H+\frac{1}{\mu}J_g(x)^\top J_g(x)$ that $\Delta x = 0$, and hence we have from Lemma 3.3.4 that $\nabla F_{BP}(x)^\top \Delta x = 0$. Moreover, $g(x) + \mu y = 0, XZ - \mu I = 0$ and Lemma 3.3.5 imply that $\langle \nabla F_{PD}(w), \Delta w\rangle = 0$. Therefore, $\langle \nabla F(w), \Delta w\rangle = 0$ from (3.3.27).

Conversely, suppose that $\langle \nabla F(w), \Delta w\rangle = 0$. Since it follows from Lemmas 3.3.4 and 3.3.5 that $\nabla F_{BP}(x)^\top \Delta x \le 0$ and $\langle \nabla F_{PD}(w), \Delta w\rangle \le 0$, the equation (3.3.27) implies that $\nabla F_{BP}(x)^\top \Delta x = 0$ and $\langle \nabla F_{PD}(w), \Delta w\rangle = 0$. It further follows from Lemmas 3.3.4 and 3.3.5 that $\Delta x = 0, g(x) + \mu y = 0$ and $XZ - \mu I = 0$. Then we have

$$\nabla_x L(w) = \nabla f(x) - J_g(x)^\top y - \mathcal{A}^*(x)Z = \nabla f(x) + \frac{1}{\mu} J_g(x)^\top g(x) - \mu \mathcal{A}^*(x) X^{-1} = 0,$$

where the last equality follows from (3.3.15). Thus, w is a shifted barrier KKT point. □

Theorem 3.3.5 guarantees that $F(w + \alpha\Delta w) < F(w)$ for sufficiently small $\alpha > 0$ if w is not a shifted barrier KKT point.

Now, we discuss how to choose an appropriate step size α such that $F(w + \alpha\Delta w) < F(w)$. The merit function F and the Newton equations (3.3.8)–(3.3.10) are well-defined only on $\mathcal{W}$. Therefore, the new point $w + \alpha\Delta w$ is required to be an interior point. Thus, we must choose a step size $\alpha \in (0, 1]$ such that $X(x + \alpha\Delta x) \succ 0$ and $Z + \alpha\Delta Z \succ 0$. To this end, we first calculate

$$\overline{\alpha}_x := \begin{cases} -\dfrac{\tau}{\lambda_{\min}(X^{-\frac{1}{2}}\Delta X X^{-\frac{1}{2}})} & \text{if } \lambda_{\min}(X^{-\frac{1}{2}}\Delta X X^{-\frac{1}{2}}) < 0 \text{ and } X \text{ is affine,} \\ 1 & \text{otherwise,} \end{cases}$$

and

$$\overline{\alpha}_z := \begin{cases} -\dfrac{\tau}{\lambda_{\min}(Z^{-\frac{1}{2}}\Delta Z Z^{-\frac{1}{2}})} & \text{if } \lambda_{\min}(Z^{-\frac{1}{2}}\Delta Z Z^{-\frac{1}{2}}) < 0, \\ 1 & \text{otherwise,} \end{cases}$$

where $\tau \in (0, 1)$ is a given constant. Set

$$\bar{\alpha} := \min\{1, \bar{\alpha}_x, \bar{\alpha}_z\}. \tag{3.3.28}$$

Then $Z + \alpha\Delta Z \succ 0$ for any $\alpha \in (0, \bar{\alpha}]$. Moreover, $X(x + \alpha\Delta x) \succ 0$ for any $\alpha \in (0, \bar{\alpha}]$ if X is affine. Note that if X is nonlinear, $X(x + \alpha\Delta x)$ is not necessarily positive definite for any $\alpha \in (0, \bar{\alpha}]$.

Next, we choose a step size $\alpha \in (0, \bar{\alpha}]$ such that $F(w + \alpha\Delta w) < F(w)$ and $X(x + \alpha\Delta x) \succ 0$. For this purpose, we adopt Armijo's line search rule which finds the smallest nonnegative integer l such that

$$F(w + \bar{\alpha}\beta^l \Delta w) \le F(w) + \varepsilon_0 \bar{\alpha}\beta^l \langle \nabla F(w), \Delta w\rangle, \quad X(x + \bar{\alpha}\beta^l \Delta x) \succ 0,$$

where $\beta, \varepsilon_0 \in (0, 1)$ are given constants. Then, set $\alpha := \bar{\alpha}\beta^l$. Note that the second condition is not necessary when X is affine.

Now, we describe a concrete Newton-type method for Step 1 of Algorithm 3.2.1. Recall that the script k denotes the k-th iteration of Algorithm 3.2.1.

Algorithm 3.3.1. *(for Step 2 of Algorithm 3.2.1)*

Step 0. *Choose* $\beta, \varepsilon_0, \tau \in (0,1)$ *and set* $j := 0$ *and* $w_0 := v_k$.

Step 1. *If* $\|r_1(w_j, \mu_k)\| \leq \sigma\mu_k$, *then set* $v_{k+1} := w_j$ *and return.*

Step 2. *Obtain the Newton direction* $\Delta w_j = (\Delta x_j, \Delta y_j, \Delta Z_j)$ *by solving the Newton equations (3.3.8)–(3.3.10).*

Step 3. *Set* $\alpha_j := \bar{\alpha}_j \beta^{l_j}$, *where* $\bar{\alpha}_j$ *is given by (3.3.28) and* l_j *is the smallest nonnegative integer such that*

$$F(w_j + \bar{\alpha}_j \beta^{l_j} \Delta w_j) \leq F(w_j) + \varepsilon_0 \bar{\alpha}_j \beta^{l_j} \langle \nabla F(w_j), \Delta w_j \rangle, \quad X(x_j + \bar{\alpha}_j \beta^{l_j} \Delta x_j) \succ 0.$$

Step 4. *Set* $w_{j+1} := w_j + \alpha_j \Delta w_j$ *and* $j := j + 1$, *and go to Step 1.* □

3.3.3 Global convergence of Algorithm 3.3.1

In this subsection, we prove the global convergence of Algorithm 3.3.1. For this purpose, we make the following assumptions.

Assumption 3.3.1.

(A1) *The functions* $f, g_1, \ldots, g_m$ *and* X *are twice continuously differentiable.*

(A2) *The sequence* $\{x_j\}$ *generated by Algorithm 3.3.1 remains in some compact set* Ω *of* $\mathbf{R}^n$.

(A3) *The sequence* $\{G_j + H_j + \frac{1}{\mu} J_g(x_j)^\top J_g(x_j)\}$ *is uniformly positive definite and the sequence* $\{G_j\}$ *is bounded.*

(A4) *The sequences* $\{T_j\}$ *and* $\{T_j^{-1}\}$ *are bounded.*

Note that Assumption 3.3.1 (A2) holds under the assumptions of Theorem 3.3.3. Assumption 3.3.1 (A3) guarantees that the Newton equations (3.3.8)–(3.3.10) have a unique solution.

Remark 3.3.6. *Assumption 3.3.1 (A1)–(A3) hold for linear SDP such that* $A_1(x_j), \ldots, A_n(x_j)$ *are linearly independent. In fact, it is clear that Assumption 3.3.1 (A1) holds. Theorem 3.3.3 guarantees that Assumption 3.3.1 (A2) holds. Moreover,* H_j *is positive definite from Remark 3.3.5 and* $G_j = 0$. *Thus, Assumption 3.3.1 (A3) holds.*

Remark 3.3.7. *Yamashita, Yabe and Harada [72] showed that their Newton-type method is convergent globally to a barrier KKT point satisfying (2.4.1) under the boundedness of the sequence* $\{y_j\}$, *in addition to Assumption 3.3.1 (A1)–(A4). However, they did not give sufficient conditions for the boundedness of* $\{y_j\}$.

Remark 3.3.8. *Kato, Yabe and Yamashita [34] also showed that a Newton-type method with the merit function* $\widetilde{F}$ *can find a shifted barrier KKT point under Assumption 3.3.1 (A1)–(A4). However, concrete sufficient conditions were not stated for Assumption 3.3.1 (A2).*

First of all, we show that the sequence $\{w_j\}$ generated by Algorithm 3.3.1 is bounded. Note that $X(x_j)$ and $\mathcal{A}(x_j)\Delta x_j$ are denoted by X_j and ΔX_j, respectively, for simplicity.

Lemma 3.3.6. *Suppose that Assumption 3.3.1 (A2) is satisfied. Then, the sequence $\{w_j = (x_j, y_j, Z_j)\}$ generated by Algorithm 3.3.1 is bounded. Furthermore, the sequences $\{X_j\}$ and $\{Z_j\}$ are uniformly positive definite.*

Proof. Since the sequence $\{F(w_j)\}$ is monotonically decreasing, we have $F(w_j) \leq F(w_0)$ for all j. Then, we obtain the desired results from Assumption 3.3.1 (A2) and Lemma 3.3.2. □

Note that Lemma 3.3.6 guarantees that Assumption 3.3.1 (A4) holds if T_j is given by HRVW/KSH/M choice, that is, $T_j = X_j^{-\frac{1}{2}}$ or NT choice, that is, $T_j = W_j^{-\frac{1}{2}}$, where $W_j := X^{\frac{1}{2}}(X^{\frac{1}{2}}ZX^{\frac{1}{2}})^{-\frac{1}{2}}X^{\frac{1}{2}}$. See Section 2.4 for details of the scaling matrix T_j.

Lemma 3.3.7. *Suppose that Assumption 3.3.1 (A2)–(A4) are satisfied. Then, the sequence $\{\Delta w_j\}$ generated by Algorithm 3.3.1 is bounded.*

Proof. It follows from Assumption 3.3.1 (A2)–(A4), Lemma 3.3.6 and Theorem 3.3.4 that the sequence $\{\Delta w_j\}$ generated by Algorithm 3.3.1 is bounded. □

We now show the global convergence of Algorithm 3.3.1. We assume that Algorithm 3.3.1 generates an infinite sequence and w_j is not a shifted barrier KKT point for all j.

Theorem 3.3.6. *Suppose that Assumption 3.3.1 (A1)–(A4) are satisfied. Then, the sequence $\{w_j = (x_j, y_j, Z_j)\}$ generated by Algorithm 3.3.1 has an accumulation point. Moreover, any accumulation point $w^* = (x^*, y^*, Z^*)$ is a shifted barrier KKT point.*

Proof. Since the sequence $\{w_j\}$ is bounded from Lemma 3.3.6, it has at least one accumulation point.

Next, we prove that any accumulation point of the sequence $\{w_j\}$ is a shifted barrier KKT point. To this end, we first show that the sequence $\{\bar{\alpha}_j\}$ generated in Step 3 is away from zero, that is, there exists a real number $\bar{\alpha}$ such that $0 < \bar{\alpha} \leq \bar{\alpha}_j$ for all j. Note that from Lemmas 3.3.6 and 3.3.7, the sequences $\{X_j\}$, $\{Z_j\}$, $\{\Delta X_j\}$ and $\{\Delta Z_j\}$ are bounded. Moreover, the sequences $\{X_j\}$ and $\{Z_j\}$ are uniformly positive definite. Thus, the sequences $\{\lambda_{\min}(X_j^{-\frac{1}{2}}\Delta X_j X_j^{-\frac{1}{2}})\}$ and $\{\lambda_{\min}(Z_j^{-\frac{1}{2}}\Delta Z_j Z_j^{-\frac{1}{2}})\}$ are also bounded. It then follows from the definition of $\bar{\alpha}_j$ that there exists a real number $\bar{\alpha}$ such that $0 < \bar{\alpha} \leq \bar{\alpha}_j$ for all j.

Next, we show $\langle \nabla F(w_j), \Delta w_j \rangle \to 0$ as $j \to \infty$. From Armijo's line search strategy in Step 3, we have

$$F(w_{j+1}) - F(w_j) \leq \varepsilon_0 \bar{\alpha}_j \beta^{l_j} \langle \nabla F(w_j), \Delta w_j \rangle .$$

Summing up the above inequality from $j = 1$ to $j = \widetilde{j}$, we have

$$F(w_{\widetilde{j}+1}) - F(w_1) \leq \varepsilon_0 \sum_{j=1}^{\widetilde{j}} \bar{\alpha}_j \beta_j \langle \nabla F(w_j), \Delta w_j \rangle .$$

Since $\langle \nabla F(w_j), \Delta w_j \rangle \leq 0$ by Theorem 3.3.5, it follows from $\bar{\alpha} \leq \bar{\alpha}_j$ that

$$F(w_{\widetilde{j}+1}) - F(w_1) \leq \varepsilon_0 \bar{\alpha} \sum_{j=1}^{\widetilde{j}} \beta^{l_j} \langle \nabla F(w_j), \Delta w_j \rangle .$$

The boundedness of the sequence $\{w_j\}$ implies that the sequence $\{F(w_j)\}$ is bounded, and hence

$$-\infty < \sum_{j=1}^{\infty} \beta^{l_j} \langle \nabla F(w_j), \Delta w_j \rangle \leq 0.$$

Then, we have

$$\lim_{j\to\infty} \beta^{l_j} \langle \nabla F(w_j), \Delta w_j \rangle = 0.$$

Now, we consider two cases: $\liminf_{j\to\infty} \beta^{l_j} > 0$ and $\liminf_{j\to\infty} \beta^{l_j} = 0$.

Case 1: $\liminf_{j\to\infty} \beta^{l_j} > 0$. Then, we have

$$\lim_{j\to\infty} \langle \nabla F(w_j), \Delta w_j \rangle = 0.$$

Case 2: $\liminf_{j\to\infty} \beta^{l_j} = 0$. In this case, there exists a subsequence $\{l_j\}_{\mathcal{J}}$ that diverges to ∞, where $\mathcal{J} \subset \{0, 1, \cdots\}$. Since the sequence $\{X(x_j)\}$ is uniformly positive definite by Lemma 3.3.6, there exists $\bar{l} \geq 0$ such that $X(x_j + \bar{\alpha}_j \beta^l \Delta x_j) \succ 0$ for all $l > \bar{l}$. Then, without loss of generality, we suppose that $l_j \geq \bar{l}$ for all $j \in \mathcal{J}$, that is, $X(x_j + \bar{\alpha}_j \beta^{l_j - 1} \Delta x_j) \succ 0$ for all $j \in \mathcal{J}$. Since $l_j - 1$ does not satisfy Armijo's line search rule in Step 3,

$$\varepsilon_0 t_j \langle \nabla F(w_j), \Delta w_j \rangle < F(w_j + t_j \Delta w_j) - F(w_j),$$

where $t_j := \bar{\alpha}_j \beta^{l_j - 1}$. Let $h(t) := F(w_j + t \Delta w_j)$. It then follows from Theorem 2.2.1 that there exists $\theta_j \in (0, 1)$ such that

$$\begin{aligned} \varepsilon_0 t_j \langle \nabla F(w_j), \Delta w_j \rangle &< F(w_j + t_j \Delta w_j) - F(w_j) \\ &= h(t_j) - h(0) \\ &= t_j h'(\theta_j t_j) \\ &= t_j \langle \nabla F(w_j + \theta_j t_j \Delta w_j), \Delta w_j \rangle , \end{aligned}$$

and hence

$$\begin{aligned} 0 < (\varepsilon_0 - 1) \langle \nabla F(w_j), \Delta w_j \rangle &< \langle \nabla F(w_j + \theta_j t_j \Delta w_j) - \nabla F(w_j), \Delta w_j \rangle \\ &\leq \| \nabla F(w_j + \theta_j t_j \Delta w_j) - \nabla F(w_j) \| \| \Delta w_j \|, \end{aligned} \tag{3.3.29}$$

where the last inequality follows from the Cauchy-Schwarz inequality. Since the subsequence $\{t_j\}_{\mathcal{J}}$ converges to 0, we have from Assumption 3.3.1 (A1) and the boundedness of the sequences $\{w_j\}$ and $\{\Delta w_j\}$ that

$$\lim_{j\to\infty, j\in\mathcal{J}} \| \nabla F(w_j + \theta_j t_j \Delta w_j) - \nabla F(w_j) \| = 0.$$

It then follows from (3.3.29) that

$$\lim_{j\to\infty, j\in\mathcal{J}} \langle \nabla F(w_j), \Delta w_j \rangle = 0.$$

From the both cases, we can conclude that

$$\lim_{j\to\infty} \langle \nabla F(w_j), \Delta w_j \rangle = 0. \tag{3.3.30}$$

By the boundedness of the sequence $\{w_j\}$ and Assumption 3.3.1 (A3) and (A4), there exist subsequences $\{w_j\}_{\mathcal{K}}, \{G_j\}_{\mathcal{K}}$ and $\{T_j\}_{\mathcal{K}}$ such that

$$\lim_{j\to\infty, j\in\mathcal{K}} w_j =: w^*, \quad \lim_{j\to\infty, j\in\mathcal{K}} G_j =: G^*, \quad \lim_{j\to\infty, j\in\mathcal{K}} T_j =: T^*,$$

where $\mathcal{K} \subset \{0, 1, \ldots\}$, $w^* \in \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{S}^p$, $G^* \in \mathbf{S}^n$ and $T^* \in \mathbf{R}^{p\times p}$. Moreover, by the definitions of the operators $T_j \odot T_j$ and $T_j^\top \odot T_j^\top$, the subsequences $\{T_j \odot T_j\}_{\mathcal{K}}$ and $\{T_j^\top \odot T_j^\top\}_{\mathcal{K}}$ converge to $T^* \odot T^*$ and $(T^*)^\top \odot (T^*)^\top$, respectively. Therefore, we have from (3.3.16) that there exists $H^* \in \mathbf{S}^n$ such that

$$\lim_{j\to\infty, j\in\mathcal{K}} H_j = H^*.$$

Note that $G^* + H^* + \frac{1}{\mu} J_g(x^*)^\top J_g(x^*)$ is positive definite from Assumption 3.3.1 (A3). It then follows from Theorem 3.3.4 that the subsequence $\{\Delta x_j\}_{\mathcal{K}}$ converges to $\Delta x^* \in \mathbf{R}^n$, where

$$\Delta x^* := -\left(G^* + H^* + \frac{1}{\mu} J_g(x^*)^\top J_g(x^*)\right)^{-1} \left(\nabla f(x^*) + \frac{1}{\mu} J_g(x^*)^\top g(x^*) - \mu \mathcal{A}^*(x^*) X(x^*)^{-1}\right).$$

Similarly, $\{\Delta y_j\}_{\mathcal{K}}$ and $\{\Delta Z_j\}_{\mathcal{K}}$ converge to $\Delta y^* \in \mathbf{R}^m$ and $\Delta Z^* \in \mathbf{S}^p$, where

$$\begin{aligned} \Delta y^* &:= -\frac{1}{\mu}(g(x^*) + \mu y^* + J_g(x^*)\Delta x^*), \\ \Delta Z^* &:= \mu X(x^*)^{-1} - Z^* - ((T^*)^\top \odot (T^*)^\top)(\widetilde{X}(x^*) \odot I)^{-1}(\widetilde{Z}^* \odot I)(T^* \odot T^*)\mathcal{A}(x^*)\Delta x^*, \end{aligned}$$

and $\widetilde{Z}^* := ((T^*)^{-\top} \odot (T^*)^{-\top}) Z^*$. It then follows from (3.3.30) that

$$\langle \nabla F(w^*), \Delta w^* \rangle = 0,$$

where $\Delta w^* := (\Delta x^*, \Delta y^*, \Delta Z^*)$. Then, Theorem 3.3.5 yields that

$$\nabla_x L(w^*) = 0, \quad g(x^*) + \mu y^* = 0 \quad \text{and} \quad X(x^*)Z^* - \mu I = 0.$$

Therefore, w^* is a shifted barrier KKT point. □

3.4 Numerical experiments

In this section, we report some numerical experiments for Algorithm 3.2.1 with Algorithm 3.3.1. We compare the proposed algorithm with the interior point method [72] based on the barrier KKT conditions (2.4.1). We present the number of iterations and the CPU time of both algorithms. The program is written in MATLAB R2010a and run on a machine with an Intel Core i7 920 2.67GHz CPU and 3.00GB RAM. The barrier parameter μ_k is updated by $\mu_{k+1} = \mu_k/10$ with $\mu_0 = 0.1$. The approximate Hessian G_k is updated by the method described in [72, Remark 3]. We employ the HRVW/KSH/M choice as the scaling matrix T_j, that is $T_j = X_j^{-\frac{1}{2}}$. Moreover, we select the following parameters:

$$\epsilon = 10^{-4}, \quad \sigma = 5.0, \quad \nu = 1.0, \quad \tau = 0.95, \quad \beta = 0.95, \quad \varepsilon_0 = 0.50.$$

We solve the following three test problems described in [72] by using the initial points indicated in [72].

Gaussian channel capacity problem:

$$\begin{array}{ll} \text{maximize} & \dfrac{1}{2}\displaystyle\sum_{i=1}^{n}\log(1+t_i), \\ \text{subject to} & \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}X_{ii}\le P,\ X_{ii}\ge 0,\ t_i\ge 0, \\ & \begin{bmatrix} 1-a_it_i & \sqrt{r_i} \\ \sqrt{r_i} & a_iX_{ii}+r_i \end{bmatrix}\succeq 0,\ i=1,\ldots,n, \end{array}$$

where decision variables are X_{ii} and t_i for $i = 1, \ldots, n$. In the experiment, the constants r_i and a_i for $i = 1, \ldots, n$ are selected randomly from the interval $[0, 1]$, and P is set to 1. Note that the objective function of the problem is concave and the constraint functions are affine.

Minimization of the minimal eigenvalue problem:

$$\begin{array}{ll} \text{minimize} & \operatorname{tr}(\Pi M(q)), \\ \text{subject to} & \operatorname{tr}(\Pi)=1, \\ & \Pi\succeq 0, \\ & q\in Q, \end{array}$$

where $Q \subset \mathbf{R}^p$, and M is a function from $\mathbf{R}^p$ to $\mathbf{S}^n$, and decision variables are $q \in \mathbf{R}^p$ and $\Pi \in \mathbf{S}^n$. In the experiment, p is set to 2 and M is given by $M(q) := q_1q_2M_1 + q_1M_2 + q_2M_3$, where $M_1, M_2, M_3 \in \mathbf{S}^n$ are constant matrices whose elements are selected randomly from the interval $[-1, 1]$. Moreover, Q is set to $[-1, 1] \times [-1, 1]$. Note that the objective function is nonconvex and the constraint functions are affine.

Nearest correlation matrix problem:

$$\begin{array}{ll} \text{minimize} & \dfrac{1}{2}\left\|X-A\right\|_F^2, \\ \text{subject to} & X\succeq \eta I, \\ & X_{ii}=1,\ i=1,\ldots,n, \end{array}$$

where decision variable is $X \in \mathbf{S}^n$, and $A \in \mathbf{S}^n$ is a constant matrix and $\eta \in \mathbf{R}$ is a positive constant. In the experiment, the elements of the matrix A are selected randomly from the interval $[-1, 1]$ with $A_{ii} = 1$ for $i = 1, \ldots, n$. Moreover, we set $\eta = 10^{-3}$. Note that the objective function is quadratic and convex, and the constraint functions are affine. Therefore, the problem is convex.

The numerical results are presented in Tables 3.1–3.3. In these tables, SDPIP denotes the primal-dual interior point method of [72]. From Tables 3.1–3.3, we see that the results obtained by using Algorithm 3.2.1 are comparable to those produced with SDPIP.

Table 3.1: Gaussian channel capacity problem

	SDPIP		Algorithm 3.2.1	
n	iteration	time(s)	iteration	time(s)
5	4	0.26	4	0.24
10	4	0.56	4	0.54
15	4	1.51	4	1.45
20	4	4.25	4	4.16
25	5	10.80	5	10.12
30	5	19.60	5	19.54
35	5	34.41	5	33.53
40	5	61.17	5	60.29

Table 3.2: Minimization of the minimal eigenvalue problem

	SDPIP		Algorithm 3.2.1	
n	iteration	time(s)	iteration	time(s)
5	4	0.29	4	0.33
10	4	6.76	4	6.87
15	4	30.85	4	31.56
20	4	67.13	5	84.10
25	4	308.13	5	417.32
30	4	781.34	5	1000.53
35	4	2804.10	5	3601.28
40	4	4339.71	5	5395.89

Table 3.3: Nearest correlation matrix problem

	SDPIP		Algorithm 3.2.1	
n	iteration	time(s)	iteration	time(s)
5	4	0.07	4	0.07
10	4	0.47	4	0.50
15	4	2.72	4	2.77
20	4	9.16	4	9.56
25	4	35.25	5	40.57
30	4	75.07	5	91.66
35	4	165.63	5	213.12
40	4	291.10	5	396.64

3.5 Concluding remarks

In this chapter, we proposed a new merit function F for the shifted barrier KKT conditions. We also showed the properties of the merit function F. In particular, we gave the level boundedness of the merit function F, which is not given in other related papers for nonlinear SDP. Moreover, we proposed Algorithm 3.3.1 to find an approximate shifted barrier KKT point, and proved its global convergence under weaker assumptions than those in [72]. In the numerical experiments, we showed that the performance of Algorithm 3.2.1 was comparable to that of the primal-dual interior point method [72] based on the barrier KKT conditions.

In this chapter, there is no discussion related to an update rule of the barrier parameter μ_k. Thus, a future work is to give its reasonable update rule.

Chapter 4

A two-step primal-dual interior point method for nonlinear semidefinite programming problems and its superlinear convergence

4.1 Introduction

In this chapter, we consider the following nonlinear semidefinite programming (SDP) problem:

$$\begin{array}{ll} \underset{x \in \mathbf{R}^n}{\text{minimize}} & f(x), \\ \text{subject to} & g(x) = 0, \quad X(x) \succeq 0, \end{array} \tag{4.1.1}$$

where $f : \mathbf{R}^n \to \mathbf{R}, g : \mathbf{R}^n \to \mathbf{R}^m, X : \mathbf{R}^n \to \mathbf{S}^p$ are twice continuously differentiable functions.

In this chapter, we propose a two-step primal-dual interior point method and show its local and superlinear convergence. The proposed method is based on the generalized shifted barrier KKT conditions. It solves two Newton equations derived from the generalized shifted barrier KKT conditions in each iteration. However, in order to reduce calculations, we replace the coefficient matrix in the second equation with that in the first one. Thus, we can solve the second equation more rapidly using some computational results obtained by solving the first equation. Despite this change, we show the superlinear convergence under the same assumptions as those in Yamashita and Yabe [71].

The present chapter is organized as follows. In Section 4.2, we first present some optimality conditions for (4.1.1) and a general framework of a primal-dual interior point method. Secondly, we propose a two-step primal-dual interior point method that uses two Newton equations having the same coefficient matrices. In Section 4.3, we prove the superlinear convergence of the proposed method. In Section 4.4, we report some numerical experiments for the proposed method. Finally, we provide some concluding remarks in Section 4.5.

4.2 Two-step primal-dual interior point method

In this section, we first present a primal-dual interior point method for finding the KKT point which satisfies the following KKT conditions:

$$\begin{bmatrix} \nabla_x L(w) \\ g(x) \\ \mathrm{svec}(X(x) \circ Z) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad X(x) \succeq 0, \quad Z \succeq 0,$$

where $w := (x, y, \mathrm{svec}(Z)) \in \mathbf{R}^l$, $l := n + m + \frac{p(p+1)}{2}$ and L is the Lagrangian function given by $L(w) = f(x) - g(x)^\top y - \langle X(x), Z \rangle$. Then, we exploit the following generalized shifted barrier KKT conditions introduced in Section 2.4 in order to construct a primal-dual interior point method:

$$r_\kappa(w, \mu) = \begin{bmatrix} \nabla_x L(w) \\ g(x) + \kappa\mu y \\ \mathrm{svec}(X(x) \circ Z - \mu I) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad X(x) \succeq 0, \quad Z \succeq 0, \tag{4.2.1}$$

where $\kappa \in [0, \infty)$ and $\mu \geq 0$. Note that if $\mu > 0$, the conditions (4.2.1) are equivalent to $r_\kappa(w, \mu) = 0$ and $w \in \mathcal{W}$. Note also that $\mathcal{W} = \{\ w \mid X(x) \succ 0,\ Z \succ 0\ \}$. By using (4.2.1), we present a framework of a primal-dual interior point method based on the generalized shifted barrier KKT conditions.

Algorithm 4.2.1.

Step 0. *(Initialize) Choose parameters $\kappa \geq 0$ and $\epsilon \in (0, 1)$, and give a sequence $\{\mu_k\}$ such that $\lim_{k\to\infty} \mu_k = 0$ and $\mu_k > 0$. Set $k := 0$.*

Step 1. *(Termination) If $\|r_\kappa(w_k, 0)\| \leq \epsilon$, then stop.*

Step 2. *(Newton step) Find an approximate generalized shifted barrier KKT point w_{k+1} such that $\|r_\kappa(w_{k+1}, \mu_k)\| \leq \mu_k$ and $w_{k+1} \in \mathcal{W}$.*

Step 3. *(Update) Set $k := k + 1$, and go to* Step 1.

The global convergence in the case where $\kappa = 0$ or $\kappa = 1$ has already shown in the previous chapter and [34, 72]. Since the global convergence for any $\kappa \in [0, \infty)$ can be also shown similarly, we omit its proof.

In this chapter, we investigate the rate of local convergence. In the following, we propose a two-step primal-dual interior point method that can find w_{k+1} in Step 2. We also show that the proposed method can find w_{k+1} in a single iteration if w_k is sufficiently close to the KKT point. To this end, we first develop a Newton equation with scaling in Subsection 4.2.1. We then provide an actual algorithm in Subsection 4.2.2.

4.2.1 Newton equation with scaling

We adopt a Newton method to find an approximate generalized shifted barrier KKT point w_{k+1} in Step 2 of Algorithm 4.2.1. As mentioned in Section 2.4, we exploit a nonsingular scaling

matrix T, that is,

$$\widetilde{X}(x) = TX(x)T^\top = (T \odot T)X(x), \qquad \widetilde{Z} = T^{-\top}ZT^{-1} = (T^{-\top} \odot T^{-\top})Z.$$

In the following, $X(x)$ and $\widetilde{X}(x)$ are denoted by X and $\widetilde{X}$, respectively, for simplicity.

Then, we consider the following scaled generalized shifted barrier KKT conditions described in Section 2.4:

$$\widetilde{r}_\kappa(w,\mu) = \begin{bmatrix} \nabla_x L(w) \\ g(x) + \kappa\mu y \\ \mathrm{svec}(\widetilde{X} \circ \widetilde{Z} - \mu I) \end{bmatrix}, \quad \widetilde{X} \succ 0, \quad \widetilde{Z} \succ 0.$$

Next, we apply the Newton method to the nonlinear equations $\widetilde{r}_\kappa(w,\mu) = 0$. Then, the Newton equations are given by

$$\begin{aligned} \nabla^2_{xx} L(w)\Delta x - J_g(x)^\top \Delta y - A(x)^\top \mathrm{svec}(\Delta Z) &= -\nabla_x L(w), && (4.2.2)\\ J_g(x)\Delta x + \kappa\mu\Delta y &= -g(x) - \kappa\mu y, && (4.2.3)\\ (\widetilde{Z} \otimes_S I)(T \otimes_S T)A(x)\Delta x + (\widetilde{X} \otimes_S I)(T^{-\top} \otimes_S T^{-\top})\mathrm{svec}(\Delta Z) &= \mathrm{svec}(\mu I - \widetilde{X} \circ \widetilde{Z}). && (4.2.4) \end{aligned}$$

Yamashita and Yabe [71] proposed the following two-step primal-dual interior point method based on the Newton equations (4.2.2)–(4.2.4) in the case where $\kappa = 0$.

Algorithm 4.2.2. *[71, scaled SDPIP]*

Step 0. *(Initialize) Choose parameters $\epsilon > 0$ and $\tau \in (0, \frac{1}{3})$, and give an initial interior point $w_0 = [x_0, y_0, \mathrm{svec}(Z_0)] \in \mathcal{W}$. Set $k := 0$.*

Step 1. *(Termination) If $\|r_0(w_k, 0)\| \le \epsilon$, then stop.*

Step 2. *(Newton steps)*

Step 2.1 *Set $\mu_k := \|r_0(w_k, 0)\|^{1+\tau}$.*

Step 2.2 *Calculate the Newton direction Δw_k by solving the Newton equations (4.2.2)–(4.2.4) at w_k, and set $\hat{w}_k := w_k + \Delta w_k$.*

Step 2.3 *Calculate the Newton direction $\Delta \hat{w}_k$ by solving the Newton equations (4.2.2)–(4.2.4) at $\hat{w}_k$, and set $w_{k+1} := \hat{w}_k + \Delta \hat{w}_k$.*

Step 3. *(Update) Set $k := k + 1$, and go to* Step 1.

Yamashita and Yabe [71] showed the superlinear convergence of Algorithm 4.2.2 under some appropriate assumptions (see Assumption 4.3.1 of Section 4.3). Note that Step 2 in this method has to solve two linear equations with different coefficient matrices.

Yamashita and Yabe [71] also showed that if T is a special matrix such as $T = X^{-\frac{1}{2}}$ and $T = W^{-\frac{1}{2}}$ $(W = X^{\frac{1}{2}}(X^{\frac{1}{2}}ZX^{\frac{1}{2}})^{-\frac{1}{2}}X^{\frac{1}{2}})$, the Newton equation (4.2.4) is written as

$$((Z \otimes_S I)A(x) + P(w))\Delta x + (X \otimes_S I)\mathrm{svec}(\Delta Z) = \mathrm{svec}(\mu I - X \circ Z), \qquad (4.2.5)$$

where the matrix $P(w) \in \mathbf{R}^{\frac{p(p+1)}{2} \times n}$ depends on T. For further details, see [71] or Appendix A. Note that for the general matrix T, there is no matrix $P(w)$ that satisfies (4.2.5). Thus, we make the following assumption on T in the rest of this chapter.

Assumption 4.2.1. *The scaling matrix T satisfies the following (S1):*

(S1) *There exists a matrix $P(w) \in \mathbf{R}^{\frac{p(p+1)}{2} \times n}$ such that the equation (4.2.4) is equivalent to the equation (4.2.5).*

See Appendix A for scaling matrices that satisfy assumption (S1).

4.2.2 Two-step primal-dual interior point method with the same coefficient matrix

We now propose a new algorithm. The proposed algorithm has a similar procedure to Algorithm 4.2.2, i.e., there exist two Newton steps in a single iteration.

First, we calculate $\hat{w}_k := w_k + \Delta w_k$ by solving the Newton equations (4.2.2)–(4.2.4) at w_k as Step 2.2 of Algorithm 4.2.2. From Assumption 4.2.1, the Newton equations are written as

$$\begin{bmatrix} \nabla^2_{xx} L(w_k) & -J_g(x_k)^\top & -A(x_k)^\top \\ J_g(x_k) & \kappa\mu_k I & 0 \\ (Z_k \otimes_S I)A(x_k) + P(w_k) & 0 & (X_k \otimes_S I) \end{bmatrix} \begin{bmatrix} \Delta x_k \\ \Delta y_k \\ \mathrm{svec}(\Delta Z_k) \end{bmatrix} = \begin{bmatrix} -\nabla_x L(w_k) \\ -g(x_k) - \kappa\mu_k y_k \\ \mathrm{svec}(\mu_k I - X_k \circ Z_k) \end{bmatrix}, \quad (4.2.6)$$

where we define $X_k := X(x_k)$ for simplicity.

Recall that the next step of Algorithm 4.2.2, i.e., Step 2.3, solves the Newton equations (4.2.2)–(4.2.4) at $\hat{w}_k$ in order to obtain $\Delta\hat{w}_k$. The coefficient matrix of these equations differs from the coefficient matrix of (4.2.6). Thus, computational costs for Step 2.3 are almost the same as those for Step 2.2.

To reduce computational costs of the second step, we generate a direction $\Delta\hat{w}_k$ by solving the following equation, which has the same coefficient matrix as that in (4.2.6).

$$\begin{bmatrix} \nabla^2_{xx} L(w_k) & -J_g(x_k)^\top & -A(x_k)^\top \\ J_g(x_k) & \kappa\mu_k I & 0 \\ (Z_k \otimes_S I)A(x_k) + P(w_k) & 0 & (X_k \otimes_S I) \end{bmatrix} \begin{bmatrix} \Delta\hat{x}_k \\ \Delta\hat{y}_k \\ \mathrm{svec}(\Delta\hat{Z}_k) \end{bmatrix} = \begin{bmatrix} -\nabla_x L(\hat{w}_k) \\ -g(\hat{x}_k) - \kappa\mu_k \hat{y}_k \\ \mathrm{svec}(\mu_k I - \hat{X}_k \circ \hat{Z}_k) \end{bmatrix}. \quad (4.2.7)$$

Note that $\hat{w}_k$ appears only in the right-hand side of (4.2.7). Summing up the above ideas, we give a new two-step primal-dual interior point method.

Algorithm 4.2.3.

Step 0. *(Initialize) Choose parameters $\kappa \geq 0$, $\epsilon > 0$ and $\tau \in (0, \frac{1}{2})$, and give an initial interior point $w_0 = [x_0, y_0, \mathrm{svec}(Z_0)] \in \mathcal{W}$. Set $k := 0$.*

Step 1. *(Termination) If $\|r_\kappa(w_k, 0)\| \leq \epsilon$, then stop.*

Step 2. *(Newton steps)*

Step 2.1 *Set* $\mu_k := \|r_\kappa(w_k, 0)\|^{1+\tau}$.

Step 2.2 *Calculate the Newton direction* Δw_k *by solving the Newton equation (4.2.6), and set* $\hat{w}_k := w_k + \Delta w_k$.

Step 2.3 *Calculate the Newton direction* $\Delta \hat{w}_k$ *by solving the Newton equation (4.2.7), and set* $w_{k+1} := \hat{w}_k + \Delta \hat{w}_k$.

Step 3. *(Update) Set* $k := k + 1$, *and go to* Step 1.

In the following, we discuss the computational costs of Step 2, i.e., the calculations of Δw_k and $\Delta \hat{w}_k$. First, note that the equation (4.2.6) can be reduced to

$$\begin{aligned}
&\begin{bmatrix} \nabla^2_{xx} L(w_k) + H_k & -J_g(x_k)^\top \\ J_g(x_k) & \kappa\mu_k I \end{bmatrix} \begin{bmatrix} \Delta x_k \\ \Delta y_k \end{bmatrix} = \begin{bmatrix} -\nabla f(x_k) + J_g(x_k)^\top y_k + \mu_k \mathcal{A}^*(x_k) X_k^{-1} \\ -g(x_k) - \kappa\mu_k y_k \end{bmatrix}, \qquad (4.2.8)\\
&\Delta Z_k = \mu_k X_k^{-1} - Z_k - (T_k^\top \odot T_k^\top)(\widetilde{X}_k \odot I)^{-1}(\widetilde{Z}_k \odot I)(T_k \odot T_k)\mathcal{A}(x_k)\Delta x_k,
\end{aligned}$$

where the (i, j)-th element of $H_k \in \mathbf{R}^{n\times n}$ is given by

$$(H_k)_{ij} = \langle A_i(x_k), (T_k^\top \odot T_k^\top)(\widetilde{X}_k \odot I)^{-1}(\widetilde{Z}_k \odot I)(T_k \odot T_k)A_j(x_k)\rangle,$$

and T_k is the scaling matrix at the k-th iteration. Similarly, we can rewrite (4.2.7) as

$$\begin{aligned}
&\begin{bmatrix} \nabla^2_{xx} L(w_k) + H_k & -J_g(x_k)^\top \\ J_g(x_k) & \kappa\mu_k I \end{bmatrix} \begin{bmatrix} \Delta \hat{x}_k \\ \Delta \hat{y}_k \end{bmatrix} \\
&\qquad = \begin{bmatrix} \mathcal{A}^*(x_k)(\mu_k X_k^{-1} - (X_k \odot I)^{-1}(\hat{X}_k \circ \hat{Z}_k)) - \nabla_x L(\hat{w}_k) \\ -g(\hat{x}_k) - \kappa\mu_k \hat{y}_k \end{bmatrix}, \qquad (4.2.9)\\
&\Delta \hat{Z}_k = \mu_k X_k^{-1} - (T_k^\top \odot T_k^\top)(\widetilde{X}_k \odot I)^{-1}(\widetilde{Z}_k \odot I)(T_k \odot T_k)\mathcal{A}(x_k)\Delta \hat{x}_k - (X_k \odot I)^{-1}(\hat{X}_k \circ \hat{Z}_k).
\end{aligned}$$

From these equations, we see that the main calculations of Step 2 are a construction of the matrix H in (4.2.8) and (4.2.9). In Algorithm 4.2.2, it is necessary to calculate the matrix H twice during Steps 2.2 and 2.3. By contrast, in Algorithm 4.2.3, we use the same matrix H in Steps 2.2 and 2.3. Thus, we can expect that Algorithm 4.2.3 can find the next point w_{k+1} faster than Algorithm 4.2.2.

4.3 Local and superlinear convergence of Algorithm 4.2.3

In this section, we show the local and superlinear convergence of Algorithm 4.2.3. First, we give some assumptions for the convergence and we define some neighborhoods of the generalized shifted barrier KKT point. Next, under these assumptions, we show that a sequence generated by Algorithm 4.2.3 is included in the neighborhoods of the generalized shifted barrier KKT point. Finally, we show the superlinear convergence of Algorithm 4.2.3.

4.3.1 Assumptions and some resulting properties

In this subsection, we first give assumptions required for the proof of the superlinear convergence. To this end, let $M(w,\mu) \in \mathbf{R}^{l\times l}$ be a Jacobian of the linear equations (4.2.2)–(4.2.4) with $T = I$, where we define $l := n + m + \frac{p(p+1)}{2}$. Then, the Jacobian $M(w,\mu)$ is expressed as

$$M(w,\mu) := M_0(w) + \kappa\mu M_I, \tag{4.3.1}$$

where

$$M_0(w) := \begin{bmatrix} \nabla^2_{xx}L(w) & -J_g(x)^\top & -A(x)^\top \\ J_g(x) & 0 & 0 \\ (Z \otimes_S I)A(x) & 0 & (X \otimes_S I) \end{bmatrix}, \quad M_I := \begin{bmatrix} 0 & 0 & 0 \\ 0 & I_m & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

We will show the superlinear convergence of Algorithm 4.2.3 under the following assumptions, which are the same as [71].

Assumption 4.3.1. *Let $w^* = [x^*, y^*, \mathrm{svec}(Z^*)]$ be a KKT point of nonlinear SDP (4.1.1).*

(A1) *There exists a positive constant ν_L such that M_0 is Lipschitz continuous on $\mathcal{V}_L := \{w \in \mathbf{R}^l \mid \|w - w^*\| \le \nu_L\}$.*

(A2) *The second-order sufficient condition holds at x^*.*

(A3) *The strict complementarity condition holds at x^*.*

(A4) *The nondegeneracy condition holds at x^*.*

Note that it follows from Theorem 2.3.3 that if (A3) holds, then (A4) holds if and only if the Lagrange multipliers $y^* \in \mathbf{R}^m$ and $Z^* \in \mathbf{S}^p$ corresponding to $x^* \in \mathbf{R}^n$ are unique.

Assumption 4.3.1 (A1) implies that there exists a positive constant L_M such that

$$\|M_0(w_1) - M_0(w_2)\|_F \le L_M\|w_1 - w_2\| \quad \text{for all } w_1, w_2 \in \mathcal{V}_L. \tag{4.3.2}$$

It follows from the definition of M_0 that

$$\|(Z_1 \otimes_S I)A(x_1) - (Z_2 \otimes_S I)A(x_2)\|_F \le \|M_0(w_1) - M_0(w_2)\|_F \quad \text{for all } w_1, w_2 \in \mathcal{V}_L, \tag{4.3.3}$$

$$\|X(x_1) \otimes_S I - X(x_2) \otimes_S I\|_F \le \|M_0(w_1) - M_0(w_2)\|_F \quad \text{for all } w_1, w_2 \in \mathcal{V}_L. \tag{4.3.4}$$

Moreover, we have from (4.3.2) and Proposition 2.2.2 (a) that

$$\|r_\kappa(w_1, 0) - r_\kappa(w_2, 0) - M_0(w_2)(w_1 - w_2)\| \le L_M\|w_1 - w_2\|^2 \quad \text{for all } w_1, w_2 \in \mathcal{V}_L.$$

Since $r_\kappa(w_1, 0) - r_\kappa(w_2, 0) - M_0(w_2)(w_1 - w_2) = r_\kappa(w_1, \mu) - r_\kappa(w_2, \mu) - M(w_2, \mu)(w_1 - w_2)$ for all $w_1, w_2 \in \mathcal{V}_L$ and $\mu \ge 0$, it then follows that

$$\|r_\kappa(w_1, \mu) - r_\kappa(w_2, \mu) - M(w_2, \mu)(w_1 - w_2)\| \le L_M\|w_1 - w_2\|^2 \tag{4.3.5}$$

for all $w_1, w_2 \in \mathcal{V}_L$ and $\mu \ge 0$. Furthermore, by the definition of M, we obtain

$$\|\mathrm{svec}(X(x_1) \circ Z_1 - X(x_2) \circ Z_2) - (Z_2 \otimes_S I)A(x_2)(x_1 - x_2) - (X(x_2) \otimes_S I)\mathrm{svec}(Z_1 - Z_2)\| \le L_M\|w_1 - w_2\|^2 \tag{4.3.6}$$

for all $w_1, w_2 \in \mathcal{V}_L$.

Yamashita and Yabe [71] showed that $M_0(w^*)$ is nonsingular under Assumption 4.3.1 (A2)–(A4).

Theorem 4.3.1. *[71, Theorem 1] Suppose that Assumption 4.3.1 (A2)–(A4) hold. Then, the matrix $M_0(w^*)$ is invertible.* □

Then, Theorems 2.2.2 and 4.3.1 guarantee that there exist a positive constant ζ and a unique continuously differentiable function $\overline{w} : (-\zeta, \zeta) \to \mathbf{R}^l$ such that $r_\kappa(\overline{w}(\mu), \mu) = 0$. Furthermore, the following lemma holds.

Lemma 4.3.1. *[71, Lemma 1] Suppose that Assumption 4.3.1 (A1)–(A4) hold. Then, there exist a positive constant γ and a unique continuously differentiable function $\bar{w} : [0, \gamma] \to \mathbf{R}^l$ such that*

$$\bar{w}(0) = w^*, \ \overline{w}(\mu) := [\overline{x}(\mu), \overline{y}(\mu), \mathrm{svec}(\overline{Z}(\mu))], \ r_\kappa(\bar{w}(\mu), \mu) = 0 \quad \text{for any } \mu \in [0, \gamma].$$

Furthermore, $X(\bar{x}(\mu)) \succ 0$ and $\bar{Z}(\mu) \succ 0$ for any $\mu \in (0, \gamma]$. □

We call $\{\overline{w}(\mu) | \mu \in [0, \gamma]\}$ the central path of (4.1.1).

Since $M_0(w^*)$ is invertible, there exists $\varepsilon \in (0, 1)$ such that any matrix $G \in \mathbf{R}^{l \times l}$ that satisfies

$$\|G - M_0(w^*)\|_F < \varepsilon \tag{4.3.7}$$

is nonsingular. From the continuity of M_0 at w^*, there exists a positive constant ν_M such that

$$\|M_0(w) - M_0(w^*)\|_F \leq \frac{1}{4}\varepsilon \quad \text{for any } w \text{ such that } \|w - w^*\| \leq \nu_M. \tag{4.3.8}$$

Thus, it follows from (4.3.7) that $M_0(w)$ is nonsingular if $\|w - w^*\| \leq \nu_M$.

Let $\nu := \min\{\nu_M, \nu_L\}$. Then, we define a subset of $\mathcal{V}_L$.

$$\mathcal{V} := \{ \ w \in \mathbf{R}^l \mid \|w - w^*\| \leq \nu \ \} \subset \mathcal{V}_L.$$

Note that M_0 is Lipschitz continuous on $\mathcal{V}$.

Next, we give a condition on μ under which $M(w, \mu)$ is invertible for any $w \in \mathcal{V}$. Now, let $w \in \mathcal{V}$. By the definition of M and the triangle inequality, we have $\|M(w, \mu) - M_0(w^*)\|_F = \|M_0(w) + \kappa\mu M_I - M_0(w^*)\|_F \leq \|M_0(w) - M_0(w^*)\|_F + \|\kappa\mu M_I\|_F$. It then follows from (4.3.8) and $\|M_I\|_F = \|I_m\|_F = \sqrt{m}$ that $\|M_0(w) - M_0(w^*)\|_F + \|\kappa\mu M_I\|_F \leq \frac{1}{4}\varepsilon + \kappa\mu\sqrt{m}$. If $\mu \leq s := \frac{\varepsilon}{4(\kappa+1)\sqrt{m}}$, then we have

$$\|M(w, \mu) - M_0(w^*)\|_F \leq \frac{1}{4}\varepsilon + \kappa\mu\sqrt{m} = \left(\frac{1}{4} + \frac{1}{4}\right)\varepsilon = \frac{1}{2}\varepsilon \quad \text{for all } w \in \mathcal{V}. \tag{4.3.9}$$

Thus, it follows from (4.3.7) that $M(w, \mu)$ is invertible for all $w \in \mathcal{V}$ and $\mu \in [0, s]$. Moreover, we may define

$$U_M := \sup\{ \ \|M(w, \mu)^{-1}\|_F \mid w \in \mathcal{V}, \ \mu \in [0, s] \ \} \quad U_y := \sup\{ \ \|y\|^2 \mid w \in \mathcal{V} \ \}.$$

Note that $U_M < \infty$ from (4.3.7) and (4.3.9). Note also that $U_y < \infty$ from the boundedness of $\mathcal{V}$. Since

$$\|r_\kappa(w,0)\| = \left\| r_\kappa(w,\mu) - \mu \begin{bmatrix} 0 \\ \kappa y \\ \mathrm{svec}(-I) \end{bmatrix} \right\|, \quad \|r_\kappa(w,\mu)\| = \left\| r_\kappa(w,0) + \mu \begin{bmatrix} 0 \\ \kappa y \\ \mathrm{svec}(-I) \end{bmatrix} \right\|,$$

we have

$$\mu\sqrt{p} - \|r_\kappa(w,\mu)\| \le \|r_\kappa(w,0)\| \le \|r_\kappa(w,\mu)\| + \mu U_1 \quad \text{for all } w \in \mathcal{V}, \mu \in [0,s], \tag{4.3.10}$$

$$\mu\sqrt{p} - \|r_\kappa(w,0)\| \le \|r_\kappa(w,\mu)\| \le \|r_\kappa(w,0)\| + \mu U_1 \quad \text{for all } w \in \mathcal{V}, \mu \in [0,s], \tag{4.3.11}$$

where $U_1 := \sqrt{\kappa U_y + p}$.

The differentiability of r_κ and X, and the boundedness of $\mathcal{V}_L$ and $[0,s]$ imply that there exist positive constants L_r and L_{XZ} such that

$$\|r_\kappa(w_1,\mu) - r_\kappa(w_2,\mu)\| \le L_r\|w_1 - w_2\| \quad \text{for all } w_1, w_2 \in \mathcal{V}_L, \mu \in [0,s], \tag{4.3.12}$$

$$\|X(x_1)Z_1 - X(x_2)Z_2\|_F \le L_{XZ}\|w_1 - w_2\| \quad \text{for all } w_1, w_2 \in \mathcal{V}_L, \mu \in [0,s]. \tag{4.3.13}$$

Next, we define a neighborhood of the central path. Let

$$\nu_N := \min\left\{ \nu,\ \frac{3}{8L_M U_M},\ \left[\frac{1}{5L_r^{1+\tau} U_M(1+U_1)}\right]^{\frac{1}{\tau}} \right\}. \tag{4.3.14}$$

Then, we define a subset of $\mathcal{V}$.

$$\mathcal{V}_N := \{\ w \in \mathbf{R}^l \mid \|w - w^*\| \le \nu_N\ \} \subset \mathcal{V}. \tag{4.3.15}$$

Note that $\tau \in (0, \frac{1}{2})$ is the constant given in Algorithm 4.2.3. Secondly, we define two subsets of $\mathcal{V}_N$.

$$\mathcal{N}_1(\mu) := \{\ w \in \mathcal{V}_N \mid \|r_\kappa(w,\mu)\| \le \mu^{1+\sigma},\ X(x) \succeq 0,\ Z \succeq 0\ \},$$
$$\mathcal{N}_2(\mu) := \{\ w \in \mathcal{V}_N \mid \|r_\kappa(w,\mu)\| \le \mu^{1+\rho},\ X(x) \succeq 0,\ Z \succeq 0\ \},$$

where σ and ρ are positive constants such that

$$\max\left\{\frac{\tau}{1-\tau}, \frac{1}{2}\right\} < \rho < 1, \quad \sigma < \frac{\rho - \tau}{1+\tau}. \tag{4.3.16}$$

Since $0 < \sigma < \frac{\rho-\tau}{1+\tau} < \frac{\rho}{1+\tau} < \rho$, we have $\mathcal{N}_2(\mu) \subset \mathcal{N}_1(\mu)$ for $\mu \in [0,1]$.

Lemma 4.3.1 shows that the generalized shifted barrier KKT point $\overline{w}(\mu)$ is unique for $\mu \in [0,\gamma]$. Then, we may regard $\mathcal{N}_1(\mu)$ and $\mathcal{N}_2(\mu)$ as neighborhoods of the generalized shifted barrier KKT point $\overline{w}(\mu)$. Thus, we define the following neighborhoods of the central path by using $\mathcal{N}_1(\mu)$ and $\mathcal{N}_2(\mu)$.

$$\Theta_1(\theta) := \cup_{\mu\in[0,\theta]}\mathcal{N}_1(\mu), \quad \Theta_2(\theta) := \cup_{\mu\in[0,\theta]}\mathcal{N}_2(\mu) \quad \text{for any } \theta \in [0, \min\{\gamma, s\}].$$

Note that since $0 < s < 1$, we have $0 \le \theta < 1$. Then,

$$\Theta_2(\theta) \subset \Theta_1(\theta) \subset \mathcal{V}_N \quad \text{for all } \theta \in [0, \min\{\gamma, s\}]. \tag{4.3.17}$$

We can consider $\Theta_1(\theta)$ and $\Theta_2(\theta)$ as neighborhoods of the central path. Moreover, we define

$$U_w(\theta) := \sup_{w\in\Theta_1(\theta),\mu\in[0,\theta]} \|w - \bar{w}(\mu)\| \quad \text{for any } \theta \in [0, \min\{\gamma, s\}],$$

which expresses the supremum of a distance between a point in $\Theta_1(\theta)$ and the central path. Now, we briefly show that there exists $\theta_1 > 0$ such that $1 - L_M U_M U_w(\theta) \geq \frac{1}{4}$ for all $\theta \in [0, \theta_1]$ and $\overline{w}(\mu) \in \mathcal{V}_N$ for all $\mu \in [0, \theta_1]$. Since $\overline{w}$ is continuous on $[0, \gamma]$ by Lemma 4.3.1, there exists $\theta_0 > 0$ such that

$$\|\overline{w}(\mu) - \overline{w}(0)\| \leq \nu_N \quad \text{for all } \mu \in [0, \theta_0]. \tag{4.3.18}$$

Using $\overline{w}(0) = w^*$, (4.3.17) and (4.3.18),

$$\begin{aligned} U_w(\theta) &= \sup_{w\in\Theta_1(\theta),\mu\in[0,\theta]} \|w - w^* + w^* - \bar{w}(\mu)\| \\ &\leq \sup_{w\in\Theta_1(\theta)} \|w - w^*\| + \sup_{\mu\in[0,\theta]} \|\bar{w}(\mu) - \overline{w}(0)\| \\ &\leq 2\nu_N \end{aligned} \tag{4.3.19}$$

for all $\theta \in [0, \min\{\gamma, s, \theta_0\}]$. Then (4.3.14) and (4.3.19) imply that $L_M U_M U_w(\theta) \leq \frac{3}{4}$ for all $\theta \in [0, \min\{\gamma, s, \theta_0\}]$. Thus,

$$1 - L_M U_M U_w(\theta) \geq \frac{1}{4} \quad \text{for all } \theta \in [0, \min\{\gamma, s, \theta_0\}]. \tag{4.3.20}$$

Moreover, from (4.3.18) and $\overline{w}(0) = w^*$,

$$\overline{w}(\mu) \in \mathcal{V}_N \quad \text{for all } \mu \in [0, \min\{\gamma, s, \theta_0\}]. \tag{4.3.21}$$

Hence, letting $\theta_1 := \min\{\gamma, s, \theta_0\}$, we have the desired results. Then, we give a condition under which $r_\kappa(w, \mu)$ provides an error bound of the generalized shifted barrier KKT point.

Lemma 4.3.2. *Suppose that Assumption 4.3.1 holds, and that $\theta \in [0, \theta_1]$. Then,*

$$\|w - \bar{w}(\mu)\| \leq U_r \|r_\kappa(w, \mu)\|, \qquad \|XZ - \mu I\|_F \leq U_R \|r_\kappa(w, \mu)\|$$

for all $w \in \Theta_1(\theta)$ and $\mu \in [0, \theta]$, where $U_r := 4U_M$ and $U_R := 4L_{XZ}U_M$.

Proof. Let $w \in \Theta_1(\theta)$ and $\mu \in [0, \theta]$. From (4.3.21), $\overline{w}(\mu) \in \mathcal{V}_N$ for all $\mu \in [0, \theta]$. Note that $\Theta_1(\theta) \subset \mathcal{V}_N \subset \mathcal{V} \subset \mathcal{V}_L$. Substituting $w_1 = w \in \mathcal{V}_L$ and $w_2 = \overline{w}(\mu) \in \mathcal{V}_L$ into (4.3.5),

$$L_M \|w - \bar{w}(\mu)\|^2 \geq \|M(\bar{w}(\mu), \mu)(w - \bar{w}(\mu)) - r_\kappa(w, \mu)\| \geq \|M(\bar{w}(\mu), \mu)(w - \bar{w}(\mu))\| - \|r_\kappa(w, \mu)\|$$

from $r_\kappa(\overline{w}(\mu), \mu) = 0$. Since $\|M(\overline{w}(\mu), \mu)(w - \overline{w}(\mu))\| \geq \frac{\|w - \overline{w}(\mu)\|}{\|M(\overline{w}(\mu), \mu)^{-1}\|_F}$, it then follows from $U_w(\theta) \geq \|w - \overline{w}(\mu)\|$ and $U_M \geq \|M(\overline{w}(\mu), \mu)^{-1}\|_F$ that

$$L_M U_w(\theta) \|w - \bar{w}(\mu)\| \geq \frac{\|w - \bar{w}(\mu)\|}{\|M(\overline{w}(\mu), \mu)^{-1}\|_F} - \|r_\kappa(w, \mu)\| \geq \frac{\|w - \bar{w}(\mu)\|}{U_M} - \|r_\kappa(w, \mu)\|.$$

As the result, we have

$$\frac{1 - L_M U_M U_w(\theta)}{U_M} \|w - \bar{w}(\mu)\| \leq \|r_\kappa(w, \mu)\|.$$

Then, since $1 - L_M U_M U_w(\theta) \geq \frac{1}{4}$ from $0 < \theta \leq \theta_1 = \min\{\gamma, s, \theta_0\}$ and (4.3.20), we obtain

$$\|w - \bar{w}(\mu)\| \leq 4U_M \|r_\kappa(w, \mu)\|. \tag{4.3.22}$$

By $U_r = 4U_M$, we have the first inequality.

Next, we show the second inequality. We have $X(\overline{x}(\mu)) \circ \overline{Z}(\mu) = \mu I$ by $r_\kappa(\overline{w}(\mu), \mu) = 0$. Since $X(\overline{x}(\mu)) \succeq 0$ and $\overline{Z}(\mu) \succeq 0$, it follows from Proposition 2.2.8 (b) that $X(\overline{x}(\mu)) \circ \overline{Z}(\mu) = \mu I$ is equivalent to $X(\overline{x}(\mu))\overline{Z}(\mu) = \mu I$. Then, (4.3.13) yields that

$$L_{XZ}\|w - \overline{w}(\mu)\| \geq \left\|\mathrm{svec}\left[XZ - X(\overline{x}(\mu))\overline{Z}(\mu)\right]\right\| = \|XZ - \mu I\|_F.$$

Combining this inequality and (4.3.22), we have $\|XZ - \mu I\|_F \leq 4L_{XZ}U_M\|r_\kappa(w, \mu)\|$. Since $U_R = 4L_{XZ}U_M$, we obtain the desired inequality. □

From Lemma 4.3.2, we can show that w^* is an isolated KKT point.

Theorem 4.3.2. *Suppose that Assumption 4.3.1 holds. If $\widetilde{w} \in \mathcal{N}_2(0)$, then $\widetilde{w} = w^*$.*

Proof. Note that $\widetilde{w} \in \mathcal{N}_2(0) = \mathcal{N}_1(0) = \Theta_1(0)$. It then follows from the definition of $\mathcal{N}_2(0)$ that $r_\kappa(\widetilde{w}, 0) = 0$. Furthermore, we have from Lemma 4.3.2 that $\|\widetilde{w} - w^*\| = \|\widetilde{w} - \overline{w}(0)\| \leq U_r\|r_\kappa(\widetilde{w}, 0)\| = 0$, that is, $\widetilde{w} = w^*$. □

4.3.2 Proof of superlinear convergence

We show the superlinear convergence of Algorithm 4.2.3 by using the properties given in Subsection 4.3.1.

First, we give an assumption related to the matrix $P(w)$, which is included in (4.2.6) and (4.2.7). To this end, we define $\theta_2 := \min\{\theta_1, (\frac{3}{4U_R})^{\frac{1}{\rho}}\}$ and

$$\Gamma(\theta) := \{\ (w, \eta) \in \mathbf{R}^l \times \mathbf{R} \mid w \in \mathcal{N}_2(\eta) \subset \Theta_2(\theta),\ w \in \mathcal{W},\ \eta \in (0, \theta]\ \} \quad \text{for } \theta \in (0, \theta_2].$$

Then, we make the following assumption on the matrix $P(w)$.

Assumption 4.3.2. *The scaling matrix T satisfies Assumption 4.2.1 (S1), that is, there exists $P(w)$ such that (4.2.4) is equivalent to (4.2.5). The matrix $P(w)$ satisfies the following (S2):*

(S2) *If $\theta \in (0, \theta_2]$, then there exists $U_P > 0$ such that $\|P(w)\|_F \leq U_P\eta^\rho$ for any $(w, \eta) \in \Gamma(\theta)$.*

When $T_k = I$, Assumption 4.3.2 (S2) holds since $P(w_k) := 0$. Furthermore, when $T_k = X_k^{-\frac{1}{2}}$ or $T_k = W_k^{-\frac{1}{2}}$ ($W_k = X_k^{\frac{1}{2}}(X_k^{\frac{1}{2}} Z_k X_k^{\frac{1}{2}})^{-\frac{1}{2}} X_k^{\frac{1}{2}}$), which are well-known scaling matrices of linear SDP, there exists the matrix $P(w_k)$ such that Assumption 4.3.2 (S2) holds. These proofs are given in Appendix A.

Assumption 4.2.1 (S1) means that the Newton equations of Steps 2.2 and 2.3 in Algorithm 4.2.3 are reduced to

$$M_P(w_k, \mu_k)\Delta w_k = -r_\kappa(w_k, \mu_k), \qquad M_P(w_k, \mu_k)\Delta \hat{w}_k = -r_\kappa(\hat{w}_k, \mu_k), \tag{4.3.23}$$

respectively, where

$$M_P(w_k,\mu_k) := M(w_k,\mu_k) + N(w_k), \quad N(w_k) := \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ P(w_k) & 0 & 0 \end{bmatrix}. \tag{4.3.24}$$

It also follows that

$$\|M_P(w_k,\mu_k) - M(w_k,\mu_k)\|_F = \|N(w_k)\|_F = \|P(w_k)\|_F. \tag{4.3.25}$$

To establish the superlinear convergence of Algorithm 4.2.3, we first show that $M_P(w_k,\mu_k)$ is nonsingular if $w_k \in \Theta_2(\theta)$ and $w_k \in \mathcal{W}$ (Lemma 4.3.4). Then, we show that $\hat{w}_k \in \mathcal{N}_1(\theta)$ and $\hat{w}_k \in \mathcal{W}$ if $w_k \in \Theta_2(\theta)$ and $w_k \in \mathcal{W}$ (Lemmas 4.3.5 and 4.3.7). Finally, we prove that $w_{k+1} \in \mathcal{N}_2(\theta)$ and $w_{k+1} \in \mathcal{W}$ if $w_k \in \Theta_2(\theta)$ and $w_k \in \mathcal{W}$ (Lemmas 4.3.8 and 4.3.9). From these results, we can easily obtain the desired theorem (Theorem 4.3.3).

In the following two lemmas, we assume that

$$0 < \theta \le \theta_3, \quad \theta_3 := \min\left\{ \theta_2, \ \left[\frac{3}{4(1+U_1)^{1+\tau}}\right]^{\frac{1}{\tau}}, \ \left[\frac{\varepsilon(\sqrt{p}-1)^\rho}{4U_P}\right]^{\frac{1+\tau}{\rho}} \right\}. \tag{4.3.26}$$

Lemma 4.3.3. *Suppose that Assumptions 4.2.1, 4.3.1 and 4.3.2 hold, and that θ satisfies (4.3.26). If $w_k \in \Theta_2(\theta)$ and $w_k \in \mathcal{W}$, then there exists $\eta_k \in (0,\theta]$ such that $(w_k,\eta_k) \in \Gamma(\theta)$ and $\eta_k \le U_\eta \mu_k^{\frac{1}{1+\tau}}$, where $U_\eta := (\sqrt{p}-1)^{-1}$. Moreover, $0 < \mu_k < \theta$.*

Proof. Note that $w_k \neq w^*$ because $w_k \in \mathcal{W}$. Then, Theorem 4.3.2 implies that $w_k \notin \mathcal{N}_2(0)$. Since $w_k \in \Theta_2(\theta)$ and $w_k \notin \mathcal{N}_2(0)$, there exists $\eta_k \in (0,\theta]$ such that $w_k \in \mathcal{N}_2(\eta_k) \subset \Theta_2(\theta)$, i.e., $(w_k,\eta_k) \in \Gamma(\theta)$. It follows from $w_k \in \Theta_2(\theta) \subset \Theta_1(\theta) \subset \mathcal{V}_N \subset \mathcal{V}$, $0 < \eta_k \le \theta \le \theta_2 \le \theta_1 \le s < 1$ and (4.3.11) that $\eta_k\sqrt{p} - \|r_\kappa(w_k,0)\| \le \|r_\kappa(w_k,\eta_k)\| \le \eta_k^{1+\rho} \le \eta_k$. Thus, we have $\eta_k(\sqrt{p}-1) \le \|r_\kappa(w_k,0)\| = \mu_k^{\frac{1}{1+\tau}}$, and hence we obtain $\eta_k \le U_\eta \mu_k^{\frac{1}{1+\tau}}$.

Next, we prove that $0 < \mu_k < \theta$. Since $\eta_k \in (0,\theta]$ and $\eta_k \le U_\eta \mu_k^{\frac{1}{1+\tau}}$, we get $0 < (\frac{\eta_k}{U_\eta})^{1+\tau} \le \mu_k$. Moreover, we have from the first part of this proof that $\|r_\kappa(w_k,\eta_k)\| \le \eta_k^{1+\rho}$. It then follows from (4.3.10) that $\|r_\kappa(w_k,0)\| \le \|r_\kappa(w_k,\eta_k)\| + \eta_k U_1 \le (\eta_k^\rho + U_1)\eta_k \le (1+U_1)\theta$. Since $\theta \le [\frac{3}{4(1+U_1)^{1+\tau}}]^{\frac{1}{\tau}}$ by (4.3.26), we get $\mu_k = \|r_\kappa(w_k,0)\|^{1+\tau} \le (1+U_1)^{1+\tau}\theta^\tau\theta \le \frac{3}{4}\theta$. Therefore, $0 < \mu_k < \theta$. □

We have from Lemma 4.3.3 that if $\theta \in (0,\theta_3]$, $w_k \in \Theta_2(\theta)$ and $w_k \in \mathcal{W}$, then

$$\exists \eta_k \in (0,\theta] \quad \text{such that} \quad (w_k,\eta_k) \in \Gamma(\theta), \quad \eta_k \le U_\eta \mu_k^{\frac{1}{1+\tau}}, \tag{4.3.27}$$

and

$$0 < \mu_k < \theta. \tag{4.3.28}$$

Then, Assumption 4.3.2 (S2) and (4.3.27) imply that

$$\|P(w_k)\|_F \le U_P \eta_k^\rho \le \mathcal{U}_P \mu_k^{\frac{\rho}{1+\tau}}, \tag{4.3.29}$$

where $\mathcal{U}_P := U_\eta^\rho U_P$.

By using (4.3.28) and (4.3.29), we prove that the Jacobian matrix $M_P(w_k,\mu_k)$ of (4.3.23) is nonsingular.

Lemma 4.3.4. *Suppose that Assumptions 4.2.1, 4.3.1 and 4.3.2 hold, and that θ satisfies (4.3.26). If $w_k \in \Theta_2(\theta)$ and $w_k \in \mathcal{W}$, then $M_P(w_k, \mu_k)$ is nonsingular.*

Proof. We have from (4.3.24) that $\|M_P(w_k,\mu_k) - M_0(w^*)\|_F \le \|M(w_k,\mu_k) - M_0(w^*)\|_F + \|N(w_k)\|_F$. Then (4.3.25) yields that

$$\|M_P(w_k,\mu_k) - M_0(w^*)\|_F \le \|M(w_k,\mu_k) - M_0(w^*)\|_F + \|P(w_k)\|_F. \tag{4.3.30}$$

We can easily see that $w_k \in \mathcal{V}$ and $\mu_k \in [0, s]$ from $w_k \in \Theta_2(\theta)$, (4.3.26) and (4.3.28). Thus, (4.3.9) yields that

$$\|M(w_k,\mu_k) - M_0(w^*)\|_F \le \frac{1}{2}\varepsilon. \tag{4.3.31}$$

On the other hand, it follows from (4.3.26), (4.3.28), $U_\eta = (\sqrt{p}-1)^{-1}$ and $\mathcal{U}_P = U_\eta^\rho U_P$ that $\mu_k \le \theta \le \theta_3 \le (\frac{\varepsilon}{4\mathcal{U}_P})^{\frac{1+\tau}{\rho}}$, that is, $\mathcal{U}_P \mu_k^{\frac{\rho}{1+\tau}} \le \frac{1}{4}\varepsilon$. Then, we have from (4.3.29) that

$$\|P(w_k)\|_F \le \frac{1}{4}\varepsilon. \tag{4.3.32}$$

By (4.3.30), (4.3.31) and (4.3.32), $\|M_P(w_k,\mu_k) - M_0(w^*)\|_F \le \frac{3}{4}\varepsilon$. Therefore, $M_P(w_k,\mu_k)$ is nonsingular from (4.3.7). □

We define

$$U_{\mathcal{M}} := \sup\left\{ \|M_P(w,\mu)^{-1}\|_F \;\middle|\; w \in \Theta_2(\theta_3),\ w \in \mathcal{W},\ \mu := \|r_\kappa(w,0)\|^{1+\tau} \right\}.$$

It then follows from Lemma 4.3.4 that if $\theta \in (0, \theta_3]$, $w_k \in \Theta_2(\theta)$ and $w_k \in \mathcal{W}$, then

$$\|M_P(w_k,\mu_k)^{-1}\|_F \le U_{\mathcal{M}} < \infty. \tag{4.3.33}$$

Now, we show that if $w_0 \in \Theta_2(\theta)$ and $w_0 \in \mathcal{W}$ for sufficiently small $\theta > 0$, then $\{w_k\} \subset \Theta_2(\theta)$ and $\{w_k\} \subset \mathcal{W}$. To this end, we first show that $\hat{w}_k$ generated by Step 2.2 of Algorithm 4.2.3 satisfies $\hat{w}_k \in \mathcal{N}_1(\mu_k) \subset \Theta_1(\theta)$ and $\hat{w}_k \in \mathcal{W}$ if $w_k \in \Theta_2(\theta)$ and $w_k \in \mathcal{W}$. In what follows, we assume that

$$0 < \theta \le \theta_4, \quad \theta_4 := \min\left\{ \theta_3,\ \left[\frac{\nu_N}{U_r + U_2}\right]^{1+\tau},\ \left(\frac{1}{U_3}\right)^{\frac{1}{h_1}} \right\}, \tag{4.3.34}$$

where

$$h_1 := \frac{\rho - \tau}{1+\tau} - \sigma, \quad U_2 := U_{\mathcal{M}}(1+U_1), \quad U_3 := U_2(L_M U_2 + \mathcal{U}_P).$$

Note that $h_1 > 0$ from (4.3.16).

In order to prove $\hat{w}_k \in \mathcal{N}_1(\mu_k)$ and $\hat{w}_k \in \mathcal{W}$, we have to show that $\hat{w}_k \in \mathcal{V}_N, \|r_\kappa(\hat{w}_k,\mu_k)\| \le \mu_k^{1+\sigma}, X(\hat{x}_k) \succ 0$ and $\hat{Z}_k \succ 0$. Thus, we first show that $\hat{w}_k \in \mathcal{V}_N$ and $\|r_\kappa(\hat{w}_k,\mu_k)\| \le \mu_k^{1+\sigma}$. Note that $\mu_k = \|r_\kappa(w_k,0)\|^{1+\tau}$ and $\hat{w}_k = [\hat{x}_k, \hat{y}_k, \mathrm{svec}(\hat{Z}_k)]$ are generated by Steps 2.1 and 2.2 of Algorithm 4.2.3, respectively.

Lemma 4.3.5. *Suppose that Assumptions 4.2.1, 4.3.1 and 4.3.2 hold, and that θ satisfies (4.3.34). If $w_k \in \Theta_2(\theta)$ and $w_k \in \mathcal{W}$, then*

$$\|\Delta w_k\| \leq U_2 \mu_k^{\frac{1}{1+\tau}},\ \|\hat{w}_k - w^*\| \leq \nu_N,\ \|r_\kappa(\hat{w}_k, \mu_k)\| \leq U_3 \mu_k^{1+\frac{\rho-\tau}{1+\tau}},\ \|r_\kappa(\hat{w}_k, \mu_k)\| \leq \mu_k^{1+\sigma}.$$

Proof. First, we show that $\|\Delta w_k\| \leq U_2 \mu_k^{\frac{1}{1+\tau}}$ and $\|\hat{w}_k - w^*\| \leq \nu_N$. It is clear that $w_k \in \mathcal{V}$ and $\mu_k \in [0, s]$ by $w_k \in \Theta_2(\theta)$, (4.3.28) and (4.3.34). Thus, it follows from (4.3.11) that

$$\|r_\kappa(w_k, \mu_k)\| \leq \|r_\kappa(w_k, 0)\| + \mu_k U_1 = \mu_k^{\frac{1}{1+\tau}} + \mu_k U_1 \leq (1 + U_1) \mu_k^{\frac{1}{1+\tau}}. \tag{4.3.35}$$

Meanwhile, we have from (4.3.23) and Lemma 4.3.4 that $\Delta w_k = -M_P(w_k, \mu_k)^{-1} r_\kappa(w_k, \mu_k)$. Furthermore, (4.3.33) and (4.3.35) yield that

$$\|\Delta w_k\| \leq \|M_P(w_k, \mu_k)^{-1}\|_F \|r_\kappa(w_k, \mu_k)\| \leq U_2 \mu_k^{\frac{1}{1+\tau}}. \tag{4.3.36}$$

By Lemma 4.3.2 and (4.3.36),

$$\|\hat{w}_k - w^*\| \leq \|w_k - w^*\| + \|\Delta w_k\| \leq U_r \|r_\kappa(w_k, 0)\| + U_2 \mu_k^{\frac{1}{1+\tau}} = (U_r + U_2)\, \mu_k^{\frac{1}{1+\tau}}.$$

Then (4.3.28) and (4.3.34) imply that $\|\hat{w}_k - w^*\| \leq (U_r + U_2)\, \mu_k^{\frac{1}{1+\tau}} \leq (U_r + U_2)\, \theta^{\frac{1}{1+\tau}} \leq \nu_N$.

Secondly, we show that $\|r_\kappa(\hat{w}_k, \mu_k)\| \leq U_3 \mu_k^{1+\frac{\rho-\tau}{1+\tau}}$ and $\|r_\kappa(\hat{w}_k, \mu_k)\| \leq \mu_k^{1+\sigma}$. We easily see that $w_k, \hat{w}_k \in \mathcal{V}_L$. It then follows from (4.3.5) that

$$\begin{aligned} L_M \|\Delta w_k\|^2 &\geq \|r_\kappa(\hat{w}_k, \mu_k) - r_\kappa(w_k, \mu_k) - M(w_k, \mu_k) \Delta w_k\| \\ &\geq \|r_\kappa(\hat{w}_k, \mu_k)\| - \|r_\kappa(w_k, \mu_k) + M(w_k, \mu_k) \Delta w_k\|. \end{aligned}$$

Moreover, (4.3.23), (4.3.25) and (4.3.29) yield that

$$\begin{aligned} L_M \|\Delta w_k\|^2 &\geq \|r_\kappa(\hat{w}_k, \mu_k)\| - \|(M(w_k, \mu_k) - M_P(w_k, \mu_k)) \Delta w_k\| \\ &\geq \|r_\kappa(\hat{w}_k, \mu_k)\| - \|P(w_k)\|_F \|\Delta w_k\| \\ &\geq \|r_\kappa(\hat{w}_k, \mu_k)\| - \mathcal{U}_P \mu_k^{\frac{\rho}{1+\tau}} \|\Delta w_k\|. \end{aligned} \tag{4.3.37}$$

Thus, we get $\|r_\kappa(\hat{w}_k, \mu_k)\| \leq L_M U_2^2 \mu_k^{1+\frac{1-\tau}{1+\tau}} + \mathcal{U}_P U_2 \mu_k^{1+\frac{\rho-\tau}{1+\tau}} \leq U_3 \mu_k^{1+\frac{\rho-\tau}{1+\tau}}$ by using (4.3.36), (4.3.37) and $\mu_k \in (0, 1)$. Using (4.3.28), we have $\|r_\kappa(\hat{w}_k, \mu_k)\| \leq U_3 \mu_k^{h_1} \mu_k^{1+\sigma} \leq U_3 \theta^{h_1} \mu_k^{1+\sigma}$. Note that $h_1 = \frac{\rho-\tau}{1+\tau} - \sigma > 0$ by (4.3.16). Since $U_3 \theta^{h_1} \leq 1$ from (4.3.34), we get $\|r_\kappa(\hat{w}_k, \mu_k)\| \leq \mu_k^{1+\sigma}$. □

Next, we show that $\hat{w}_k \in \mathcal{W}$ if we choose θ such that

$$0 < \theta \leq \theta_5, \quad \theta_5 := \min \left\{ \theta_4,\ \left(\frac{3}{4}\right)^{\frac{1}{\rho}},\ \left(\frac{3}{4U_3}\right)^{\frac{1+\tau}{\rho-\tau}} \right\}. \tag{4.3.38}$$

For this purpose, we present the following lemma.

Lemma 4.3.6. *The following three properties hold.*

(a) *Let μ, α and K_1 be positive numbers. Furthermore, let A be a matrix included in $\mathbf{S}^p$. If $\mu \in (0, (\frac{3}{4K_1})^{\frac{1}{\alpha}}]$ and $\|A - \mu I\|_F \leq K_1 \mu^{1+\alpha}$, then $A \succ 0$.*

(b) *Let μ, β and K_2 be positive numbers. Furthermore, let $\Phi : [0,1] \to \mathbf{S}^p$ be a function. If $\mu \in (0, (\frac{3}{4K_2})^{\frac{1}{\beta}}]$, $\Phi(0) \succ 0$ and $\|t^{-1}[\Phi(t) - (1-t)\Phi(0)] - \mu I\|_F \leq K_2\mu^{1+\beta}$ for any $t \in (0,1]$, then $\Phi(t) \succ 0$ for all $t \in (0,1]$.*

(c) *Let $w \in \mathcal{W}$, $d_x \in \mathbf{R}^n$ and $D_Z \in \mathbf{S}^p$. Furthermore, let $\Phi : [0,1] \to \mathbf{S}^p$ be defined by $\Phi(t) := X(x + td_x) \circ (Z + tD_Z)$. If $\Phi(t) \succ 0$ for all $t \in (0,1]$, then $X(x + d_x) \succ 0$ and $Z + D_Z \succ 0$.*

Proof. We first prove (a). Since $\mu \in (0, (\frac{3}{4K_1})^{\frac{1}{\alpha}}]$ and $\|A - \mu I\|_F \leq K_1\mu^{1+\alpha}$, we have $\|A - \mu I\|_2 \leq \|A - \mu I\|_F \leq K_1\mu^{1+\alpha} = K_1\mu^\alpha\mu \leq \frac{3}{4}\mu$, where the extreme left-hand side inequality follows from Proposition 2.2.1 (a). Thus, we have $v^\top Av = v^\top(A - \mu I)v + \mu\|v\|^2 \geq (\mu - \|A - \mu I\|_2)\|v\|^2 \geq \frac{1}{4}\mu\|v\|^2 > 0$ for all $v(\neq 0) \in \mathbf{R}^p$, where the first inequality follows from the Cauchy-Schwarz inequality and the definition of $\|\cdot\|_2$. Therefore, this inequality implies that $A \succ 0$.

Secondly, we show (b). It follows from (a) that $t^{-1}[\Phi(t) - (1-t)\Phi(0)] \succ 0$ for all $t \in (0,1]$. If $t = 1$, then $\Phi(1) \succ 0$. On the other hand, if $t \in (0,1)$, then $\Phi(t) \succ (1-t)\Phi(0) \succ 0$ for all $t \in (0,1)$. Therefore, $\Phi(t) \succ 0$ for all $t \in (0,1]$.

Finally, we give the proof of (c), that is, we show that $X(x + td_x) \succ 0$ for any $t \in (0,1]$. To this end, suppose the opposite, i.e., there exists $\bar{t} \in (0,1]$ such that $X(x + \bar{t}d_x)$ is not positive definite. Note that $w \in \mathcal{W}$ implies that $X(x) \succ 0$. It follows from the continuity of eigenvalues of $X(x)$ that there exists $\widetilde{t} \in (0, \bar{t}]$ such that $\lambda_{\min}(X(x + \widetilde{t}d_x)) = 0$. Thus, $X(x + \widetilde{t}d_x)$ is singular, that is, there exists $v_0 \neq 0$ such that $X(x + \widetilde{t}d_x)v_0 = 0$. Then, we obtain $v_0^\top\Phi(\widetilde{t})v_0 = \frac{1}{2}\left[v_0^\top X(x + \widetilde{t}d_x)(Z + \widetilde{t}D_Z)v_0 + v_0^\top(Z + \widetilde{t}D_Z)X(x + \widetilde{t}d_x)v_0\right] = 0$. However, this contradicts $\Phi(\widetilde{t}) \succ 0$ for any $t \in (0,1]$. Similarly, $Z + tD_Z \succ 0$ for any $t \in (0,1]$. Therefore, $t = 1$ implies that $X(x + d_x) \succ 0$ and $Z + D_Z \succ 0$. □

Lemma 4.3.7. *Suppose that Assumptions 4.2.1, 4.3.1 and 4.3.2 hold, and that θ satisfies (4.3.38). If $w_k \in \Theta_2(\theta)$ and $w_k \in \mathcal{W}$, then $\hat{w}_k \in \mathcal{W}$.*

Proof. Let $\Phi : [0,1] \to \mathbf{S}^p$ be defined by $\Phi(t) := X(x_k + t\Delta x_k) \circ (Z_k + t\Delta Z_k)$. From $w_k \in \mathcal{W}$ and Lemma 4.3.6 (c), it suffices to show that $\Phi(t) \succ 0$ for all $t \in (0,1]$. Now, we see that $w_k \in \mathcal{V}_L$ by $w_k \in \Theta_2(\theta)$. Moreover, since Lemma 4.3.5 yield that $\hat{w}_k \in \mathcal{V}_L$, we have $w_k + t\Delta w_k \in \mathcal{V}_L$ for all $t \in (0,1]$. Thus, substituting $w_1 = w_k + t\Delta w_k,\ w_2 = w_k$ into (4.3.6),

$$\begin{aligned} t^2L_M\|\Delta w_k\|^2 &\geq \|\mathrm{svec}[X(x_k + t\Delta x_k) \circ (Z_k + t\Delta Z_k) - X(x_k) \circ Z_k] \\ &\qquad -t[(Z_k \otimes_S I)A(x_k)\Delta x_k + (X(x_k) \otimes_S I)\mathrm{svec}(\Delta Z_k)]\| \\ &= \|\mathrm{svec}[\Phi(t) - (1-t)\Phi(0) - t\mu_k I] + tP(w_k)\Delta x_k\| \\ &\geq \|\Phi(t) - (1-t)\Phi(0) - t\mu_k I\|_F - t\|P(w_k)\|_F\|\Delta x_k\|, \end{aligned} \tag{4.3.39}$$

where the equality follows from $(Z_k \otimes_S I)A(x_k)\Delta x_k + (X(x_k) \otimes_S I)\mathrm{svec}(\Delta Z_k) = \mathrm{svec}(\mu_k I - X(x_k) \circ Z_k) - P(w_k)\Delta x_k$ in the Newton equation (4.2.6). It follows from $\|\Delta x_k\| \leq \|\Delta w_k\|$, (4.3.29) and (4.3.39) that

$$t^2L_M\|\Delta w_k\|^2 \geq \|\Phi(t) - (1-t)\Phi(0) - t\mu_k I\|_F - t\mathcal{U}_P\mu_k^{\frac{\rho}{1+\tau}}\|\Delta w_k\|.$$

Since $\|\Delta w_k\| \leq U_2 \mu_k^{\frac{1}{1+\tau}}$ by Lemma 4.3.5, we have from $t, \mu_k \in (0,1]$ that

$$\|\Phi(t) - (1-t)\Phi(0) - t\mu_k I\|_F \leq tL_M U_2^2 \mu_k^{\frac{2}{1+\tau}} + t\mathcal{U}_P U_2 \mu_k^{\frac{1+\rho}{1+\tau}} \leq tU_3 \mu_k^{1+\frac{\rho-\tau}{1+\tau}}.$$

Dividing both sides by $t \in (0,1]$, we obtain

$$\left\| \frac{\Phi(t) - (1-t)\Phi(0)}{t} - \mu_k I \right\|_F \leq U_3 \mu_k^{1+\frac{\rho-\tau}{1+\tau}}. \tag{4.3.40}$$

Meanwhile, we have from (4.3.27) and the definition of $\Gamma(\theta)$ that there exists $\eta_k \in (0, \theta]$ such that $w_k \in \mathcal{N}_2(\eta_k)$. In addition, $\eta_k \in (0, (\frac{3}{4})^{\frac{1}{\rho}}]$ by (4.3.38). Then, the definitions of $r_\kappa(w_k, \eta_k)$ and $\mathcal{N}_2(\eta_k)$ imply that $\|\Phi(0) - \eta_k I\|_F \leq \|r_\kappa(w_k, \eta_k)\| \leq \eta_k^{1+\rho}$. Thus, $\Phi(0) \succ 0$ by Lemma 4.3.6 (a). Moreover, (4.3.28) and (4.3.38) yield that $\mu_k \in (0, (\frac{3}{4U_3})^{\frac{1+\tau}{\rho-\tau}}]$. It then follows from (4.3.40) and Lemma 4.3.6 (b) that $\Phi(t) \succ 0$ for all $t \in (0,1]$. □

We summarize the results of Lemmas 4.3.5 and 4.3.7. Suppose that $\theta \in (0, \theta_5]$, $w_k \in \Theta_2(\theta)$ and $w_k \in \mathcal{W}$. Lemmas 4.3.5 and 4.3.7 imply that

$$\hat{w}_k \in \mathcal{N}_1(\mu_k),\ \hat{w}_k \in \mathcal{W},\ \|\Delta w_k\| \leq U_2 \mu_k^{\frac{1}{1+\tau}},\ \|r_\kappa(\hat{w}_k, \mu_k)\| \leq U_3 \mu_k^{1+\frac{\rho-\tau}{1+\tau}}. \tag{4.3.41}$$

Note that $w_k, \hat{w}_k \in \mathcal{V}_L$. We have from (4.3.2) and (4.3.41) that

$$\|M_0(\hat{w}_k) - M_0(w_k)\|_F = \|M_0(w_k + \Delta w_k) - M_0(w_k)\|_F \leq L_M \|\Delta w_k\| \leq \mathcal{U}_M \mu_k^{\frac{1}{1+\tau}}, \tag{4.3.42}$$

where $\mathcal{U}_M := L_M U_2$.

Next, we show that the sequence $\{w_k\}$ generated by Algorithm 4.2.3 is included in $\Theta_2(\theta)$ and $\mathcal{W}$. In what follows, suppose that θ satisfies

$$0 < \theta \leq \theta_6, \quad \theta_6 := \min \left\{ \theta_5,\ \frac{\nu_N}{U_5},\ \left(\frac{1}{U_6}\right)^{\frac{1}{h_2}} \right\}, \tag{4.3.43}$$

where

$$h_2 := \frac{2\rho - \tau}{1+\tau} - \rho, \quad U_4 := U_{\mathcal{M}} U_3, \quad U_5 := U_r(U_1 + U_3) + U_4, \quad U_6 := U_4(L_M U_4 + \mathcal{U}_M + \mathcal{U}_P).$$

Note that $h_2 > 0$ from (4.3.16). First of all, we show $w_{k+1} \in \mathcal{V}_N$ and $\|r_\kappa(w_{k+1}, \mu_k)\| \leq \mu_k^{1+\rho}$.

Lemma 4.3.8. *Suppose that Assumptions 4.2.1, 4.3.1 and 4.3.2 hold, and that θ satisfies (4.3.43). If $w_k \in \Theta_2(\theta)$ and $w_k \in \mathcal{W}$, then*

$$\hat{w}_k \in \mathcal{N}_1(\mu_k) \subset \Theta_1(\theta), \quad \mathcal{N}_2(\mu_k) \subset \Theta_2(\theta),$$
$$\|\Delta \hat{w}_k\| \leq U_4 \mu_k^{1+\frac{\rho-\tau}{1+\tau}},\ \|w_{k+1} - w^*\| \leq \nu_N,\ \|r_\kappa(w_{k+1}, \mu_k)\| \leq \mu_k^{1+\rho}.$$

Proof. Note that $w_{k+1} = \hat{w}_k + \Delta \hat{w}_k$. First, we show that $\hat{w}_k \in \mathcal{N}_1(\mu_k) \subset \Theta_1(\theta)$ and $\mathcal{N}_2(\mu_k) \subset \Theta_2(\theta)$. Since $0 < \mu_k < \theta$ by (4.3.28), the definitions of $\Theta_1(\theta)$ and $\Theta_2(\theta)$ imply that $\mathcal{N}_1(\mu_k) \subset \cup_{\mu \in [0,\theta]} \mathcal{N}_1(\mu) = \Theta_1(\theta)$ and $\mathcal{N}_2(\mu_k) \subset \cup_{\mu \in [0,\theta]} \mathcal{N}_2(\mu) = \Theta_2(\theta)$, respectively. Furthermore, using $\hat{w}_k \in \mathcal{N}_1(\mu_k)$ in (4.3.41), we have the desired result.

Next, we prove that $\|\Delta\hat{w}_k\| \leq U_4\mu_k^{1+\frac{\rho-\tau}{1+\tau}}$ and $\|w_{k+1}-w^*\| \leq \nu_N$. We have from (4.3.23) and Lemma 4.3.4 that $\Delta\hat{w}_k = -M_P(w_k,\mu_k)^{-1}r_\kappa(\hat{w}_k,\mu_k)$. Moreover, (4.3.33) and (4.3.41) yield that

$$\|\Delta\hat{w}_k\| \leq \|M_P(w_k,\mu_k)^{-1}\|_F\|r_\kappa(\hat{w}_k,\mu_k)\| \leq U_{\mathcal{M}}\|r_\kappa(\hat{w}_k,\mu_k)\| \leq U_4\mu_k^{1+\frac{\rho-\tau}{1+\tau}}. \tag{4.3.44}$$

On the other hand, it is clear that $\hat{w}_k \in \mathcal{V}$ and $\mu_k \in [0,s]$ from $w_k \in \Theta_2(\theta)$, (4.3.28) and (4.3.43). Thus, substituting $w=\hat{w}_k$ and $\mu=\mu_k$ into (4.3.10), and using (4.3.41) and $\mu_k \in (0,1)$, we get

$$\|r_\kappa(\hat{w}_k,0)\| \leq U_1\mu_k + \|r_\kappa(\hat{w}_k,\mu_k)\| \leq U_1\mu_k + U_3\mu_k^{1+\frac{\rho-\tau}{1+\tau}} \leq (U_1+U_3)\mu_k. \tag{4.3.45}$$

It follows from Lemma 4.3.2 that $\|w_{k+1}-w^*\| \leq \|\hat{w}_k-w^*\| + \|\Delta\hat{w}_k\| \leq U_r\|r_\kappa(\hat{w}_k,0)\| + \|\Delta\hat{w}_k\|$. Using (4.3.44), (4.3.45) and $\mu_k \in (0,1)$, we obtain $\|w_{k+1}-w^*\| \leq [U_r(U_1+U_3)+U_4]\,\mu_k = U_5\mu_k$. Moreover, since $U_5\mu_k \leq U_5\theta \leq \nu_N$ by (4.3.28) and (4.3.43), we have $\|w_{k+1}-w^*\| \leq \nu_N$.

Finally, we prove that $\|r_\kappa(w_{k+1},\mu_k)\| \leq \mu_k^{1+\rho}$. It follows from (4.3.23) and (4.3.24) that $r_\kappa(\hat{w}_k,\mu_k) = -M_P(w_k,\mu_k)\Delta\hat{w}_k = -(M(w_k,\mu_k)+N(w_k))\Delta\hat{w}_k$. Then, since $\hat{w}_k, w_{k+1} \in \mathcal{V}_L$ and $\mu_k \geq 0$, we substitute $w_1 = w_{k+1}, w_2 = \hat{w}_k$ and $\mu=\mu_k$ into (4.3.5), that is,

$$\begin{aligned}
L_M\|\Delta\hat{w}_k\|^2 &\geq \|r_\kappa(w_{k+1},\mu_k) - r_\kappa(\hat{w}_k,\mu_k) - M(\hat{w}_k,\mu_k)\Delta\hat{w}_k\| \\
&\geq \|r_\kappa(w_{k+1},\mu_k)\| - \|M(\hat{w}_k,\mu_k) - M(w_k,\mu_k) - N(w_k)\|_F\|\Delta\hat{w}_k\| \\
&\geq \|r_\kappa(w_{k+1},\mu_k)\| - \|M(\hat{w}_k,\mu_k) - M(w_k,\mu_k)\|_F\|\Delta\hat{w}_k\| - \|N(w_k)\|_F\|\Delta\hat{w}_k\| \\
&= \|r_\kappa(w_{k+1},\mu_k)\| - \|M_0(\hat{w}_k) - M_0(w_k)\|_F\|\Delta\hat{w}_k\| - \|P(w_k)\|_F\|\Delta\hat{w}_k\|,
\end{aligned}$$

where the last equality follows from (4.3.1) and (4.3.25). Using (4.3.29) and (4.3.42), we get $\|r_\kappa(w_{k+1},\mu_k)\| \leq L_M\|\Delta\hat{w}_k\|^2 + \mathcal{U}_M\mu_k^{\frac{1}{1+\tau}}\|\Delta\hat{w}_k\| + \mathcal{U}_P\mu_k^{\frac{\rho}{1+\tau}}\|\Delta\hat{w}_k\|$, and hence $\|r_\kappa(w_{k+1},\mu_k)\| \leq L_MU_4^2\mu_k^{2+\frac{2(\rho-\tau)}{1+\tau}} + \mathcal{U}_MU_4\mu_k^{1+\frac{1+\rho-\tau}{1+\tau}} + \mathcal{U}_PU_4\mu_k^{1+\frac{2\rho-\tau}{1+\tau}} \leq U_6\mu_k^{1+\frac{2\rho-\tau}{1+\tau}}$ from (4.3.44) and $\mu_k \in (0,1)$. Since (4.3.43) implies $U_6\theta^{h_2} \leq 1$, we have from (4.3.28) that $U_6\mu_k^{1+\frac{2\rho-\tau}{1+\tau}} = U_6\mu_k^{h_2}\mu_k^{1+\rho} \leq U_6\theta^{h_2}\mu_k^{1+\rho} \leq \mu_k^{1+\rho}$. Note that $h_2 = \frac{2\rho-\tau}{1+\tau} - \rho > 0$ by (4.3.16). Thus, $\|r_\kappa(w_{k+1},\mu_k)\| \leq \mu_k^{1+\rho}$. □

Finally, we prove that the sequence $\{w_k\}$ generated by Algorithm 4.2.3 is included in $\mathcal{W}$. Let $\widetilde{\theta}$ be defined by

$$\widetilde{\theta} := \min\left\{\theta_6,\ \left(\frac{3}{4}\right)^{\frac{1}{\sigma}},\ \left(\frac{3}{4U_7}\right)^{\frac{1+\tau}{2\rho-\tau}}\right\}, \tag{4.3.46}$$

where $U_7 := U_4(L_MU_4 + 2\mathcal{U}_M + \mathcal{U}_P)$.

Lemma 4.3.9. *Suppose that Assumptions 4.2.1, 4.3.1 and 4.3.2 hold. If $w_k \in \Theta_2(\widetilde{\theta})$ and $w_k \in \mathcal{W}$, then $w_{k+1} \in \mathcal{W}$.*

Proof. Let $\Phi : [0,1] \to \mathbf{S}^p$ be defined by $\Phi(t) := X(\hat{x}_k + t\Delta\hat{x}_k) \circ (\hat{Z}_k + t\Delta\hat{Z}_k)$. We see that $\hat{w}_k \in \mathcal{W}$ by (4.3.41). Then, from Lemma 4.3.6 (c), it suffices to prove that $\Phi(t) \succ 0$ for all $t \in (0,1]$. We easily see that $\hat{w}_k, \hat{w}_k + \Delta\hat{w}_k \in \mathcal{V}_L$ by Lemmas 4.3.5 and 4.3.8. It then follows that $\hat{w}_k + t\Delta\hat{w}_k \in \mathcal{V}_L$ for all $t \in (0,1]$. Thus, substituting $w_1 = \hat{w}_k + t\Delta\hat{w}_k,\ w_2 = \hat{w}_k$ into (4.3.6), we

have

$$\begin{aligned}
t^2 L_M \|\Delta\hat{w}_k\|^2 &\geq \|\mathrm{svec}[X(\hat{x}_k + t\Delta\hat{x}_k) \circ (\hat{Z}_k + t\Delta\hat{Z}_k) - ((1-t)+t)X(\hat{x}_k) \circ \hat{Z}_k \\
&\qquad -t[(\hat{Z}_k \otimes_S I)A(\hat{x}_k)\Delta\hat{x}_k + (X(\hat{x}_k) \otimes_S I)\mathrm{svec}(\Delta\hat{Z}_k)]\| \\
&= \|\mathrm{svec}[\Phi(t) - (1-t)\Phi(0)] - t\mathrm{svec}[X(\hat{x}_k) \circ \hat{Z}_k] \\
&\qquad -t[(\hat{Z}_k \otimes_S I)A(\hat{x}_k)\Delta\hat{x}_k + (X(\hat{x}_k) \otimes_S I)\mathrm{svec}(\Delta\hat{Z}_k)]\| \\
&\geq \|\Phi(t) - (1-t)\Phi(0) - t\mu_k I\|_F - t\|P(w_k)\|_F\|\Delta\hat{w}_k\| \\
&\qquad -t\|(\hat{Z}_k \otimes_S I)A(\hat{x}_k) - (Z_k \otimes_S I)A(x_k)\|_F\|\Delta\hat{w}_k\| \\
&\qquad -t\|X(\hat{x}_k) \otimes_S I - X(x_k) \otimes_S I\|_F\|\Delta\hat{w}_k\| \\
&\geq \|\Phi(t) - (1-t)\Phi(0) - t\mu_k I\|_F - t\|P(w_k)\|_F\|\Delta\hat{w}_k\| \\
&\qquad -2t\|M_0(\hat{w}_k) - M_0(w_k)\|_F\|\Delta\hat{w}_k\|,
\end{aligned}$$

where the second inequality follows from $\mathrm{svec}[X(\hat{x}_k) \circ \hat{Z}_k] = \mathrm{svec}(\mu_k I) - (Z_k \otimes_S I)A(x_k)\Delta\hat{x}_k - (X(x_k) \otimes_S I)\mathrm{svec}(\Delta\hat{Z}_k) - P(w_k)\Delta\hat{x}_k$ in the Newton equation (4.2.7), and the last inequality follows from (4.3.3) and (4.3.4). Then, we exploit (4.3.29), (4.3.42) and Lemma 4.3.8, i.e., $\|P(w_k)\|_F \leq \mathcal{U}_P \mu_k^{\frac{\rho}{1+\tau}}$, $\|M_0(\hat{w}_k) - M_0(w_k)\|_F \leq \mathcal{U}_M \mu_k^{\frac{1}{1+\tau}}$ and $\|\Delta\hat{w}_k\| \leq U_4 \mu_k^{1+\frac{\rho-\tau}{1+\tau}}$. As the result, we get

$$\begin{aligned}
\|\Phi(t) - (1-t)\Phi(0) - t\mu_k I\|_F &\leq t^2 L_M \|\Delta\hat{w}_k\|^2 + t\|P(w_k)\|_F\|\Delta\hat{w}_k\| \\
&\qquad +2t\|M_0(\hat{w}_k) - M_0(w_k)\|_F\|\Delta\hat{w}_k\| \\
&\leq t^2 L_M U_4^2 \mu_k^{2+\frac{2(\rho-\tau)}{1+\tau}} + 2t\mathcal{U}_M U_4 \mu_k^{1+\frac{1+\rho-\tau}{1+\tau}} + t\mathcal{U}_P U_4 \mu_k^{1+\frac{2\rho-\tau}{1+\tau}} \\
&\leq t U_7 \mu_k^{1+\frac{2\rho-\tau}{1+\tau}},
\end{aligned}$$

where the last inequality follows from $t, \mu_k \in (0, 1]$. Dividing both sides by $t \in (0, 1]$, we obtain

$$\left\| \frac{\Phi(t) - (1-t)\Phi(0)}{t} - \mu_k I \right\|_F \leq U_7 \mu_k^{1+\frac{2\rho-\tau}{1+\tau}}. \tag{4.3.47}$$

On the other hand, we have from (4.3.28) and (4.3.46) that $0 < \mu_k < \min\{(\frac{3}{4})^{\frac{1}{\sigma}}, (\frac{3}{4U_7})^{\frac{1+\tau}{2\rho-\tau}}\}$. In addition, Lemma 4.3.8 yields that $\hat{w}_k \in \mathcal{N}_1(\mu_k)$. Then, the definitions of $r_\kappa(\hat{w}_k, \mu_k)$ and $\mathcal{N}_1(\mu_k)$ imply that $\|\Phi(0) - \mu_k I\|_F \leq \|r_\kappa(\hat{w}_k, \mu_k)\| \leq \mu_k^{1+\sigma}$. Thus, $\Phi(0) \succ 0$ by Lemma 4.3.6 (a). It then follows from (4.3.47) and Lemma 4.3.6 (b) that $\Phi(t) \succ 0$ for all $t \in (0, 1]$. □

Using Lemmas 4.3.8 and 4.3.9, we prove that $\{w_k\}$ converges to w^* superlinearly.

Theorem 4.3.3. *Suppose that Assumptions 4.2.1, 4.3.1 and 4.3.2 hold. If $w_0 \in \Theta_2(\widetilde{\theta})$ and $w_0 \in \mathcal{W}$, the sequence $\{w_k\}$ generated by Algorithm 4.2.3 converges to w^* superlinearly.*

Proof. Note that $\widetilde{\theta} \leq \theta_i$ $(i = 1, \ldots, 6)$ from the definitions of $\theta_1, \ldots, \theta_6$ and $\widetilde{\theta}$. First, we show the following relations by the mathematical induction: For all positive integer k,

$$w_k \in \mathcal{N}_2(\mu_{k-1}) \subset \Theta_2(\widetilde{\theta}), \quad w_k \in \mathcal{W}. \tag{4.3.48}$$

Since $w_0 \in \Theta_2(\widetilde{\theta})$ and $w_0 \in \mathcal{W}$, we have from Lemmas 4.3.8 and 4.3.9 that $w_1 \in \mathcal{N}_2(\mu_0) \subset \Theta_2(\widetilde{\theta})$ and $w_1 \in \mathcal{W}$. Next, let $k \geq 2$. Suppose that $w_k \in \mathcal{N}_2(\mu_{k-1}) \subset \Theta_2(\widetilde{\theta})$ and $w_k \in \mathcal{W}$. Then, it

follows from Lemmas 4.3.8 and 4.3.9 that $w_{k+1} \in \mathcal{N}_2(\mu_k) \subset \Theta_2(\widetilde{\theta})$ and $w_{k+1} \in \mathcal{W}$. Therefore, the proof of (4.3.48) is complete.

Secondly, we prove that $\{w_k\}$ converges to w^* superlinearly. Let k be an arbitrary positive integer. Note that $0 < \mu_k < \widetilde{\theta} \leq \theta_1 = \min\{\gamma, s, \theta_0\} < 1$ from (4.3.28). Then, note also that $w_{k+1} \in \mathcal{N}_2(\mu_k) \subset \Theta_2(\widetilde{\theta}) \subset \Theta_1(\widetilde{\theta}) \subset \mathcal{V}_N \subset \mathcal{V}$ by (4.3.15), (4.3.17) and (4.3.48). Lemma 4.3.2 and (4.3.10) yield that $\|w_{k+1} - w^*\| \leq U_r \|r_\kappa(w_{k+1}, 0)\| \leq U_r(\|r_\kappa(w_{k+1}, \mu_k)\| + U_1\mu_k) \leq U_r(\mu_k^\rho + U_1)\mu_k \leq U_r(1 + U_1)\mu_k$. Thus, we obtain $\|w_{k+1} - w^*\| \leq U_r(1 + U_1)\|r_\kappa(w_k, 0) - r_\kappa(w^*, 0)\|^{1+\tau}$ by $\mu_k = \|r_\kappa(w_k, 0)\|^{1+\tau}$ and $r_\kappa(w^*, 0) = 0$. It then follows from $w_k, w^* \in \mathcal{V}_L$ and (4.3.12) that

$$\|w_{k+1} - w^*\| \leq U_r(1 + U_1)\|r_\kappa(w_k, 0) - r_\kappa(w^*, 0)\|^{1+\tau} \leq L_r^{1+\tau} U_r(1 + U_1)\|w_k - w^*\|^{1+\tau}. \quad (4.3.49)$$

Using (4.3.14), $w_k \in \mathcal{V}_N$ and $U_r = 4U_M$ that

$$L_r^{1+\tau} U_r(1 + U_1)\|w_k - w^*\|^\tau \leq L_r^{1+\tau} U_r(1 + U_1)\nu_N^\tau \leq \frac{U_r}{5U_M} = \frac{4}{5}. \quad (4.3.50)$$

It follows from (4.3.49) and (4.3.50) that $\|w_{k+1} - w^*\| \leq \frac{4}{5}\|w_k - w^*\|$, and hence $\{\|w_k - w^*\|\}$ converges to 0. Since $\lim_{k\to\infty} L_r^{1+\tau} U_r(1 + U_1)\|w_k - w^*\|^\tau = 0$, we have from (4.3.49) that

$$\lim_{k\to\infty} \frac{\|w_{k+1} - w^*\|}{\|w_k - w^*\|} = 0.$$

Therefore, $\{w_k\}$ converges to w^* superlinearly. □

4.4 Numerical experiments

In this section, we report several numerical experiments for Algorithm 4.2.3. We compare the proposed method with Yamashita and Yabe's two-step method [71], that is, Algorithm 4.2.2. We provide the number of iterations and the CPU time of Algorithms 4.2.2 and 4.2.3. The program is written in MATLAB R2010a and tun on a machine with an Intel Core i7 920 2.67GHz CPU and 3.00GB RAM. We adopt the HRVW/KSH/M choice as the scaling matrix T_k, that is, $T_k = X_k^{-\frac{1}{2}}$. Moreover, we select the following parameters:

$$\begin{aligned} &\text{Algorithm 4.2.2}: \quad \epsilon = 10^{-6},\ \tau = 0.33, \\ &\text{Algorithm 4.2.3}: \quad \epsilon = 10^{-6},\ \kappa = 1,\ \tau = 0.49. \end{aligned}$$

The test problems used in the experiments are the Gaussian channel capacity problem and the nearest correlation problem. These problems are exactly the same ones given in Section 3.4. Note that these problems satisfy Assumption 4.2.1 because they are convex programming. For these details, see Section 3.4. We choose an initial point of Algorithms 4.2.2 and 4.2.3 as follows. First, we solve the test problem by using Algorithm 3.2.1, and obtain a point $w = [x, y, \mathrm{svec}(Z)] \in \mathbf{R}^l$ such that $\|r_1(w, 0)\| \leq 10^{-3}$ and $w \in \mathcal{W}$. Next, we set w as an initial point.

Tables 4.1 and 4.2 show the numerical results. From Tables 4.1 and 4.2, all the iteration counts of Algorithm 4.2.3 were less than those of Algorithm 4.2.2. Therefore, we can guess that the computational cost of Algorithm 4.2.3 was less than half of Algorithm 4.2.2. Indeed, in most experiments, Algorithm 4.2.3 was able to find a solution in less than half the time of Algorithm 4.2.2.

Table 4.1: Gaussian channel capacity problem

	Algorithm 4.2.2		Algorithm 4.2.3	
n	iteration	time(s)	iteration	time(s)
5	4	0.07	3	0.08
10	6	0.24	3	0.09
15	8	1.16	4	0.32
20	12	4.49	4	0.89
25	5	4.39	2	0.93
30	5	8.65	2	1.89
35	6	19.40	3	5.22
40	6	34.61	3	8.99

Table 4.2: Nearest correlation matrix problem

	Algorithm 4.2.2		Algorithm 4.2.3	
n	iteration	time(s)	iteration	time(s)
5	3	0.05	2	0.03
10	4	0.48	2	0.12
15	4	2.01	3	0.88
20	5	7.71	3	2.41
25	3	19.13	2	6.05
30	3	47.38	2	14.11
35	4	121.27	3	45.77
40	4	215.30	3	80.02

4.5 Concluding remarks

In this chapter, we proposed a two-step primal-dual interior point method (Algorithm 4.2.3) based on the generalized shifted barrier KKT conditions (4.2.1) for the nonlinear SDP and proved the superlinear convergence of the proposed method. In particular, in order to reduce calculations, we replaced the coefficient matrix in the second equation with that in the first one. Therefore, we can expect that the proposed method can find the next point faster than the existing methods [40, 48, 71]. In the numerical experiments, we actually showed that Algorithm 4.2.3 can find a solution faster than Algorithm 4.2.2.

As a future work, it is desired to prove the superlinear convergence of a one-step method with scaling.

Appendix A

In this Appendix, we show that there exists $P(w)$ such that Assumption 4.2.1 holds for $T = X^{-\frac{1}{2}}$ and $T = W^{-\frac{1}{2}}$, and Assumption 4.3.2 also holds.

In what follows, we define $E(\eta) := XZ - \eta I$ ($\eta \in \mathbf{R}$). First, we give two inequalities which evaluate $E(\eta)$ and X^{-1} over $\Theta_2(\theta)$. Secondly, we also give an inequality which evaluates $A \otimes_S B$ for any $A, B \in \mathbf{R}^{p\times p}$. These inequalities play important roles in evaluation of $P(w)$.

Lemma A.1. *Suppose that Assumption 4.3.1 holds, and that $\theta \in (0, \theta_2]$. Then, we obtain $\|E(\eta)\|_F \le U_R\eta^{1+\rho}$ and $\|X^{-1}\|_F \le U_X\eta^{-1}$ for any $(w,\eta) \in \Gamma(\theta)$, where $U_X := 4pU_Z$ and $U_Z := \sup\{\|Z\|_F | w \in \Theta_2(\theta_2), w \in \mathcal{W}\}$.*

Proof. For any $(w,\eta) \in \Gamma(\theta)$, we have from the definition of $\Gamma(\theta)$ that $w \in \mathcal{N}_2(\eta) \subset \Theta_2(\theta)$, $w \in \mathcal{W}$ and $\eta \in (0,\theta]$. Thus, we also have from the definition of $\mathcal{N}_2(\eta)$ that $\|r_\kappa(w,\eta)\| \le \eta^{1+\rho}$. Moreover, $w \in \Theta_2(\theta) \subset \Theta_1(\theta)$, $\eta \in (0,\theta] \subset [0,\theta_1]$ and Lemma 4.3.2 yield that $\|E(\eta)\|_F \le U_R\|r_\kappa(w,\eta)\| \le U_R\eta^{1+\rho}$. It then follows from $\eta \le \theta \le \theta_2 \le (\frac{3}{4U_R})^{\frac{1}{\rho}}$ that $\|I - \eta^{-1}XZ\|_F = \eta^{-1}\|E(\eta)\|_F \le U_R\eta^\rho \le \frac{3}{4}$. Thus, $\eta\frac{\|X^{-1}\|_F}{\|Z\|_F} \le \eta\|Z^{-1}X^{-1}\|_F = \|(I-(I-\eta^{-1}XZ))^{-1}\|_F \le \frac{p}{1-\|I-\eta^{-1}XZ\|_F} \le 4p$, where the second inequality follows from Proposition 2.2.1 (d). Since $w \in \Theta_2(\theta) \subset \Theta_2(\theta_2)$ and $w \in \mathcal{W}$, we obtain $0 < \|Z\|_F \le U_Z$. Hence, we get $\|X^{-1}\|_F \le 4pU_Z\eta^{-1}$. Letting $U_X = 4pU_Z$, we obtain the desired inequality. □

(i) HRVW/KSH/M ($T = X^{-\frac{1}{2}}$)

First, we discuss the case where $T = X^{-\frac{1}{2}}$. We define

$$F(w,\eta) := \frac{1}{2}\left[E(\eta) \otimes_S X^{-1} - I \otimes_S (X^{-1}E(\eta))\right] A(x).$$

Then, letting $P(w) := F(w,\eta)$, we see that Assumption 4.2.1 (S1) holds. Note that we can choose η arbitrarily.

Next, we show that Assumption 4.3.2 (S2) holds. Suppose that Assumption 4.3.1 holds, and that $\theta \in (0,\theta_2]$. For any $(w,\eta) \in \Gamma(\theta)$, it follows from Proposition 2.2.10 that

$$\|F(w,\eta)\|_F \le \frac{1}{2}U_A\left[\|E(\eta)\otimes_S X^{-1}\|_F + \|I \otimes_S (X^{-1}E(\eta))\|_F\right] \le C_2U_A\|E(\eta)\|_F\|X^{-1}\|_F, \quad \text{(A.1)}$$

where $U_A := \sup\{\|A(x)\|_F | w \in \Theta_2(\theta_2), w \in \mathcal{W}\}$ and $C_2 := \frac{1+\sqrt{p}}{2}\sqrt{\frac{p(p+1)}{2}}$. We get $\|P(w)\|_F = \|F(w,\eta)\|_F \le C_2U_AU_RU_X\eta^\rho$ from Lemma A.1 and (A.1). Therefore, letting $U_P := C_2U_AU_RU_X$, we see that Assumption 4.3.2 (S2) holds.

(ii) NT ($T = W^{-\frac{1}{2}}$)

In this part, we discuss the case where we choose $T = W^{-\frac{1}{2}}$, where $W = X^{\frac{1}{2}}(X^{\frac{1}{2}}ZX^{\frac{1}{2}})^{-\frac{1}{2}}X^{\frac{1}{2}}$. We define

$$\begin{aligned}
G(w,\eta) := &-\frac{1}{2}(I \otimes_S (E(\eta)^\top X^{-1}))A(x) + \eta(I \otimes_S (X^{-\frac{1}{2}}H(w,\eta)X^{-\frac{1}{2}}))A(x) + \frac{1}{2\eta}(E(\eta)\otimes_S Z)A(x)\\
&-\frac{1}{4\eta}(E(\eta)\otimes_S (E(\eta)^\top X^{-1}))A(x) + \frac{1}{2}(E(\eta)\otimes_S (X^{-\frac{1}{2}}H(w,\eta)X^{-\frac{1}{2}}))A(x)\\
&+((X^{\frac{1}{2}}H(w,\eta)X^{-\frac{1}{2}})\otimes_S Z)A(x) - \frac{1}{2}((X^{\frac{1}{2}}H(w,\eta)X^{-\frac{1}{2}})\otimes_S (E(\eta)^\top X^{-1}))A(x)\\
&+\eta((X^{\frac{1}{2}}H(w,\eta)X^{-\frac{1}{2}})\otimes_S (X^{-\frac{1}{2}}H(w,\eta)X^{-\frac{1}{2}}))A(x),
\end{aligned}$$

where $H(w,\eta) := \left(I + \frac{1}{\eta}X^{-\frac{1}{2}}E(\eta)X^{\frac{1}{2}}\right)^{\frac{1}{2}} - I - \frac{1}{2\eta}X^{-\frac{1}{2}}E(\eta)X^{\frac{1}{2}}$. Then, letting $P(w) := G(w,\eta)$, we see that Assumption 4.2.1 (S1) holds. Note that we can choose η arbitrarily.

Next, we prove that Assumption 4.3.2 (S2) holds. Suppose that Assumption 4.3.1 holds, and that $\theta \in (0, \theta_2]$. For any $(w,\eta) \in \Gamma(\theta)$, it follows from Proposition 2.2.10 that there exists $C_3 > 0$ such that

$$\begin{aligned}
&\|G(w,\eta)\|_F \\
&\quad \le C_3 U_A(\|I\|_F\|E(\eta)\|_F\|X^{-1}\|_F + \eta\|I\|_F\|X^{-\frac{1}{2}}\|_F^2\|H(w,\eta)\|_F + \eta^{-1}\|E(\eta)\|_F\|Z\|_F \\
&\qquad +\eta^{-1}\|E(\eta)\|_F^2\|X^{-1}\|_F + \|X^{-\frac{1}{2}}\|_F^2\|H(w,\eta)\|_F\|E(\eta)\|_F + \|Z\|_F\|X^{\frac{1}{2}}\|_F\|H(w,\eta)\|_F\|X^{-\frac{1}{2}}\|_F \\
&\qquad +\|X^{\frac{1}{2}}\|_F\|H(w,\eta)\|_F\|X^{-\frac{1}{2}}\|_F\|E(\eta)\|_F\|X^{-1}\|_F + \eta\|X^{-\frac{1}{2}}\|_F^3\|H(w,\eta)\|_F^2\|X^{\frac{1}{2}}\|_F). \qquad \text{(A.2)}
\end{aligned}$$

In what follows, we evaluate $\|X^{-\frac{1}{2}}\|_F$ and $\|H(w,\eta)\|_F$. First, it follows from Lemma A.1 that

$$\|X^{-\frac{1}{2}}\|_F = \sqrt{\mathrm{tr}(X^{-1})} = \sqrt{\langle X^{-1}, I\rangle} \le \sqrt{\|I\|_F\|X^{-1}\|_F} \le \sqrt{p^{\frac{1}{2}}U_X}\eta^{-\frac{1}{2}}. \qquad \text{(A.3)}$$

Next, we evaluate $\|H(w,\eta)\|_F$. For this purpose, we first evaluate $\eta^{-1}X^{-\frac{1}{2}}E(\eta)X^{\frac{1}{2}}$ which constitutes $H(w,\eta)$. Since $E(\eta) = XZ - \eta I$ implies $X^{-\frac{1}{2}}E(\eta)X^{\frac{1}{2}} = X^{\frac{1}{2}}ZX^{\frac{1}{2}} - \eta I \in \mathbf{S}^p$, we obtain $\|X^{-\frac{1}{2}}E(\eta)X^{\frac{1}{2}}\|_F = \sqrt{\mathrm{tr}(X^{-\frac{1}{2}}E(\eta)X^{\frac{1}{2}}X^{-\frac{1}{2}}E(\eta)X^{\frac{1}{2}})} = \sqrt{\mathrm{tr}(E(\eta)^2)} = \sqrt{\langle E(\eta)^\top, E(\eta)\rangle}$. Then, the Cauchy-Schwarz inequality yields that

$$\|\eta^{-1}X^{-\frac{1}{2}}E(\eta)X^{\frac{1}{2}}\|_F \le \eta^{-1}\sqrt{\|E(\eta)^\top\|_F\|E(\eta)\|_F} = \eta^{-1}\|E(\eta)\|_F. \qquad \text{(A.4)}$$

Since $\eta^{-1}X^{-\frac{1}{2}}E(\eta)X^{\frac{1}{2}}$ is symmetric, all eigenvalues $\lambda_1, \ldots, \lambda_p$ are real numbers. Moreover, there exists an orthogonal matrix V such that $\eta^{-1}X^{-\frac{1}{2}}E(\eta)X^{\frac{1}{2}} = VDV^\top$, where $D = \mathrm{diag}[\lambda_1, \ldots, \lambda_p]$. From Lemma A.1, $0 \le \eta \le \theta \le \theta_2 \le (\frac{3}{4U_R})^{\frac{1}{\rho}}$ and (A.4),

$$1 > \frac{3}{4} \ge U_R\eta^\rho \ge \eta^{-1}\|E(\eta)\|_F \ge \|\eta^{-1}X^{-\frac{1}{2}}E(\eta)X^{\frac{1}{2}}\|_F = \sqrt{\sum_{i=1}^p \lambda_i^2} \ge |\lambda_i|,$$

for $i = 1, \ldots, p$. As the result, the matrix $I + D$ is symmetric positive definite, i.e., the existence of $(I+D)^{\frac{1}{2}}$ is guaranteed. Considering the diagonalization of $\eta^{-1}X^{-\frac{1}{2}}E(\eta)X^{\frac{1}{2}}$,

$$\begin{aligned}
\|H(w,\eta)\|_F &= \left\|\left(VV^\top + VDV^\top\right)^{\frac{1}{2}} - VV^\top - \frac{1}{2}VDV^\top\right\|_F \\
&= \left\|V(I+D)^{\frac{1}{2}}V^\top - VV^\top - \frac{1}{2}VDV^\top\right\|_F \\
&\le \|V\|_F^2\left\|(I+D)^{\frac{1}{2}} - I - \frac{1}{2}D\right\|_F \\
&= p\sqrt{\sum_{i=1}^p\left(\sqrt{1+\lambda_i} - 1 - \frac{1}{2}\lambda_i\right)^2}. \qquad \text{(A.5)}
\end{aligned}$$

Let $\varphi : (-1, 1) \to \mathbf{R}$ be defined by $\varphi(u) := \sqrt{1+u}$. Since φ is a twice continuously differentiable function defined on the bounded convex set, it follows from Proposition 2.2.2 (b) that there

exists a positive constant L_φ such that $L_\varphi u^2 \geq |\varphi(u) - \varphi(0) - \varphi'(0)u| = \left|\sqrt{1+u} - 1 - \frac{1}{2}u\right|$. Then, Lemma A.1, (A.4) and (A.5) yield that

$$\begin{aligned}
\|H(w,\eta)\|_F &\leq pL_\varphi \sqrt{\sum_{i=1}^{p} \lambda_i^4} \\
&= pL_\varphi \left\| \left(\eta^{-1} X^{-\frac{1}{2}} E(\eta) X^{\frac{1}{2}}\right)^2 \right\|_F \\
&\leq pL_\varphi \left\| \eta^{-1} X^{-\frac{1}{2}} E(\eta) X^{\frac{1}{2}} \right\|_F^2 \\
&\leq pL_\varphi \|E(\eta)\|_F^2 \eta^{-2} \\
&\leq pL_\varphi U_R^2 \eta^{2\rho}.
\end{aligned}$$

This inequality, Lemma A.1, (A.2), (A.3), the boundedness of $\|Z\|_F$ and $\|X^{\frac{1}{2}}\|_F$ imply that there exists $C_4 > 0$ such that $\|G(w,\eta)\|_F \leq C_4(\eta^\rho + \eta^{2\rho} + \eta^{3\rho} + \eta^{2\rho-\frac{1}{2}} + \eta^{3\rho-\frac{1}{2}} + \eta^{4\rho-\frac{1}{2}}) \leq 6C_4\eta^\rho$, where the second inequality follows from $\rho \leq 2\rho - \frac{1}{2}$ in (4.3.16) and $0 < \eta \leq \theta_2 \leq \theta_1 \leq s < 1$. Letting $U_P := 6C_4$, we prove that Assumption 4.3.2 (S2) holds.

These results show that $T = X^{-\frac{1}{2}}$ and $T = W^{-\frac{1}{2}}$ satisfy Assumptions 4.2.1 and 4.3.2.

Chapter 5

A block coordinate descent method for a maximum likelihood estimation problem of mixture distributions

5.1 Introduction

When some observational data are supposed to obey a certain distribution with parameters, it is important to estimate valid parameters from the data. A maximum likelihood estimation is one of estimation procedures of the parameters. In this chapter, we focus on the maximum likelihood estimation in mixture distributions, which is frequently used in statistics, pattern recognition and machine learning [8, 43].

The EM algorithm is known to be one of the most powerful methods for the estimation [16, 26, 51, 67], and has been studied actively even in recent years. It is an iterative method which consists of the Expectation Step (E-Step) and the Maximization Step (M-Step). The E-Step calculates an expectation of the likelihood and the M-Step maximizes the expectation with respect to parameters.

Recently, many researchers have actively studied the maximum likelihood estimation with regularization methods. For example, when we add the L_1 regularization to the likelihood function, we may choose important parameters in the model. The L_1 regularization is used for a sparse precision matrix selection in a Gaussian distribution [38, 73]. Since the (i, j)-th element of a precision matrix expresses a relation between the i-th and j-th elements of probability variables, a sparse precision matrix plays a critical role in the graphical modeling [20]. Ruan, Yuan and Zou [53] proposed the EM algorithm for Gaussian mixtures with the L_1 regularization, and succeeded in estimating parameters with sparse precision matrices.

In this chapter, we first define a maximum likelihood estimation problem, whose objective function consists of not only a log-likelihood function but also some proper lower semicontinuous quasiconvex functions. If we exploit the L_1 regularization term and/or indicator functions of constraint sets as the additional proper lower semicontinuous quasiconvex functions, we can estimate parameters with the regularization and/or the constraint. Especially, parameter esti-

mations with lower constraints on mixture coefficients are one of contributions in this thesis. Thanks to such constraints, we can obtain some theoretically and practically nice properties. Meanwhile, an estimation problem considered in this chapter is more general than that in the existing papers, such as [53, 67].

The estimation problem in this chapter might not be solved by the usual EM algorithm. Then, we consider a block coordinate descent (BCD) method. At each iteration of the BCD method, the objective function is minimized among a few parameters while all the other parameters are fixed.

Since the log-likelihood function is not separable for each parameter in mixture distributions, we first construct a separable problem related to the original one. Then, we apply a BCD method to the separable problem, where the block corresponds to a set of parameters in a single distribution. As seen in Section 2.5, Tseng [62] showed that a BCD method for a nondifferentiable minimization problem has the global convergence property under some reasonable conditions. Using the results in Section 2.5, we prove the global convergence of the proposed BCD method when we add certain lower bound constraints on mixture coefficients. In addition, we discuss efficient implementations for some concrete problems, such as the maximum likelihood estimation with box constraints on mixture coefficients.

The present chapter is organized as follows. In Section 5.2, we introduce maximum likelihood estimation problems for mixture distributions. In particular, we present a general class of maximum likelihood estimation problems for mixture distributions that has a log-likelihood function and/or some proper lower semicontinuous quasiconvex functions, such as the L_1 regularization and/or indicator functions of constraint sets. In Section 5.3, we present a BCD method for the proposed maximum likelihood estimation problem, and discuss its global convergence. In Section 5.4, we discuss how to solve subproblems in the proposed BCD method for some special cases. In Section 5.5, we report some numerical results for the maximum likelihood estimation with some additional constraints. Finally, we make some concluding remarks in Section 5.6.

5.2 Maximum likelihood estimation for mixture distributions

In this section, we introduce maximum likelihood estimation problems for mixture distributions.

Assume that probability variables $x \in \mathbf{R}^d$ obey a probability distribution $p(x)$. If $p(x)$ is expressed as a weighted linear combination of distributions $p_i(x|\theta_i)$ $(i = 1, \ldots, m)$:

$$p(x) := \sum_{i=1}^{m} \alpha_i p_i(x|\theta_i),$$

then $p(x)$ is called a *mixture distribution* which has parameters α_i, θ_i $(i = 1, \ldots, m)$, where $p_i(x|\theta_i)$ $(i = 1, \ldots, m)$ are called *mixture components*, α_i $(i = 1, \ldots, m)$ are called *mixture coefficients* satisfying

$$\sum_{i=1}^{m} \alpha_i = 1, \quad 0 \leq \alpha_i, \quad i = 1, \ldots, m,$$

and θ_i $(i = 1, \ldots, m)$ are parameters of the distributions $p_i(x|\theta_i)$ $(i = 1, \ldots, m)$, respectively. Let Θ_i $(i = 1, \ldots, m)$ be sets of the parameters θ_i $(i = 1, \ldots, m)$, respectively. Throughout

this chapter, we suppose that the sets Θ_i $(i = 1, \ldots, m)$ are closed convex. Moreover, let $\mathcal{V}_i$ $(i = 1, \ldots, m)$ be inner product spaces such that $\Theta_i \subset \mathcal{V}_i$ $(i = 1, \ldots, m)$, respectively. Note that $\theta_i \in \Theta_i \subset \mathcal{V}_i$ $(i = 1, \ldots, m)$. Then, we define Θ and $\mathcal{V}$ as follows:

$$\Theta := \Theta_1 \times \cdots \times \Theta_m, \quad \mathcal{V} := \mathcal{V}_1 \times \cdots \times \mathcal{V}_m.$$

Note that since Θ_i $(i = 1, \ldots, m)$ are closed convex sets, so is Θ. Note also that since $\mathcal{V}_i$ $(i = 1, \ldots, m)$ are inner product spaces, so is $\mathcal{V}$.

Suppose that we have observational data $X := [x_1, \ldots, x_n] \in \mathbf{R}^{d \times n}$. Then, we wish to model the data X using the mixture distribution $p(x)$ with the parameters α_i, θ_i $(i = 1, \ldots, m)$. To this end, we consider an estimation of the parameters α_i, θ_i $(i = 1, \ldots, m)$. In the remainder of this chapter, we exploit the following notation in order to specify that the parameters of $p(x)$ are α_i, θ_i $(i = 1, \ldots, m)$:

$$p(x|\alpha, \theta) := \sum_{i=1}^{m} \alpha_i p_i(x|\theta_i), \tag{5.2.1}$$

where $\alpha := [\alpha_1, \ldots, \alpha_m]^\top \in \mathbf{R}^m$ and $\theta := [\theta_1, \ldots, \theta_m] \in \Theta$.

A joint probability for the observational data X is given by

$$P(X|\alpha, \theta) := \prod_{k=1}^{n} p(x_k|\alpha, \theta).$$

We call $P(X|\alpha, \theta)$ a *likelihood*. Moreover, a maximizer (α^*, θ^*) of the likelihood $P(X|\alpha, \theta)$ is called a *maximum likelihood estimator*. In what follows, an estimation of parameters means that we obtain the maximum likelihood estimator (α^*, θ^*). Since a maximization of a likelihood is difficult in general, we usually maximize the following log-likelihood function:

$$L(\alpha, \theta) := \log P(X|\alpha, \theta) = \sum_{k=1}^{n} \log \left(\sum_{i=1}^{m} \alpha_i p_i(x_k|\theta_i) \right).$$

We sometimes want to maximize the log-likelihood function L with regularizations and/or constraints on some parameters in (α, θ). Thus, we consider the following maximization problem:

$$\begin{array}{ll} \text{maximize} & L(\alpha, \theta) - f(\alpha, \theta), \\ \text{subject to} & \alpha \in \Omega_l,\ \theta_i \in \Theta_i,\ i = 1, \ldots, m, \end{array} \tag{5.2.2}$$

where the function $f : \mathbf{R}^m \times \mathcal{V} \to \mathbf{R}$ is proper lower semicontinuous quasiconvex, and the set Ω_l is defined by

$$\Omega_l := \left\{ \alpha \in \mathbf{R}^m \;\middle|\; \sum_{i=1}^{m} \alpha_i = 1,\ l_i \leq \alpha_i,\ i = 1, \ldots, m \right\}.$$

where $l = [l_1, \ldots, l_m]^\top \in \mathbf{R}^m$ is a constant vector such that $l_i \in [0, 1]$ $(i = 1, \ldots, m)$ and $\sum_{i=1}^{m} l_i < 1$. In what follows, we call problem (5.2.2) a maximum likelihood estimation problem. Note that the function f is regarded as a generalization of the L_1 regularization and indicator functions of constraint sets. Note also that $l = 0$ in [26, 51, 53, 67]. To the author's best

knowledge, this is the first time to consider the lower bounds $l_i \leq \alpha_i$ $(i = 1, \ldots, m)$ in the maximum likelihood estimation for mixture distributions. As seen in Sections 5.3 and 5.5, the lower bounds with positive constants l_i $(i = 1, \ldots, m)$ bring in both theoretically and practically nice effects.

We now give two concrete cases of problem (5.2.2).

Example 1. The maximum likelihood estimation with constraints on mixture coefficients

We discuss the maximum likelihood estimation with constraints on mixture coefficients. We assume that the mixture coefficients satisfy $\alpha_i \in [l_i, u_i]$ $(i = 1, \ldots, m)$, where $l_i, u_i \in (0, 1]$ $(i = 1, \ldots, m)$, $\sum_{i=1}^m l_i < 1$ and $\sum_{i=1}^m u_i \geq 1$. Let $\mathcal{U}_u := \{\ \alpha \in \mathbf{R}^m \mid \alpha_i \leq u_i,\ i = 1, \ldots, m\ \}$. Then, we may define the function f of problem (5.2.2) as

$$f(\alpha, \theta) := \begin{cases} 0 & \text{if } \alpha \in \mathcal{U}_u \cap \Omega_l, \\ +\infty & \text{otherwise.} \end{cases}$$

As described above, the constraints $l_i \leq \alpha_i$ $(i = 1, \ldots, m)$ play a critical role in the theoretical and practical aspects. In the theoretical aspect, these constraints enable us to show the global convergence of a BCD method proposed in Section 5.3. In the practical aspect, these constraints bring in some valid parameter estimations when the amount of the observational data is small.

Example 2. The maximum likelihood estimation with the L_1 regularization for Gaussian mixtures

Suppose that the mixture components $p_i(x|\theta_i)$ $(i = 1, \ldots, m)$ in (5.2.1) are Gaussian:

$$\mathcal{N}(x|\mu_i, \Lambda_i^{-1}) := \frac{\sqrt{\det \Lambda_i}}{(2\pi)^{d/2}} \exp\left[-\frac{1}{2}(x - \mu_i)^\top \Lambda_i (x - \mu_i)\right], \quad i = 1, \ldots, m,$$

where $\mu_i \in \mathbf{R}^d$ and $\Lambda_i \in \mathbf{S}^d$ denote a mean vector and a precision matrix. Note that a precision matrix is the inverse of a covariance matrix. Then, $\theta_i = [\mu_i, \Lambda_i]$ $(i = 1, \ldots, m)$. Several researchers, such as Friedman, Hastie and Tibshirani [20] and Lu [39], proposed maximum likelihood estimation problems with the L_1 regularization. We apply such ideas to the maximum likelihood estimation for mixture distributions. Then, we may consider the following problem:

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_{k=1}^n \log\left(\sum_{i=1}^m \alpha_i \mathcal{N}(x_k|\mu_i, \Lambda_i^{-1})\right) - \sum_{i=1}^m \rho_i \|\Lambda_i\|_1, \\ \text{subject to} & \alpha \in \Omega_0,\ \underline{\lambda}_i I \preceq \Lambda_i \preceq \overline{\lambda}_i I,\ i = 1, \ldots, m, \end{array} \tag{5.2.3}$$

where ρ_i, $\underline{\lambda}_i$, $\overline{\lambda}_i$ $(i = 1 \ldots, m)$ are constants such that $\rho_i \in [0, \infty)$, $\underline{\lambda}_i \in [0, \infty)$, $\overline{\lambda}_i \in (0, \infty]$, $\underline{\lambda}_i < \overline{\lambda}_i$ $(i = 1, \ldots, m)$. Note that we allow $\overline{\lambda}_i$ $(i = 1, \ldots, m)$ to be $+\infty$. Note also that $\underline{\lambda}_i = 0$, $\overline{\lambda}_i = \infty$ $(i = 1, \ldots, m)$ in [20, 39].

Thanks to the L_1 regularization term $\sum_{i=1}^m \rho_i \|\Lambda_i\|_1$, we can obtain a maximum likelihood estimator with sparse precision matrices. The sparse precision matrix plays an important role

in the graphical modeling. For details, see [20, 38, 39, 53]. Let $\mathcal{S}_i := \{\ S \in \mathbf{S}^d \mid \underline{\lambda}_i I \preceq S \preceq \overline{\lambda}_i I\ \}$. Then, problem (5.2.3) is written as (5.2.2) with

$$f(\alpha,\theta) := \sum_{i=1}^{m} f_i(\Lambda_i), \quad f_i(\Lambda_i) := \begin{cases} \rho_i\|\Lambda_i\|_1 & \text{if } \Lambda_i \in \mathcal{S}_i, \\ +\infty & \text{otherwise,} \end{cases} \quad \Theta_i := \mathbf{R}^d \times \mathcal{S}_i, \quad i = 1,\dots,m.$$

5.3 Block coordinate descent method and its global convergence

In this section, we present a BCD method solving maximum likelihood estimation problem (5.2.2). To this end, we first construct a separable problem suitable to the proposed BCD method. Next, we give conditions under which the proposed BCD method has the global convergence property.

If a BCD method is directly applied to problem (5.2.2), then it may solve the following subproblems at each step:

$$\begin{aligned}
\alpha^{t+1} &\in \operatorname*{argmax}_{\alpha\in\Omega_l} \left\{L(\alpha,\theta_1^t,\dots,\theta_m^t) - f(\alpha,\theta_1^t,\dots,\theta_m^t)\right\}, \\
\theta_1^{t+1} &\in \operatorname*{argmax}_{\theta_1\in\Theta_1} \left\{L(\alpha^{t+1},\theta_1,\theta_2^t,\dots,\theta_m^t) - f(\alpha^{t+1},\theta_1,\theta_2^t,\dots,\theta_m^t)\right\}, \\
\theta_2^{t+1} &\in \operatorname*{argmax}_{\theta_2\in\Theta_2} \left\{L(\alpha^{t+1},\theta_1^{t+1},\theta_2,\theta_3^t,\dots,\theta_m^t) - f(\alpha^{t+1},\theta_1^{t+1},\theta_2,\theta_3^t,\dots,\theta_m^t)\right\}, \\
&\vdots \\
\theta_m^{t+1} &\in \operatorname*{argmax}_{\theta_m\in\Theta_m} \left\{L(\alpha^{t+1},\theta_1^{t+1},\dots,\theta_{m-1}^{t+1},\theta_m) - f(\alpha^{t+1},\theta_1^{t+1},\dots,\theta_{m-1}^{t+1},\theta_m)\right\},
\end{aligned}$$

where the superscript t denotes the t-th iteration. We see that the subproblems cannot be solved in parallel because the log-likelihood function L has a weighted linear combination of the mixture components $p_i(x_k|\theta_i)$ $(i = 1,\dots,m)$ in the antilogarithm part. Thus, we construct a separable problem associated with (5.2.2) in order to solve subproblems in parallel.

We assume that the function f is separable with respect to $\alpha,\theta_1,\dots,\theta_m$, that is,

$$f(\alpha,\theta) = f_0(\alpha) + \sum_{i=1}^{m} f_i(\theta_i), \tag{5.3.1}$$

where f_0 is a proper lower semicontinuous quasiconvex function for adding some constraints on mixture coefficients α_i $(i = 1,\dots,m)$, and f_i $(i = 1,\dots,m)$ are also proper lower semicontinuous quasiconvex functions for adding some constraints on parameters θ_i $(i = 1,\dots,m)$, respectively. Then, we consider the following minimization problem instead of problem (5.2.2):

$$\begin{aligned}
&\text{minimize} && D(W,\alpha,\theta) + f_0(\alpha) + \sum_{i=1}^{m} f_i(\theta_i), \\
&\text{subject to} && W \in M,\ \alpha \in \Omega_l,\ \theta_i \in \Theta_i,\ i = 1,\dots,m,
\end{aligned} \tag{5.3.2}$$

where decision variables of problem (5.3.2) are α,θ and W, the function $D : \mathbf{R}_+^{m\times n}\times\mathbf{R}_+^m\times\Theta \to \mathbf{R}$

and the set M are defined by

$$D(W,\alpha,\theta) := \sum_{i=1}^{m}\sum_{k=1}^{n} W_{ik}\{\log W_{ik} - \log\alpha_i - \log p_i(x_k|\theta_i)\}, \tag{5.3.3}$$

$$M := \left\{ W \in \mathbf{R}_+^{m\times n} \;\middle|\; \sum_{i=1}^{m} W_{ik} = 1,\ k = 1,\dots,n \right\}, \tag{5.3.4}$$

respectively. If we apply a BCD method to problem (5.3.2), then the objective function of (5.3.2) is separable for α and θ_i $(i = 1,\dots,m)$ when W is fixed. The details are discussed later.

Now we mention that a BCD method can find a solution of (5.2.2) if it solves problem (5.3.2). For each $(\alpha,\theta) \in \mathbf{R}_+^m \times \Theta$, we consider the following problem:

$$\begin{aligned} &\text{minimize} && D(W,\alpha,\theta),\\ &\text{subject to} && W \in M. \end{aligned} \tag{5.3.5}$$

Then, we define a function $g : \mathbf{R}_+^m \times \Theta \to \mathbf{R}$ as

$$g(\alpha,\theta) := \min_{W\in M} D(W,\alpha,\theta).$$

For each $(\alpha,\theta) \in \mathbf{R}_+^m \times \Theta$, the function $D(\cdot,\alpha,\theta)$ is strictly convex on the compact set M defined by (5.3.4), and hence Proposition 2.2.6 implies that problem (5.3.5) has a unique optimum. In what follows, we denote the unique optimum by $\mathcal{W}(\alpha,\theta)$. Note that $g(\alpha,\theta) = D(\mathcal{W}(\alpha,\theta),\alpha,\theta)$ for any $(\alpha,\theta) \in \mathbf{R}_+^m \times \Theta$. The next lemma shows that $g(\alpha,\theta) = -L(\alpha,\theta)$, i.e., problem (5.2.2) is equivalent to

$$\begin{aligned} &\text{minimize} && g(\alpha,\theta) + f_0(\alpha) + \sum_{i=1}^{m} f_i(\theta_i),\\ &\text{subject to} && \alpha \in \Omega_l,\ \theta_i \in \Theta_i,\ i = 1,\dots,m. \end{aligned} \tag{5.3.6}$$

Although the equivalence is implicitly given in [26], we provide its proof for the completeness of this thesis.

Lemma 5.3.1. *For each $(\alpha,\theta) \in \mathbf{R}_+^m \times \Theta$,*

$$g(\alpha,\theta) = -L(\alpha,\theta), \quad \mathcal{W}_{ik}(\alpha,\theta) = \frac{\alpha_i p_i(x_k|\theta_i)}{p(x_k|\alpha,\theta)}, \quad i = 1,\dots,m,\ k = 1,\dots,n,$$

where $\mathcal{W}_{ik}(\alpha,\theta)$ denotes the (i,k)-th element of $\mathcal{W}(\alpha,\theta)$.

Proof. Let $(\alpha,\theta) \in \mathbf{R}_+^m \times \Theta$. We denote $\mathcal{W}(\alpha,\theta)$ by W^* for simplicity. The KKT conditions for problem (5.3.5) with (α,θ) are written as

$$\sum_{i=1}^{m} W_{ik}^* = 1, \quad \log W_{ik}^* + 1 - \log\alpha_i p_i(x_k|\theta_i) - u_k^* = 0, \quad i = 1,\dots,m,\ k = 1,\dots,n,$$

where u_k^* $(k = 1,\dots,n)$ are Lagrange multipliers for $\sum_{i=1}^m W_{ik}^* = 1$ $(k = 1,\dots,n)$, respectively. Then, we obtain

$$W_{ik}^* = \alpha_i p_i(x_k|\theta_i)\exp(u_k^* - 1), \quad i = 1,\dots,m,\ k = 1,\dots,n. \tag{5.3.7}$$

It further follows from $\sum_{i=1}^m W_{ik}^* = 1$ $(k = 1, \ldots, n)$ that

$$1 = \sum_{i=1}^m W_{ik}^* = \exp(u_k^* - 1) \sum_{i=1}^m \alpha_i p_i(x_k|\theta_i) = \exp(u_k^* - 1) p(x_k|\alpha, \theta), \quad k = 1, \ldots, n,$$

and hence

$$\exp(u_k^* - 1) = \frac{1}{p(x_k|\alpha, \theta)}, \quad k = 1, \ldots, n. \tag{5.3.8}$$

Then, (5.3.7) and (5.3.8) yield that

$$W_{ik}^* = \frac{\alpha_i p_i(x_k|\theta_i)}{p(x_k|\alpha, \theta)}, \quad i = 1, \ldots, m, \ k = 1, \ldots, n.$$

Moreover, we have

$$\begin{aligned}
g(\alpha, \theta) &= D(W^*, \alpha, \theta) \\
&= \sum_{i=1}^m \sum_{k=1}^n W_{ik}^* \left\{ \log \frac{\alpha_i p_i(x_k|\theta_i)}{p(x_k|\alpha, \theta)} - \log \alpha_i - \log p_i(x_k|\theta_i) \right\} \\
&= \sum_{i=1}^m \sum_{k=1}^n W_{ik}^* \left\{ \log \alpha_i p_i(x_k|\theta_i) - \log p(x_k|\alpha, \theta) - \log \alpha_i p_i(x_k|\theta_i) \right\} \\
&= -\sum_{k=1}^n \left(\sum_{i=1}^m W_{ik}^* \right) \log p(x_k|\alpha, \theta) \\
&= -L(\alpha, \theta),
\end{aligned}$$

where the last equality follows from $\sum_{i=1}^m W_{ik}^* = 1$ $(k = 1, \ldots, n)$. □

From Lemma 5.3.1, problem (5.2.2) is equivalent to problem (5.3.6). On the other hand, if $(\overline{W}, \overline{\alpha}, \overline{\theta})$ is a global optimum of problem (5.3.2), then $(\overline{\alpha}, \overline{\theta})$ is that of problem (5.3.6). Therefore, we can obtain a global optimum of problem (5.2.2) by solving problem (5.3.2). Moreover, under some assumptions, we can show that if $(\overline{W}, \overline{\alpha}, \overline{\theta})$ is a stationary point of (5.3.2), then $(\overline{\alpha}, \overline{\theta})$ is that of problem (5.2.2). In order to prove it, we first provide gradients of g with respect to α and θ, respectively.

Lemma 5.3.2. *Suppose that the following statements hold:*

(i) *The function D is differentiable at $(\mathcal{W}(\overline{\alpha}, \overline{\theta}), \overline{\alpha}, \overline{\theta})$;*

(ii) *For each $i \in \{1, \ldots, m\}$ and $x_k \in \{x_1, \ldots, x_n\}$, the function $p_i(x_k|\cdot)$ is continuously differentiable on $\mathrm{int}\Theta_i$ and $0 < p_i(x_k|\theta_i)$ for all $\theta_i \in \Theta_i$.*

Then the function g is differentiable at $(\overline{\alpha}, \overline{\theta})$, and its gradients with respect to α and θ are

$$\nabla_\alpha g(\overline{\alpha}, \overline{\theta}) = \nabla_\alpha D(\mathcal{W}(\overline{\alpha}, \overline{\theta}), \overline{\alpha}, \overline{\theta}), \qquad \nabla_\theta g(\overline{\alpha}, \overline{\theta}) = \nabla_\theta D(\mathcal{W}(\overline{\alpha}, \overline{\theta}), \overline{\alpha}, \overline{\theta}).$$

Proof. Let $\overline{u} := (\overline{\alpha}, \overline{\theta})$ and $U := \mathbf{R}_+^m \times \Theta$. Moreover, let $\overline{W} := \mathcal{W}(\overline{u})$. Assumption (i) implies that $(\overline{W}, \overline{u}) \in \mathbf{R}_{++}^{m \times n} \times \mathrm{int}U$, that is,

$$\overline{W} \in \mathbf{R}_{++}^{m \times n}, \quad \overline{u} \in \mathrm{int}U. \tag{5.3.9}$$

Let $i \in \{1, \ldots, m\}$ and $k \in \{1, \ldots, n\}$. Lemma 5.3.1, assumption (ii) and (5.3.9) imply that $\mathcal{W}_{ik} : U \to \mathbf{R}$ is continuous at $\overline{u} \in \mathrm{int}U$, that is,

$$\forall \varepsilon > 0, \quad \exists r_{ik} > 0 \quad \text{such that} \quad |\mathcal{W}_{ik}(\overline{u} + d) - \mathcal{W}_{ik}(\overline{u})| < \varepsilon, \quad \forall d \in B(0, r_{ik}). \tag{5.3.10}$$

Now, we obtain that $0 < \overline{W}_{ik} = \mathcal{W}_{ik}(\overline{u})$ from (5.3.9), and hence if we choose ε in (5.3.10) as $0 < \varepsilon < \mathcal{W}_{ik}(\overline{u})$, then

$$0 < \mathcal{W}_{ik}(\overline{u}) - \varepsilon < \mathcal{W}_{ik}(\overline{u} + d), \quad \forall d \in B(0, r_{ik}). \tag{5.3.11}$$

Meanwhile, we have from (5.3.9) that there exists $r_0 > 0$ such that $B(\overline{u}, r_0) \subset \mathrm{int}U$, that is,

$$\overline{u} + d \in \mathrm{int}U, \quad \forall d \in B(0, r_0). \tag{5.3.12}$$

Then, let r be a positive number such that $r < r_0$ and $r < r_{ik}$ $(i = 1, \ldots, m,\ k = 1, \ldots, n)$. Moreover, let $d \in B(0, r)$ be arbitrary, where $d \neq 0$. Now, it is clear that $d \in B(0, r) \subset B(0, r_{ik})$ $(i = 1, \ldots, m,\ k = 1, \ldots, n)$ and $d \in B(0, r) \subset B(0, r_0)$. Thus, it follows from (5.3.11) and (5.3.12) that

$$\mathcal{W}(\overline{u} + d) \in \mathbf{R}^{m \times n}_{++}, \quad \overline{u} + d \in \mathrm{int}U. \tag{5.3.13}$$

Note that $g(\overline{u}) = D(\overline{W}, \overline{u})$ because $\overline{W}$ is a global optimum of problem (5.3.5) with $(\overline{\alpha}, \overline{\theta})$. Then, we have from the definition of g and (5.3.13) that

$$g(\overline{u} + d) - g(\overline{u}) \leq D(\overline{W}, \overline{u} + d) - D(\overline{W}, \overline{u}) = \langle \nabla_u D(\overline{W}, \overline{u}), d \rangle + \|d\| \varphi(d),$$

where $\varphi : U \to \mathbf{R}$ is a certain function such that $\varphi(d) \to 0$ if $d \to 0$. Furthermore, it follows from $d \neq 0$ that

$$\frac{g(\overline{u} + d) - g(\overline{u}) - \langle \nabla_u D(\overline{W}, \overline{u}), d \rangle}{\|d\|} \leq \varphi(d),$$

and hence, by $\|d\| \to 0$,

$$\limsup_{d \to 0} \frac{g(\overline{u} + d) - g(\overline{u}) - \langle \nabla_u D(\overline{W}, \overline{u}), d \rangle}{\|d\|} \leq 0. \tag{5.3.14}$$

Let $u_d := \overline{u} + d$, and let $W_d := \mathcal{W}(u_d)$. Then, the definition of g and (5.3.13) yield that

$$D(W_d, u_d) - D(W_d, \overline{u}) \leq D(W_d, u_d) - D(\overline{W}, \overline{u}) = g(\overline{u} + d) - g(\overline{u}). \tag{5.3.15}$$

On the other hand, we have by the definition of D, assumption (ii) and (5.3.9) that there exists $\nabla_u D(\widetilde{W}, \widetilde{u})$ for all $(\widetilde{W}, \widetilde{u}) \in \mathbf{R}^{m \times n}_{++} \times \mathrm{int}U$, and

$$\nabla_u D(\widetilde{W}, \widetilde{U}) \to \nabla_u D(\overline{W}, \overline{u}) \quad (\widetilde{W} \to \overline{W},\ \widetilde{u} \to \overline{u}). \tag{5.3.16}$$

Thus, there exists $\nabla_u D(W_d, \overline{u} + \lambda d)$ for all $\lambda \in [0, 1]$ because $W_d \in \mathbf{R}^{m \times n}_{++}$ and $\overline{u} + \lambda d \in \mathrm{int}U$ for all $\lambda \in [0, 1]$ by (5.3.9) and (5.3.13). Then, Theorem 2.2.1 implies that there exists $t \in (0, 1)$ such that

$$D(W_d, u_d) - D(W_d, \overline{u}) = \langle \nabla_u D(W_d, \overline{u} + td), d \rangle. \tag{5.3.17}$$

We have by the Cauchy-Schwarz inequality, (5.3.15) and (5.3.17) that

$$-\|d\|\|\nabla_u D(W_d, \overline{u}+td) - \nabla_u D(\overline{W}, \overline{u})\| \leq g(\overline{u}+d) - g(\overline{u}) - \langle \nabla_u D(\overline{W}, \overline{u}), d\rangle.$$

It then follows from $d \neq 0$ that

$$-\|\nabla_u D(W_d, \overline{u}+td) - \nabla_u D(\overline{W}, \overline{u})\| \leq \frac{g(\overline{u}+d) - g(\overline{u}) - \langle \nabla_u D(\overline{W}, \overline{u}), d\rangle}{\|d\|}. \tag{5.3.18}$$

Meanwhile, it is clear that $u_d \to \overline{u}$ and $\overline{u}+td \to \overline{u}$ when $d \to 0$. Moreover, (5.3.10) means that $\mathcal{W} : U \to \mathbf{R}^{m\times n}$ is continuous at $\overline{u} \in \mathrm{int}U$, that is $W_d = \mathcal{W}(u_d) \to \mathcal{W}(\overline{u}) = \overline{W}$ when $d \to 0$. Then, (5.3.16) yields that $\nabla_u D(W_d, \overline{u}+td) \to \nabla_u D(\overline{W}, \overline{u})$ when $d \to 0$. It then follows from (5.3.18) that

$$0 \leq \liminf_{d\to 0} \frac{g(\overline{u}+d) - g(\overline{u}) - \langle \nabla_u D(\overline{W}, \overline{u}), d\rangle}{\|d\|}. \tag{5.3.19}$$

Combining (5.3.14) and (5.3.19),

$$\lim_{d\to 0} \frac{g(\overline{u}+d) - g(\overline{u}) - \langle \nabla_u D(\overline{W}, \overline{u}), d\rangle}{\|d\|} = 0,$$

that is, $\nabla g(\overline{u}) = \nabla_u D(\overline{W}, \overline{u})$. Now, we see that

$$\nabla g(\overline{u}) = \begin{bmatrix} \nabla_\alpha g(\overline{\alpha}, \overline{\theta}) \\ \nabla_\theta g(\overline{\alpha}, \overline{\theta}) \end{bmatrix}, \quad \nabla_u D(\overline{W}, \overline{u}) = \begin{bmatrix} \nabla_\alpha D(\overline{W}, \overline{\alpha}, \overline{\theta}) \\ \nabla_\theta D(\overline{W}, \overline{\alpha}, \overline{\theta}) \end{bmatrix}.$$

Therefore, $\nabla_\alpha g(\overline{\alpha}, \overline{\theta}) = \nabla_\alpha D(\overline{W}, \overline{\alpha}, \overline{\theta})$ and $\nabla_\theta g(\overline{\alpha}, \overline{\theta}) = \nabla_\theta D(\overline{W}, \overline{\alpha}, \overline{\theta})$. □

Next, under some assumptions, we show that if $(\overline{W}, \overline{\alpha}, \overline{\theta}) \in M \times \Omega_l \times \Theta$ is a stationary point of problem (5.3.2), then $(\overline{W}, \overline{\alpha}, \overline{\theta})$ is that of problem (5.2.2).

Theorem 5.3.1. *Let $(\overline{W}, \overline{\alpha}, \overline{\theta})$ be a stationary point of problem (5.3.2). Suppose that the following statements hold:*

(i) *The function D is differentiable at $(\overline{W}, \overline{\alpha}, \overline{\theta})$;*

(ii) *For each $i \in \{1, \ldots, m\}$ and $x_k \in \{x_1, \ldots, x_n\}$, the function $p_i(x_k|\cdot)$ is continuously differentiable on $\mathrm{int}\Theta_i$ and $0 < p_i(x_k|\theta_i)$ for all $\theta_i \in \Theta_i$.*

Then $(\overline{W}, \overline{\alpha}, \overline{\theta})$ is a stationary point of problem (5.2.2).

Proof. Let $(W, \alpha, \theta) \in M \times \Omega_l \times \Theta$ be arbitrary. Now, $(\overline{W}, \overline{\alpha}, \overline{\theta}) \in M \times \Omega_l \times \Theta$ is a stationary point of problem (5.3.2), and hence

$$0 \leq \langle \nabla_W D(\overline{W}, \overline{\alpha}, \overline{\theta}), W - \overline{W}\rangle + \langle \nabla_\alpha D(\overline{W}, \overline{\alpha}, \overline{\theta}), \alpha - \overline{\alpha}\rangle + \langle \nabla_\theta D(\overline{W}, \overline{\alpha}, \overline{\theta}), \theta - \overline{\theta}\rangle$$
$$+ f_0'(\overline{\alpha}; \alpha - \overline{\alpha}) + \sum_{i=1}^m f_i'(\overline{\theta}_i; \theta_i - \overline{\theta}_i). \tag{5.3.20}$$

Note that $(\alpha, \theta) \in \Omega_l \times \Theta$ is arbitrary. Substituting $(\alpha, \theta) = (\overline{\alpha}, \overline{\theta})$ into (5.3.20), for all $W \in M$,

$$0 \leq \langle \nabla_W D(\overline{W}, \overline{\alpha}, \overline{\theta}), W - \overline{W}\rangle.$$

Thus, $\overline{W}$ is a stationary point of problem (5.3.5). Moreover, we have from Proposition 2.2.6 that $\overline{W} = \mathcal{W}(\overline{\alpha}, \overline{\theta})$. It then follows from Lemma 5.3.2 and assumptions (i) and (ii) that

$$\nabla_\alpha g(\overline{\alpha}, \overline{\theta}) = \nabla_\alpha D(\overline{W}, \overline{\alpha}, \overline{\theta}), \qquad \nabla_\theta g(\overline{\alpha}, \overline{\theta}) = \nabla_\theta D(\overline{W}, \overline{\alpha}, \overline{\theta}). \tag{5.3.21}$$

On the other hand, we substitute $W = \overline{W}$ into (5.3.20) because $W \in M$ is arbitrary. As the result, for all $(\alpha, \theta) \in \Omega_l \times \Theta$,

$$0 \leq \langle \nabla_\alpha D(\overline{W}, \overline{\alpha}, \overline{\theta}), \alpha - \overline{\alpha} \rangle + \langle \nabla_\theta D(\overline{W}, \overline{\alpha}, \overline{\theta}), \theta - \overline{\theta} \rangle + f_0'(\overline{\alpha}; \alpha - \overline{\alpha}) + \sum_{i=1}^m f_i'(\overline{\theta}_i; \theta_i - \overline{\theta}_i). \tag{5.3.22}$$

By (5.3.21) and (5.3.22), for all $(\alpha, \theta) \in \Omega_l \times \Theta$,

$$0 \leq \langle \nabla_\alpha g(\overline{\alpha}, \overline{\theta}), \alpha - \overline{\alpha} \rangle + \langle \nabla_\theta g(\overline{\alpha}, \overline{\theta}), \theta - \overline{\theta} \rangle + f_0'(\overline{\alpha}; \alpha - \overline{\alpha}) + \sum_{i=1}^m f_i'(\overline{\theta}_i; \theta_i - \overline{\theta}_i).$$

Thus, $(\overline{\alpha}, \overline{\theta})$ is a stationary point of problem (5.3.6). It then follows from Lemma 5.3.1 that $(\overline{\alpha}, \overline{\theta})$ is a stationary point of problem (5.2.2). □

We now apply a BCD method to problem (5.3.2). Let (α^t, θ^t) be given. The BCD method first solves problem (5.3.5) with (α^t, θ^t), that is, $W^t := \mathcal{W}(\alpha^t, \theta^t)$. From Lemma 5.3.1, the solution W^t is given by

$$W_{ik}^t = \frac{\alpha_i^t p_i(x_k|\theta_i^t)}{p(x_k|\alpha^t, \theta^t)}, \quad i = 1, \ldots, m, \ k = 1, \ldots, n. \tag{5.3.23}$$

Next, it solves the following subproblems with respect to α and θ_i $(i = 1, \ldots, m)$ independently:

$$\begin{aligned} &\text{minimize} && -\sum_{i=1}^m \sum_{k=1}^n W_{ik}^t \log \alpha_i + f_0(\alpha), \\ &\text{subject to} && \alpha \in \Omega_l, \end{aligned} \tag{5.3.24}$$

$$\begin{aligned} &\text{minimize} && -\sum_{k=1}^n W_{ik}^t \log p_i(x_k|\theta_i) + f_i(\theta_i), \\ &\text{subject to} && \theta_i \in \Theta_i. \end{aligned} \tag{5.3.25}$$

Note that the functions f_0 and f_i $(i = 1, \ldots, m)$ are given by (5.3.1). Summing up the above discussion, the BCD method is described as follows.

Algorithm 5.3.1.

Step 0. *Choose an initial point* $(\alpha^0, \theta^0) \in \mathbf{R}^m \times \Theta$*, and set* $t := 0$.

Step 1. *Calculate* W^t *by (5.3.23).*

Step 2. *Obtain a solution* α^{t+1} *to problem (5.3.24).*

Step 3. *For each* $i \in \{1, \ldots, m\}$*, obtain a solution* θ_i^{t+1} *to problem (5.3.25).*

Step 4. *If an appropriate termination criterion is satisfied, then stop. Otherwise, set* $t := t + 1$ *and go to Step 1.*

Next, we discuss the global convergence of Algorithm 5.3.1. First, we give some assumptions for problem (5.3.2). These assumptions are sufficient conditions under which Algorithm 5.3.1 has the global convergence property.

Assumption 5.3.1.

(A1) *The constant vector* $l \in \mathbf{R}^m$ *satisfies that* $l = \underline{\alpha}$*, where* $\underline{\alpha} \in \mathbf{R}^m$ *is a certain vector such that* $0 < \underline{\alpha}_i$ $(i = 1, \ldots, m)$ *and* $\sum_{i=1}^m \underline{\alpha}_i < 1$*.*

(A2) *The functions* f_0 *and* f_i $(i = 1, \ldots, m)$ *are proper lower semicontinuous quasiconvex on* $\mathbf{R}^m$ *and* $\mathcal{V}_i$ $(i = 1, \ldots, m)$*, respectively.*

(A3) *For each* $i \in \{1, \ldots, m\}$ *and* $x_k \in \{x_1, \ldots, x_n\}$*, the function* $-\log p_i(x_k|\cdot)$ *is quasiconvex and hemivariate on* Θ_i*. Moreover, the function* $p_i(x_k|\cdot)$ *is continuously differentiable on* $\mathrm{int}\Theta_i$ *and* $0 < p_i(x_k|\theta_i)$ *for all* $\theta_i \in \Theta_i$*.*

Next, we provide a theorem that guarantees the global convergence of Algorithm 5.3.1.

Theorem 5.3.2. *Suppose that Assumption 5.3.1 (A1)–(A3) hold. Suppose also that Algorithm 5.3.1 generates an infinite sequence* $\{(W^t, \alpha^{t+1}, \theta^{t+1})\}$ *that has an accumulation point* $(\overline{W}, \overline{\alpha}, \overline{\theta})$*. If the function* D *is differentiable at* $(\overline{W}, \overline{\alpha}, \overline{\theta})$*, then* $(\overline{W}, \overline{\alpha}, \overline{\theta})$ *is a stationary point of problem (5.3.2).*

Proof. By using $\underline{\alpha} \in \mathbf{R}^m$ in Assumption 5.3.1 (A1), we consider the following problem:

$$\text{minimize} \quad F(W, \alpha, \theta), \tag{5.3.26}$$

where

$$F(W, \alpha, \theta) := \overline{D}(W, \alpha, \theta) + \delta_M(W) + \overline{f}_0(\alpha) + \sum_{i=1}^m f_i(\theta_i),$$

$$\overline{D}(W, \alpha, \theta) := \begin{cases} D(W, \alpha, \theta) & \text{if } W \in \mathbf{R}_+^{m \times n},\ \alpha \in \mathbf{R}_{++}^m,\ \theta_i \in \Theta_i,\ i = 1, \ldots, m, \\ +\infty & \text{otherwise,} \end{cases}$$

$$\delta_M(W) := \begin{cases} 0 & \text{if } W \in M, \\ +\infty & \text{otherwise,} \end{cases}$$

$$\overline{f}_0(\alpha) := \begin{cases} f_0(\alpha) & \text{if } \alpha \in \Omega_{\underline{\alpha}}, \\ +\infty & \text{otherwise.} \end{cases}$$

Note that $M \times \Omega_l \times \Theta$ is a closed convex set. Thus, by Proposition 2.2.7 and problem (5.3.26), it suffices to show that $(\overline{W}, \overline{\alpha}, \overline{\theta})$ is a stationary point of problem (5.3.26). First, we have from Assumption 5.3.1 (A2) and (A3) that assumptions (i)–(iii) of Proposition 2.5.2 are satisfied. Next, we obtain $\mathrm{dom}\overline{D} = \mathbf{R}_+^{m \times n} \times \mathbf{R}_{++}^m \times \Theta_1 \times \ldots \times \Theta_m$ by the definition of $\overline{D}$, that is, assumption (iv) of Proposition 2.5.2 holds. Then, Proposition 2.5.2 implies that $(\overline{W}, \overline{\alpha}, \overline{\theta})$ is a coordinatewise minimum point of F. On the other hand, we have $(\overline{W}, \overline{\alpha}, \overline{\theta}) \in \mathbf{R}_{++}^{m \times n} \times \mathbf{R}_{++}^m \times \mathrm{int}\Theta$ because D is differentiable at $(\overline{W}, \overline{\alpha}, \overline{\theta})$. It then follows from the definition of $\overline{D}$ that $\overline{D}$ is differentiable at $(\overline{W}, \overline{\alpha}, \overline{\theta})$. Therefore, these results and Proposition 2.5.1 imply that $(\overline{W}, \overline{\alpha}, \overline{\theta})$ is a stationary point of problem (5.3.26). □

From Theorems 5.3.1 and 5.3.2, Algorithm 5.3.1 can get a stationary point of (5.2.2) if it solves (5.3.2).

Remark 5.3.1. *For the global convergence of Algorithm 5.3.1, we should get exact solutions of subproblems (5.3.24) and (5.3.25). As shown in Section 5.4, we can obtain them in some special cases.*

We now discuss when Assumption 5.3.1 (A1)–(A3) hold.

(1) Assumption 5.3.1 (A1) guarantees that $0 < \underline{\alpha}_i \leq \overline{\alpha}_i$ $(i = 1, \ldots, m)$. However, when we have a large amount of the observational data, the vector $\overline{\alpha}$ satisfies such conditions in many cases even if we do not suppose the existence of the vector $\underline{\alpha}$.

(2) Some distributions, such as logistic distributions, satisfy Assumption 5.3.1 (A2). For these details, see [11, Chapter 7].

(3) When we employ the L_1 regularization and/or indicator functions of closed convex sets as f_0 and f_i $(i = 1, \ldots, m)$, Assumption 5.3.1 (A3) holds.

Unfortunately, a Gaussian distribution $\mathcal{N}(x|\mu_i, \Lambda_i^{-1})$ does not satisfy the quasiconvexity condition in Assumption 5.3.1 (A2). However, under some reasonable assumptions, we can construct a global convergent BCD method for Gaussian mixtures. Note that $\theta_i = [\mu_i, \Lambda_i]$, $\Theta_i = \mathbf{R}^d \times \mathbf{S}^d_{++}$ and $\mathcal{V}_i = \mathbf{R}^d \times \mathbf{S}^d$ for each $i \in \{1, \ldots, m\}$ when mixture components are Gaussian. In addition, we use notations $\mu := [\mu_1, \ldots, \mu_m]$ and $\Lambda := [\Lambda_1, \ldots, \Lambda_m]$. For each $i \in \{1, \ldots, m\}$, we assume that the function f_i is separable with respect to μ_i and Λ_i, that is,

$$f_i(\theta_i) = f_i^{\mu}(\mu_i) + f_i^{\Lambda}(\Lambda_i), \tag{5.3.27}$$

where f_i^{μ} and f_i^{Λ} are proper lower semicontinuous quasiconvex on $\mathbf{R}^m$ and $\mathbf{S}^d$, respectively. Then, we execute the following two steps instead of Step 3 in Algorithm 5.3.1.

Step 3-1. For each $i \in \{1, \ldots, m\}$, obtain a solution μ_i^{t+1} of the following problem:

$$\begin{array}{ll} \text{minimize} & -\displaystyle\sum_{k=1}^{n} W_{ik}^t \log \mathcal{N}(x_k|\mu_i, (\Lambda_i^t)^{-1}) + f_i^{\mu}(\mu_i), \\ \text{subject to} & \mu_i \in \mathbf{R}^d. \end{array} \tag{5.3.28}$$

Step 3-2. For each $i \in \{1, \ldots, m\}$, obtain a solution Λ_i^{t+1} of the following problem:

$$\begin{array}{ll} \text{minimize} & -\displaystyle\sum_{k=1}^{n} W_{ik}^t \log \mathcal{N}(x_k|\mu_i^{t+1}, \Lambda_i^{-1}) + f_i^{\Lambda}(\Lambda_i), \\ \text{subject to} & \Lambda_i \succeq 0. \end{array} \tag{5.3.29}$$

Note that the modified method is also a BCD method. We call it Algorithm 5.3.2 in the remainder of this chapter.

The next assumptions are sufficient conditions under which Algorithm 5.3.2 has the global convergence property.

Assumption 5.3.2.

(A1) *The mixture components satisfy that* $p_i(x|\theta_i) = \mathcal{N}(x|\mu_i, \Lambda_i^{-1})$, $\theta_i = [\mu_i, \Lambda_i]$ $(i = 1, \ldots, m)$.

(A2) *The constant vector* $l \in \mathbf{R}^m$ *satisfies that* $l = \underline{\alpha}$, *where* $\underline{\alpha} \in \mathbf{R}^m$ *is a certain vector such that* $0 < \underline{\alpha}_i$ $(i = 1, \ldots, m)$ *and* $\sum_{i=1}^m \underline{\alpha}_i < 1$.

(A3) *Problem (5.3.2) has constraints* $0 \prec \underline{\lambda}_i I \preceq \Lambda_i$ $(i = 1, \ldots, m)$, *where* $0 < \underline{\lambda}_i$ $(i = 1, \ldots, m)$.

(A4) *The function* f_0 *is proper lower semicontinuous quasiconvex, and the functions* f_i $(i = 1, \ldots, m)$ *are written as (5.3.27) with proper lower semicontinuous quasiconvex functions* f_i^μ, f_i^Λ $(i = 1, \ldots, m)$.

Then, we give a theorem that guarantees the global convergence of Algorithm 5.3.2.

Theorem 5.3.3. *Suppose that Assumption 5.3.2 (A1)–(A4) hold. Suppose also that Algorithm 5.3.2 generates an infinite sequence* $\{(W^t, \alpha^{t+1}, \mu^{t+1}, \Lambda^{t+1})\}$ *that has an accumulation point* $(\overline{W}, \overline{\alpha}, \overline{\mu}, \overline{\Lambda})$. *Then,* $(\overline{W}, \overline{\alpha}, \overline{\mu}, \overline{\Lambda})$ *is a stationary point of problem (5.3.2).*

Proof. First, we show that D is differentiable at $(\overline{W}, \overline{\alpha}, \overline{\mu}, \overline{\Lambda})$. By (5.3.23) and Assumption 5.3.2 (A1) and (A2), we obtain that $0 < \overline{\alpha}_i$ $(i = 1, \ldots, m)$ and $0 < \overline{W}_{ik}$ $(i = 1, \ldots, m,\ k = 1, \ldots, n)$. Meanwhile, it follows from Assumption 5.3.2 (A3) that $\overline{\Lambda}_i \in \mathbf{S}_{++}^d$ $(i = 1, \ldots, m)$. These results imply that D is differentiable at $(\overline{W}, \overline{\alpha}, \overline{\mu}, \overline{\Lambda})$.

Secondly, we consider the following problem by using $\underline{\alpha} \in \mathbf{R}^m$ and $\underline{\lambda}_i \in \mathbf{R}$ $(i = 1, \ldots, m)$ in Assumption 5.3.2 (A2) and (A3):

$$\text{minimize} \quad F(W, \alpha, \mu, \Lambda), \tag{5.3.30}$$

where

$$F(W, \alpha, \mu, \Lambda) := \overline{D}(W, \alpha, \mu, \Lambda) + \delta_M(W) + \overline{f}_0(\alpha) + \sum_{i=1}^m f_i^\mu(\mu_i) + \sum_{i=1}^m \overline{f}_i^\Lambda(\Lambda_i),$$

$$\overline{D}(W, \alpha, \mu, \Lambda) := \begin{cases} D(W, \alpha, \mu, \Lambda) & \text{if } W \in \mathbf{R}_+^{m \times n},\ \alpha \in \mathbf{R}_{++}^m,\ \mu_i \in \mathbf{R}^d,\ \Lambda_i \in \mathbf{S}_{++}^d,\ i = 1, \ldots, m, \\ +\infty & \text{otherwise,} \end{cases}$$

$$\delta_M(W) := \begin{cases} 0 & \text{if } W \in M, \\ +\infty & \text{otherwise,} \end{cases}$$

$$\overline{f}_0(\alpha) := \begin{cases} f_0(\alpha) & \text{if } \alpha \in \Omega_{\underline{\alpha}}, \\ +\infty & \text{otherwise,} \end{cases}$$

$$\overline{f}_i^\Lambda(\Lambda_i) := \begin{cases} 0 & \text{if } \underline{\lambda}_i I \preceq \Lambda_i, \\ +\infty & \text{otherwise,} \end{cases} \quad i = 1, \ldots, m.$$

From the differentiability of D at $(\overline{W}, \overline{\alpha}, \overline{\mu}, \overline{\Lambda})$ and problem (5.3.30), this theorem can be shown in a way similar to the proof of Theorem 5.3.2. □

5.4 Implementation issue for special cases

In this section, we describe efficient solution methods solving subproblems (5.3.24), (5.3.28) and (5.3.29) for special cases such as Examples 1 and 2 of Section 5.2.

5.4.1 Maximum likelihood estimation with constraints on mixture coefficients

We discuss the maximum likelihood estimation with box constraints on mixture coefficients as described in Example 1 of Section 5.2. Since the update of the mixture coefficients appears only in subproblem (5.3.24) of Step 2, we only discuss how to solve subproblem (5.3.24).

Subproblem (5.3.24) has simple constraints $\sum_{i=1}^m \alpha_i = 1,\ l_i \le \alpha_i \le u_i\ (i = 1, \ldots, m)$. There exist efficient methods that solve special convex problems with the constraints in $O(m)$ [12, 47]. Although the objective function in (5.3.24) is different from those in [12, 47], we can construct an $O(m)$ method for (5.3.24) by using the ideas of [12, 47]. For the completeness of this thesis, we provide a concrete $O(m)$ method for (5.3.24).

For simplicity, we consider only the lower constraints $l_i \le \alpha_i\ (i = 1, \ldots, m)$, and let $W_{ik} := W_{ik}^t$. Note that we can construct a method for the problem with $l_i \le \alpha_i \le u_i\ (i = 1, \ldots, m)$ as in [12]. We assume that $l_i \in (0, 1]\ (i = 1, \ldots, m)$ and $\sum_{i=1}^m l_i < 1$.

For solving subproblem (5.3.24), we may find its KKT point. Let $(\alpha^*, \lambda^*, \gamma^*) \in \mathbf{R}^m \times \mathbf{R} \times \mathbf{R}^m$ be a KKT point satisfying

$$\sum_{i=1}^m \alpha_i^* = 1, \tag{5.4.1}$$

$$\frac{N_i}{\alpha_i^*} - \lambda^* + \gamma_i^* = 0, \quad \alpha_i^* - l_i \ge 0, \quad \gamma_i^* \ge 0, \quad \gamma_i^*(\alpha_i^* - l_i) = 0, \quad i = 1, \ldots, m, \tag{5.4.2}$$

where $N_i := \sum_{k=1}^n W_{ik}$. Note that λ^* and $\gamma_i^*\ (i = 1, \ldots, m)$ are Lagrange multipliers for $\sum_{i=1}^m \alpha_i^* = 1$ and $\alpha_i^* - l_i \ge 0\ (i = 1, \ldots, m)$, respectively.

As shown below, a partition of the set $\{N_i/l_i\}$ plays an important role to find $(\alpha^*, \lambda^*, \gamma^*)$. Thus, we define the following partitions $I(t)$ and $J(t)$, and the related functions.

$$I(t) := \{\ i\ \mid t \ge N_i/l_i\ \}, \qquad J(t) := \{\ i\ \mid t < N_i/l_i\ \},$$

$$\mu_{\min}(t) := \begin{cases} \max\limits_{i \in I(t)} \{N_i/l_i\} & \text{if } I(t) \ne \emptyset, \\ -\infty & \text{if } I(t) = \emptyset, \end{cases} \qquad \mu_{\max}(t) := \begin{cases} \min\limits_{i \in J(t)} \{N_i/l_i\} & \text{if } J(t) \ne \emptyset, \\ +\infty & \text{if } J(t) = \emptyset, \end{cases}$$

$$\alpha_i(t) := \begin{cases} l_i & \text{if } i \in I(t), \\ N_i/t & \text{if } i \in J(t), \end{cases} \qquad \gamma_i(t) := \begin{cases} t - N_i/l_i & \text{if } i \in I(t), \\ 0 & \text{if } i \in J(t), \end{cases}$$

$$\mu(t) := \sum_{i \in J(t)} N_i \Bigg/ \left(1 - \sum_{i \in I(t)} l_i\right), \qquad S(t) := \sum_{i=1}^m \alpha_i(t).$$

For the partitions and functions, the following properties hold.

Lemma 5.4.1. *Let $(\alpha^*, \lambda^*, \gamma^*)$ be a KKT point satisfying (5.4.1) and (5.4.2). Then, the following (i)–(v) hold.*

(i) *Let $I^* := \{\ i \mid \alpha_i^* = l_i\ \}$ and $J^* := \{\ i \mid \gamma_i^* = 0,\ \alpha_i^* \ne l_i\ \}$. Then,*

$$\alpha_i^* = \begin{cases} l_i & \text{if } i \in I^*, \\ N_i/\lambda^* & \text{if } i \in J^*, \end{cases} \qquad \lambda^* = \sum_{i \in J^*} N_i \Bigg/ \left(1 - \sum_{i \in I^*} l_i\right), \quad I(\lambda^*) = I^*, \quad J(\lambda^*) = J^*,$$

that is, $\alpha_i^* = \alpha_i(\lambda^*), \lambda^* = \mu(\lambda^*)$. *Moreover,* $\mu(\lambda^*) \in [\mu_{\min}(\lambda^*), \mu_{\max}(\lambda^*))$.

(ii) *If* $\mu(t^*) \in [\mu_{\min}(t^*), \mu_{\max}(t^*))$, *then* $(\alpha(\mu(t^*)), \mu(t^*), \gamma(\mu(t^*)))$ *is a KKT point of (5.3.24).*

(iii) *The function* S *is strictly monotonically decreasing. In addition,* $S(\lambda^*) = 1$.

(iv) *If* $\kappa \in [\mu_{\min}(t), \mu_{\max}(t))$, *then* $\mu(\kappa) = \mu(t), \mu_{\min}(\kappa) = \mu_{\min}(t)$ *and* $\mu_{\max}(\kappa) = \mu_{\max}(t)$.

(v) *If* $\mu(t) \notin [\mu_{\min}(t), \mu_{\max}(t))$, *then we have either* $S(\mu_{\min}(t)) < 1$ *or* $S(\mu_{\max}(t)) \geq 1$.

Proof. (i) Since $\gamma_i^*(\alpha_i^* - l_i) = 0$ from (5.4.2), we have $\gamma_i^* = 0$ or $\alpha_i^* = l_i$. It then follows from $N_i/\alpha_i^* - \lambda^* + \gamma_i^* = 0$ in (5.4.2) that $\alpha_i^* = N_i/\lambda^*$ whenever $\gamma_i^* = 0$. From the definitions of I^* and J^*, we have

$$\alpha_i^* = \begin{cases} l_i & \text{if } i \in I^*, \\ N_i/\lambda^* & \text{if } i \in J^*. \end{cases}$$

Since $1 = \sum_{i=1}^m \alpha_i^* = \sum_{i \in I^*} l_i + \sum_{i \in J^*} N_i/\lambda^*$ from (5.4.1),

$$\lambda^* = \sum_{i \in J^*} N_i \Bigg/ \left(1 - \sum_{i \in I^*} l_i\right).$$

Next, we show that $I(\lambda^*) = I^*$ and $J(\lambda^*) = J^*$. Since $\gamma_i^* \geq 0$ and $N_i/\alpha_i^* - \lambda^* + \gamma_i^* = 0$ by (5.4.2), we have $\alpha_i^* \geq N_i/\lambda^*$. Then, for each $i \in I^*$, we obtain $l_i = \alpha_i^* \geq N_i/\lambda^*$, and hence $\lambda^* \geq N_i/l_i$. As the result, we obtain $i \in I(\lambda^*)$. Conversely, if $i \in I(\lambda^*)$, then $\lambda^* \geq N_i/l_i$, and hence $l_i \geq N_i/\lambda^*$. Now, we assume that $i \in J^*$, that is, $\gamma_i^* = 0$ and $\alpha_i^* \neq l_i$. It then follows from $N_i/\alpha_i^* - \lambda^* + \gamma_i^* = 0$ and $\alpha_i^* \geq l_i$ in (5.4.2) that $N_i/\lambda^* = \alpha_i^* > l_i$. Thus, $N_i/\lambda^* > l_i \geq N_i/\lambda^*$, which is contradictory. Therefore, $i \notin J^*$, that is, $i \in I^*$ because

$$I^* \cup J^* = \{1, \ldots, m\} \quad \text{and} \quad I^* \cap J^* = \emptyset. \tag{5.4.3}$$

Consequently, we have $I(\lambda^*) = I^*$. The relation $J(\lambda^*) = J^*$ is obtained from (5.4.3). Moreover, we get $\alpha_i^* = \alpha_i(\lambda^*)$ and $\lambda^* = \mu(\lambda^*)$ from the definitions of $\alpha_i(t)$ and $\mu(t)$.

The definitions of $\mu_{\min}(t)$ and $\mu_{\max}(t)$ imply that $t \in [\mu_{\min}(t), \mu_{\max}(t))$ for all $t \in \mathbf{R}$. Therefore, $\mu(\lambda^*) = \lambda^* \in [\mu_{\min}(\lambda^*), \mu_{\max}(\lambda^*))$.
(ii) We show that $(\alpha(\mu(t^*)), \mu(t^*), \gamma(\mu(t^*)))$ satisfies the KKT conditions (5.4.1) and (5.4.2). From the definitions of $\alpha_i(t)$ and $\gamma_i(t)$, we obtain

$$\alpha_i(\mu(t^*)) = \begin{cases} l_i & \text{if } i \in I(\mu(t^*)), \\ N_i/\mu(t^*) & \text{if } i \in J(\mu(t^*)), \end{cases} \quad \gamma_i(\mu(t^*)) = \begin{cases} \mu(t^*) - N_i/l_i & \text{if } i \in I(\mu(t^*)), \\ 0 & \text{if } i \in J(\mu(t^*)). \end{cases} \tag{5.4.4}$$

Moreover, $t^* \in [\mu_{\min}(t^*), \mu_{\max}(t^*))$ by the definitions of $\mu_{\min}(t)$ and $\mu_{\max}(t)$. It then follows from the assumption $\mu(t^*) \in [\mu_{\min}(t^*), \mu_{\max}(t^*))$ that $I(t^*) = I(\mu(t^*))$ and $J(t^*) = J(\mu(t^*))$. Thus, (5.4.1) follows from the definitions of $\alpha_i(\mu(t^*))$ and $\mu(t^*)$. Next, we show (5.4.2). Suppose that $i \in I(t^*)$. It then follows from $I(t^*) = I(\mu(t^*))$, (5.4.4) and the definition of $\mu_{\min}(t^*)$ that

$$\frac{N_i}{\alpha_i(\mu(t^*))} - \mu(t^*) + \gamma_i(\mu(t^*)) = \frac{N_i}{l_i} - \mu(t^*) + \mu(t^*) - \frac{N_i}{l_i} = 0,$$
$$\alpha_i(\mu(t^*)) - l_i = l_i - l_i = 0,$$
$$\gamma_i(\mu(t^*)) = \mu(t^*) - \frac{N_i}{l_i} \geq \mu_{\min}(t^*) - \frac{N_i}{l_i} \geq 0.$$

Similarly, if $i \in J(t^*)$, then

$$\frac{N_i}{\alpha_i(\mu(t^*))} - \mu(t^*) + \gamma_i(\mu(t^*)) = \mu(t^*) - \mu(t^*) + 0 = 0,$$
$$\gamma_i(\mu(t^*)) = 0,$$
$$\alpha_i(\mu(t^*)) - l_i = \frac{N_i}{\mu(t^*)} - l_i \geq \frac{N_i}{\mu_{\max}(t^*)} - l_i \geq 0.$$

Therefore, the KKT conditions (5.4.1) and (5.4.2) hold.
(iii) Let $t_1 < t_2$. Since $I(t_1) \subseteq I(t_2)$, $J(t_2) \subseteq J(t_1)$ and $I(t_1) \cup J(t_1) = I(t_2) \cup J(t_2)$, we have $I(t_2) = I(t_1) \cup (J(t_1) \backslash J(t_2))$. Thus,

$$S(t_1) = \sum_{i \in I(t_1)} l_i + \sum_{i \in J(t_1) \backslash J(t_2)} \frac{N_i}{t_1} + \sum_{i \in J(t_2)} \frac{N_i}{t_1} > \sum_{i \in I(t_1)} l_i + \sum_{i \in J(t_1) \backslash J(t_2)} l_i + \sum_{i \in J(t_2)} \frac{N_i}{t_2} = S(t_2),$$

where the inequality follows from $t_1 < t_2$. Moreover, since $\alpha_i^* = \alpha_i(\lambda^*)$ by (i), $S(\lambda^*) = \sum_{i=1}^m \alpha_i(\lambda^*) = \sum_{i=1}^m \alpha_i^* = 1$.
(iv) We have the desired results from the definitions of $\mu(t), \mu_{\min}(t)$ and $\mu_{\max}(t)$.
(v) In order to prove by contradiction, we suppose that $S(\mu_{\min}(t)) \geq 1$ and $S(\mu_{\max}(t)) < 1$. We obtain $\lambda^* \in [\mu_{\min}(t), \mu_{\max}(t))$ by (iii). It then follows from (iv) that $\mu(t) = \mu(\lambda^*)$. Since $\mu(\lambda^*) = \lambda^*$ from (i), we get $\mu(t) = \lambda^* \in [\mu_{\min}(t), \mu_{\max}(t))$. However, this result contradicts $\mu(t) \notin [\mu_{\min}(t), \mu_{\max}(t))$. □

From Lemma 5.4.1 (i), there exists $t^* \in \mathbf{R}$ such that

$$\mu_{\min}(t^*) \leq \mu(t^*) < \mu_{\max}(t^*). \tag{5.4.5}$$

Conversely, if we find t^* satisfying (5.4.5), then we can obtain the solution α^* of problem (5.3.24) by Lemma 5.4.1 (ii). Thus, we consider how to find t^* satisfying (5.4.5).

Now, suppose that $\bar{t}$ satisfies (5.4.5). Since $\mu_{\min}(\bar{t}) \in [\mu_{\min}(\bar{t}), \mu_{\max}(\bar{t}))$, we have $\mu(\mu_{\min}(\bar{t})) = \mu(\bar{t}), \mu_{\min}(\mu_{\min}(\bar{t})) = \mu_{\min}(\bar{t})$ and $\mu_{\max}(\mu_{\min}(\bar{t})) = \mu_{\max}(\bar{t})$ from Lemma 5.4.1 (iv). It then follows from (5.4.5) that $\mu(\mu_{\min}(\bar{t})) \in [\mu_{\min}(\mu_{\min}(\bar{t})), \mu_{\max}(\mu_{\min}(\bar{t})))$, i.e., $\mu_{\min}(\bar{t})$ also satisfies (5.4.5). Note that $\mu_{\min}(\bar{t})$ is included in the set $\{N_i/l_i\}$. Therefore we can find t^* only in the set $\{N_i/l_i\}$.

Then, we choose a median[1] c of the set $\{N_i/l_i\}$ as a candidate of t^*. If $\mu(c) \in [\mu_{\min}(c), \mu_{\max}(c))$, then we may regard c as t^* satisfying (5.4.5). In what follows, we discuss the case where $\mu(c) \notin [\mu_{\min}(c), \mu_{\max}(c))$. By Lemma 5.4.1 (v), we may consider the following two cases:

Case 1: $S(\mu_{\min}(c)) < 1$.

It then follows from Lemma 5.4.1 (iii) that $\lambda^* < \mu_{\min}(c)$, and hence $t^* \in \{\, N_i/l_i \mid i \in I(c) \,\}$.

Case 2: $S(\mu_{\max}(c)) \geq 1$.

It then follows from Lemma 5.4.1 (iii) that $\lambda^* \geq \mu_{\max}(c)$, and hence $t^* \in \{\, N_i/l_i \mid i \in J(c) \,\}$.

Thus, the set $\{N_i/l_i\}$ of the candidates of t^* is reduced to $\{\, N_i/l_i \mid i \in I(c) \,\}$ or $\{\, N_i/l_i \mid i \in J(c) \,\}$, that is, the number of candidates becomes half. When Case 1 occurs, we choose a median of

[1] A scalar c in $\left\{ \frac{N_1}{l_1}, \ldots, \frac{N_m}{l_m} \right\}$ is called the median of $\left\{ \frac{N_1}{l_1}, \ldots, \frac{N_m}{l_m} \right\}$ if $|\{\, i \mid N_i/l_i < c \,\}| < \left\lceil \frac{m}{2} \right\rceil \leq |\{\, i \mid N_i/l_i \leq c \,\}|$.

$\{ N_i/l_i \mid i \in I(c) \}$ as the next candidate of t^*, and we again carry out the same procedure. After the procedure, we obtain a KKT point or reduce the number of candidates of t^* to half further. Similarly, we apply the same procedure to $\{ N_i/l_i \mid i \in J(c) \}$ when Case 2 occurs. Repeating the procedure, we can find a KKT point.

Now, we discuss the computational complexity of the above method. Note that a median of $\{N_1/l_1, \ldots, N_m/l_m\}$ can be found in $O(m)$ time [9]. Moreover, the function values $\mu(c_1), \mu_{\min}(c_1), \mu_{\max}(c_1)$ and $S(c_1)$ are evaluated in $O(m)$ time. Thus, the first iteration of the method is executed in $O(m)$ time. Next, we perform the second iteration for $m/2$ elements, that is, $\{ N_i/l_i \mid i \in I(c) \}$ or $\{ N_i/l_i \mid i \in J(c) \}$. Then, the median c_2 of the set is found in $O(m/2)$ time. On the other hand, since $\mu(c_2), \mu_{\min}(c_2), \mu_{\max}(c_2)$ and $S(c_2)$ are defined with m elements, the direct evaluations of these function values take $O(m)$ time. Fortunately, we can reduce the time by using products of the previous iteration. To see this, consider the case where $S(\mu_{\min}(c_1)) < 1$. Then, c_2 is a median of $\{ N_i/l_i \mid i \in I(c_1) \}$ and $\mu(c_2)$ is given by

$$\mu(c_2) = \sum_{i \in J(c_1) \cup J(c_2)} N_i \Bigg/ \left(1 - \sum_{i \in I(c_2)} l_i\right) = \left(\sum_{i \in J(c_1)} N_i + \sum_{i \in J(c_2)} N_i\right) \Bigg/ \left(1 - \sum_{i \in I(c_2)} l_i\right).$$

Note that $\sum_{i \in J(c_1)} N_i$ has been calculated for $\mu(c_1)$ in the first iteration. Therefore, we can omit its calculation, and hence we have to calculate only $\sum_{i \in J(c_2) N_i}$ and $\sum_{i \in J(c_2)} l_i$ to evaluate $\mu(c_2)$. These calculations take $O(m/2)$ time. Similarly, $\mu_{\min}(c_2), \mu_{\max}(c_2)$ and $S(c_2)$ are evaluated in $O(m/2)$ time. Consequently, the second iteration is done in $O(m/2)$ time. Repeating these calculations, the worst computational time is $O(m + m/2 + m/4 + \ldots) = O(m)$. Note that, when $l_i = 0$ $(i = 1, \ldots, m)$, the solution $\alpha^* \in \mathbf{R}^m$ is calculated by $\alpha_i^* = N_i/n$, which takes $O(m)$ time. The computational complexity for $l_i > 0$ $(i = 1, \ldots, m)$ is the same as that for $l_i = 0$ $(i = 1, \ldots, m)$.

5.4.2 Maximum likelihood estimation for Gaussian mixtures

Now, we consider the case where mixture components are Gaussian, i.e., $p_i(x|\theta_i) = \mathcal{N}(x|\mu_i, \Lambda_i^{-1})$, $\theta_i = [\mu_i, \Lambda_i]$ $(i = 1, \ldots, m)$.

The maximum likelihood estimation for Gaussian mixtures is equivalent to problem (5.3.2) with

$$f_0(\alpha) := \begin{cases} 0 & \text{if } \alpha \in \Omega_0, \\ +\infty & \text{otherwise,} \end{cases} \quad f_i^{\mu}(\mu_i) := 0, \quad f_i^{\Lambda}(\Lambda_i) := \begin{cases} 0 & \text{if } \Lambda_i \succeq 0, \\ +\infty & \text{otherwise,} \end{cases} \quad i = 1, \ldots, m. \tag{5.4.6}$$

Then, $\alpha_i^{t+1}, \mu_i^{t+1}$ and Λ_i^{t+1} in Steps 2, 3-1 and 3-2 of Algorithm 5.3.2 are given by

$$\alpha_i^{t+1} = \frac{N_i^t}{n}, \quad \mu_i^{t+1} = \frac{1}{N_i^t} \sum_{k=1}^{n} W_{ik}^t x_k, \quad \Lambda_i^{t+1} = \left(\frac{1}{N_i^t} \sum_{k=1}^{n} W_{ik}^t (x_k - \mu_i^{t+1})(x_k - \mu_i^{t+1})^\top\right)^{-1}, \tag{5.4.7}$$

where $N_i^t := \sum_{k=1}^{n} W_{ik}^t$. We see that (5.4.7) is equivalent to the EM algorithm. Note that the equivalence has already been pointed out in [67].

5.4.3 Maximum likelihood estimation for Gaussian mixtures with constraints on precision matrices

In this subsection, we consider the maximum likelihood estimation for Gaussian mixtures that has additional constraints on the precision matrices such that $\underline{\lambda}_i I \preceq \Lambda_i \preceq \overline{\lambda}_i I$ $(i = 1, \ldots, m)$, where $\underline{\lambda}_i, \overline{\lambda}_i \in \mathbf{R}$ $(i = 1, \ldots, m)$ are constants such that $0 < \underline{\lambda}_i < \overline{\lambda}_i$.

In this case, we should replace f_i^Λ in (5.4.6) with

$$f_i^\Lambda(\Lambda_i) := \begin{cases} 0 & \text{if } \underline{\lambda}_i I \preceq \Lambda_i \preceq \overline{\lambda}_i I, \\ \infty & \text{otherwise,} \end{cases} \qquad i = 1, \ldots, m.$$

Note that α_i^{t+1} and μ_i^{t+1} are also given by (5.4.7) because subproblems with respect to α_i and μ_i are same as those in Subsection 5.4.2. On the other hand, subproblem (5.3.29) with respect to Λ_i is different, and it is expressed as

$$\begin{array}{ll} \underset{\Lambda_i \in \mathbf{S}^d}{\text{minimize}} & \operatorname{tr}\left(A_i^t \Lambda_i\right) - \log\det \Lambda_i, \\ \text{subject to} & \underline{\lambda}_i I \preceq \Lambda_i \preceq \overline{\lambda}_i I, \end{array} \tag{5.4.8}$$

where

$$A_i^t := \frac{1}{N_i^t} \sum_{k=1}^n W_{ik}^t (x_k - \mu_i^{t+1})(x_k - \mu_i^{t+1})^\top, \tag{5.4.9}$$

and N_i^t is given in Subsection 5.4.2.

Thanks to the constraints $\underline{\lambda}_i I \preceq \Lambda_i$ $(i = 1, \ldots, m)$, Assumption 5.3.2 (A3) holds. Moreover, if we also add the constraints on mixture coefficients as described in Subsection 5.4.1, Assumption 5.3.2 (A2) also holds. Therefore, such constraints guarantee the global convergence of Algorithm 5.3.2.

Now, we discuss how to solve (5.4.8). As shown below, we can provide a solution of (5.4.8) analytically. For simplicity, let $A := A_i^t$, $\Lambda := \Lambda_i$, $\underline{\lambda} := \underline{\lambda}_i$ and $\overline{\lambda} := \overline{\lambda}_i$ in the rest of this subsection.

Since problem (5.4.8) is convex, $\Lambda_i^* \in \mathbf{S}^d$ satisfying the following KKT conditions is an optimal solution:

$$\begin{array}{lll} A - (\Lambda^*)^{-1} + U^* - V^* = 0, & (\underline{\lambda} I - \Lambda^*)U^* = 0, & (\overline{\lambda} I - \Lambda^*)V^* = 0, \\ \underline{\lambda} I \preceq \Lambda^* \preceq \overline{\lambda} I, & 0 \preceq U^*, & 0 \preceq V^*, \end{array} \tag{5.4.10}$$

where $U^* \in \mathbf{S}^d$ and $V^* \in \mathbf{S}^d$ are Lagrange multipliers for $\underline{\lambda} I \preceq \Lambda^*$ and $\Lambda^* \preceq \overline{\lambda} I$, respectively. We have from (5.4.10) that Λ^*, U^*, V^* and A commute mutually. Then, Proposition 2.2.1 (e) yields that Λ^*, U^*, V^* and A are simultaneously diagonalizable, that is, there exists an orthogonal matrix $P \in \mathbf{S}^d$ such that

$$\begin{array}{ll} P^\top \Lambda^* P = \operatorname{diag}(\lambda_1^*, \ldots, \lambda_d^*), & P^\top U^* P = \operatorname{diag}(u_1^*, \ldots, u_d^*), \\ P^\top V^* P = \operatorname{diag}(v_1^*, \ldots, v_d^*), & P^\top A P = \operatorname{diag}(a_1, \ldots, a_d), \end{array}$$

where $\lambda_j^*, u_j^*, v_j^*$ and a_j $(j = 1, \ldots, d)$ are eigenvalues of matrices Λ^*, U^*, V^* and A, respectively. Pre- and post-multiplying (5.4.10) by $P^\top$ and P, respectively,

$$\begin{array}{lll} a_j - (\lambda_j^*)^{-1} + u_j^* - v_j^* = 0, & (\underline{\lambda} - \lambda_j^*)u_j^* = 0, & (\overline{\lambda} - \lambda_j^*)v_j^* = 0, \\ \underline{\lambda} \le \lambda_j^* \le \overline{\lambda}, & 0 \le u_j^*, & 0 \le v_j^* \end{array} \tag{5.4.11}$$

for $j = 1, \ldots, d$. Therefore, we have from (5.4.11) that

$$\Lambda^* = P \operatorname{diag}(\lambda_1^*, \ldots, \lambda_d^*)\, P^\top, \quad \lambda_j^* = \begin{cases} \overline{\lambda} & \text{if } 1/\overline{\lambda} \geq a_j, \\ 1/a_j & \text{if } 1/\overline{\lambda} \leq a_j \leq 1/\underline{\lambda}, \\ \underline{\lambda} & \text{if } 1/\underline{\lambda} \leq a_j, \end{cases} \quad j = 1, \ldots, d. \tag{5.4.12}$$

In order to obtain Λ^*, we may conduct the following procedure. We first get the eigenvalues a_j $(j = 1, \ldots, d)$ and the orthogonal matrix P by diagonalizing A. Next, we calculate Λ^* by (5.4.12).

5.4.4 Maximum likelihood estimation for Gaussian mixtures with sparse precision matrices

We also discuss the maximum likelihood estimation for Gaussian mixtures in Subsection 5.4.3. However, we add the L_1 regularization in order to obtain precision matrices being sparse. In this case, we should replace f_i^Λ in (5.4.6) with

$$f_i^\Lambda(\Lambda_i) := \begin{cases} \rho_i \|\Lambda_i\|_1 & \text{if } \underline{\lambda}_i I \preceq \Lambda_i \preceq \overline{\lambda}_i I, \\ +\infty & \text{otherwise,} \end{cases} \quad i = 1, \ldots, m,$$

where $\rho_1, \ldots, \rho_m$ are positive constants.

Note that α_i^{t+1} and μ_i^{t+1} are also given by (5.4.7) as mentioned in Subsection 5.4.3. On the other hand, subproblem (5.3.29) with respect to Λ_i is different, and it is written as

$$\begin{array}{ll} \text{minimize} & \operatorname{tr}\left(A_i^t \Lambda_i\right) - \log\det\Lambda_i + \tau_i^t \|\Lambda_i\|_1, \\ \text{subject to} & \underline{\lambda}_i I \preceq \Lambda_i \preceq \overline{\lambda}_i I, \end{array} \tag{5.4.13}$$

where A_i^t is given by (5.4.9). We can obtain the solution Λ_i^{t+1} of problem (5.4.13) by the existing methods such as [38, 39, 73].

5.5 Numerical experiments

In this section, we report two numerical experiments for the models discussed in Subsections 5.4.1 and 5.4.3. The program was coded in MATLAB R2010a and run on a machine with an Intel Core i7 920 2.67GHz CPU and 3.00GB RAM.

Experiment 1 for the model discussed in Subsection 5.4.1

In the Experiment 1, we investigate the validity of the model discussed in Subsection 5.4.1.

Throughout the Experiment 1, we used the observational data $X = [x_1, \ldots, x_n] \in \mathbf{R}^{1\times n}$ and the test data $\widetilde{X} := [\widetilde{x}_1, \ldots, \widetilde{x}_{10000}] \in \mathbf{R}^{1\times 10000}$ generated by the following Gaussian mixture with $d = 1$ and $m = 5$:

$$p(x) = \frac{1}{5}\mathcal{N}(x|-10, 5) + \frac{1}{5}\mathcal{N}(x|-8, 5) + \frac{1}{5}\mathcal{N}(x|0, 5) + \frac{1}{5}\mathcal{N}(x|8, 5) + \frac{1}{5}\mathcal{N}(x|10, 5). \tag{5.5.1}$$

For the given observational data X, we estimated parameters of the Gaussian mixture with $d = 1$ and $m = 5$, that is, we solved the following problem by Algorithm 5.3.2:

$$\begin{aligned} &\text{maximize} && \sum_{k=1}^{n} \log\left(\sum_{i=1}^{5} \alpha_i \mathcal{N}(x_k|\mu_i, \Lambda_i^{-1})\right), \\ &\text{subject to} && \sum_{i=1}^{5} \alpha_i = 1,\ l_i \le \alpha_i,\ 0 \le \Lambda_i,\ i = 1, \ldots, 5. \end{aligned} \tag{5.5.2}$$

In the Experiment 1, we estimated parameters by exploiting three models with $l_i = 0$ $(i = 1, \ldots, 5)$, $l_i = 0.1$ $(i = 1, \ldots, 5)$ and $l_i = 0.15$ $(i = 1, \ldots, 5)$ in (5.5.2).

An initial point of Algorithm 5.3.2 was chosen as follows. We set $\alpha_i^0 = 1, \Lambda_i^0 = 1$ $(i = 1, \ldots, 5)$. A mean μ^0 was set to the computational result of K-means algorithm (`kmeans`) in MATLAB. Algorithm 5.3.2 was stopped when $|D(W^{t+1}, \alpha^t, \mu^t, \Lambda^t) - D(W^t, \alpha^{t-1}, \mu^{t-1}, \Lambda^{t-1})| < 10^{-5}$, where the function D is defined by (5.3.3).

Tables 5.1 and 5.2 show the results when the number of the observational data is 30 and 100, respectively. In each case, we carried out the maximum likelihood estimation 15 times for 15 different observational data. In two tables, we report the log-likelihoods for the observational data and the test data. Note that we used the same test data $\widetilde{X} \in \mathbf{R}^{1\times 10000}$ in all experiments. Since the amount of the test data is sufficiently large, we may consider that the estimation with bigger log-likelihood for the test data is better than that with small one. For each experiment in Table 5.1, numbers with boldface type indicate the highest log-likelihood among the various l_i. Furthermore, "$*$" in Tables indicates that Algorithm 5.3.2 was stopped by numerical difficulty. The reason for the difficulty is that the mixture coefficient α_i became too small.

From Table 5.1, we see that the model with $l_i = 0$ is better than the models with $l_i = 0.1$ and $l_i = 0.15$ from the viewpoint of the log-likelihood **for the observational data**. The results are quite natural because the feasible set with $l_i = 0$ is larger than those with $l_i = 0.1$ or $l_i = 0.15$. On the other hand, from the viewpoint of the log-likelihood **for the test data**, the models with $l_i = 0.1$ and $l_i = 0.15$ are better than the model with $l_i = 0$ for many trials. In particular, the model with $l_i = 0.15$ tends to be the best. This is because the true mixture coefficient is 0.2 as in (5.5.1). Moreover, the estimation of the model with $l_i = 0$ is overfitting for the small observational data. This can be seen in Figures 5.1, 5.2 and Table 5.3 that present the details of the numerical result for No. 3 in Table 5.1. Figures 5.1 and 5.2 are probability density functions obtained by the models with $l_i = 0$ and $l_i = 0.15$, respectively. In the both figures, the black dash line indicates the probability density function of the true mixture distribution (5.5.1), and the black line indicates the estimated probability density function. Table 5.3 presents the estimated parameters. From Table 5.3, we see that α_5 and Λ_5^{-1} of the model with $l_i = 0$ are very small. Thus, the probability density function value in Figure 5.1 becomes very large around $\mu_5 = 4.9377$. This phenomenon sometimes occurred when the amount of the observational data is small. See [8, Section 9.2.1] for its details. On the other hand, such a singular phenomenon did not happen on the model with $l_i = 0.15$ (Figure 5.2).

From Table 5.2, we do not see big differences in the log-likelihoods for the test data among the models. The reason for these results is that we were able to estimate parameters correctly regardless of the value of l_i because we had sufficient amount of the observational data.

From these results, even if the amount of the observational data is small, the model with l_i close to the true value is expected to avoid the overfitting and find an appropriate estimation.

Experiment 2 for the model discussed in Subsection 5.4.3

In the Experiment 2, we use the model discussed in Subsection 5.4.3, and study its effectivity.

In this experiment, we used the observational data $X = [x_1, \ldots, x_n] \in \mathbf{R}^{d\times n}$ and the test data $\widetilde{X} = [\widetilde{x}_1, \ldots, \widetilde{x}_{10000}] \in \mathbf{R}^{d\times 10000}$. These data are generated by the following Gaussian mixture:

$$p(x) = \sum_{i=1}^{10} \frac{1}{10} \mathcal{N}(x|\hat{\mu}_i, \hat{\Lambda}_i^{-1}),$$

where the elements of $\hat{\mu}_i$ were selected randomly from the interval $[-1, 1]$, and $\hat{\Lambda}_i^{-1}$ $(i = 1, \ldots, 10)$ are selected as follows. First, we generated a matrix $A_i \in \mathbf{R}^d$ $(i = 1, \ldots, 10)$ whose elements are normally distributed with mean 0 and variance 1. Then we set $\hat{\Lambda}_i^{-1} := (A_i^\top A_i)^{\frac{1}{2}}$ $(i = 1, \ldots, 10)$. For the observational data X, we solved the following model by Algorithm 5.3.2 in order to estimate parameters α_i, μ_i and Λ_i^{-1} $(i = 1, \ldots, 10)$:

$$\begin{aligned} \text{maximize} \quad & \sum_{k=1}^{500} \log \left(\sum_{i=1}^{10} \alpha_i \mathcal{N}(x_k|\mu_i, \Lambda_i^{-1}) \right), \\ \text{subject to} \quad & \sum_{i=1}^{10} \alpha_i = 1,\ 10^{-3} \le \alpha_i,\ \underline{\lambda}_i I \preceq \Lambda_i \preceq \overline{\lambda}_i I,\ i = 1, \ldots, 10. \end{aligned} \tag{5.5.3}$$

In the experiments, we estimated parameters by using two models with $(\underline{\lambda}_i, \overline{\lambda}_i) = (0, \infty)$ and $(\underline{\lambda}_i, \overline{\lambda}_i) = (10^{-3}, 10^3)$ in (5.5.3). In the following, the models (A) and (B) indicate the model with $(\underline{\lambda}_i, \overline{\lambda}_i) = (0, \infty)$ and $(\underline{\lambda}_i, \overline{\lambda}_i) = (10^{-3}, 10^3)$, respectively.

An initial point of Algorithm 5.3.2 was selected as follows. We chose $\alpha_i^0 = 1, \Lambda_i^0 = I$ $(i = 1, \ldots, 10)$, and set μ^0 as the computational result of K-means algorithm (`kmeans`) in MATLAB. Moreover, we used the same termination criterion as the Experiment 1.

Tables 5.4 and 5.5 show the results when the dimension d of the observational data X is 10 and 30, respectively. In each case, we conducted the maximum likelihood estimation 10 times by using observational data. In No. 1 of Tables 5.4 and 5.5, we exploited the observational data X such that $n = 100$. In the subsequent estimations, we added 100 observational data into the previous ones, and used those data as the observational data. Note that we exploited the same test data in each dimension d. In Tables 5.4 and 5.5, we report the log-likelihoods for both the observational and test data divided by the numbers of data, respectively. As with the Experiment 1, " $*$ " indicates that Algorithm 5.3.2 was stopped by numerical difficulty.

As seen in Table 5.4 when $d = 10$, we do not see big differences between the both models. On the other hand, as seen in Table 5.5, the differences appeared between the models (A) and (B). Although the model (A) could not estimate parameters when the amount of the observational data is small, the model (B) could estimate parameters owing to the constraints $\underline{\lambda}_i I \preceq \Lambda_i \preceq \overline{\lambda}_i I$ $(i = 1, \ldots, m)$.

Table 5.1: Comparison of log-likelihoods (The amount of data is 30.)

	$l_i = 0$ $(i = 1, \ldots, 5)$		$l_i = 0.1$ $(i = 1, \ldots, 5)$		$l_i = 0.15$ $(i = 1, \ldots, 5)$	
No.	Observation	Test	Observation	Test	Observation	Test
1	-92.6920	-3.8819e+004	-93.9322	-3.8646e+004	-94.0707	**-3.8355e+004**
2	-85.1689	-4.2377e+004	-85.3814	**-4.2056e+004**	-86.7102	-4.2398e+004
3	-89.0016	-3.5282e+004	-94.2949	-3.4429e+004	-94.2979	**-3.4427e+004**
4	-83.7478	-4.8337e+004	-83.9552	-4.8651e+004	-90.0500	**-3.6657e+004**
5	-90.9218	-3.5632e+004	-91.1861	-3.6142e+004	-94.1681	**-3.5011e+004**
6	-87.3364	-3.6083e+004	-89.0853	-3.5634e+004	-93.4515	**-3.3895e+004**
7	-94.0355	-3.4853e+004	-94.1238	-3.4554e+004	-94.3455	**-3.4372e+004**
8	-85.6939	-3.8715e+004	-85.8422	-3.8203e+004	-86.2903	**-3.7063e+004**
9	-93.2788	-3.6004e+004	-97.6769	**-3.3735e+004**	-97.7062	-3.3868e+004
10	-86.7174	-3.8413e+004	-86.9268	-3.8697e+004	-90.2407	**-3.6068e+004**
11	-89.2880	**-3.5906e+004**	-89.4100	-3.6042e+004	-90.0250	-3.6390e+004
12	$*$	$*$	-88.0714	-3.8834e+003	-90.1522	**-3.5275e+003**
13	-87.4854	-3.9886e+004	-91.6894	-3.9187e+004	-94.7792	**-3.6992e+004**
14	$*$	$*$	-100.3292	-3.3846e+004	-101.2753	**-3.3538e+004**
15	-94.9501	-3.4860e+004	-94.9503	-3.4865e+004	-95.2875	**-3.4847e+004**

Table 5.2: Comparison of log-likelihoods (The amount of data is 100.)

	$l_i = 0$ $(i = 1, \ldots, 5)$		$l_i = 0.1$ $(i = 1, \ldots, 5)$		$l_i = 0.15$ $(i = 1, \ldots, 5)$	
No.	Observation	Test	Observation	Test	Observation	Test
1	-333.3528	-3.3148e+004	-333.3910	-3.3173e+004	-333.3504	-3.3173e+004
2	-320.6859	-3.3347e+004	-322.3568	-3.2981e+004	-323.1440	-3.3170e+004
3	-321.1681	-3.3312e+004	-321.1681	-3.3312e+004	-321.1942	-3.3346e+004
4	-321.6798	-3.3584e+004	-325.7542	-3.3380e+004	-327.4315	-3.3076e+004
5	-317.9389	-3.3457e+004	-318.2513	-3.3593e+004	-319.8712	-3.3566e+004
6	-321.1656	-3.3005e+004	-321.8685	-3.2909e+004	-321.5855	-3.3038e+004
7	-316.4248	-3.3408e+004	-321.0262	-3.3135e+004	-321.8640	-3.3118e+004
8	-315.8101	-3.4256e+004	-317.1999	-3.3815e+004	-317.8504	-3.4037e+004
9	$*$	$*$	-326.2708	-3.3251e+004	-325.2491	-3.3916e+004
10	-316.0259	-3.3110e+004	-316.9840	-3.3029e+004	-316.9493	-3.3088e+004
11	-305.3455	-3.3814e+004	-307.2223	-3.3443e+004	-308.8353	-3.3541e+004
12	-334.4943	-3.3415e+004	-334.9925	-3.3392e+004	-338.3864	-3.3058e+004
13	-307.9100	-3.4562e+004	-307.9106	-3.4554e+004	-308.6720	-3.4483e+004
14	-312.5304	-3.3674e+004	-312.5304	-3.3674e+004	-312.5319	-3.3678e+004
15	-306.6108	-3.4756e+004	-307.6602	-3.4569e+004	-315.3187	-3.3014e+004

Figure 5.1: Results of No. 3 in Table 5.1

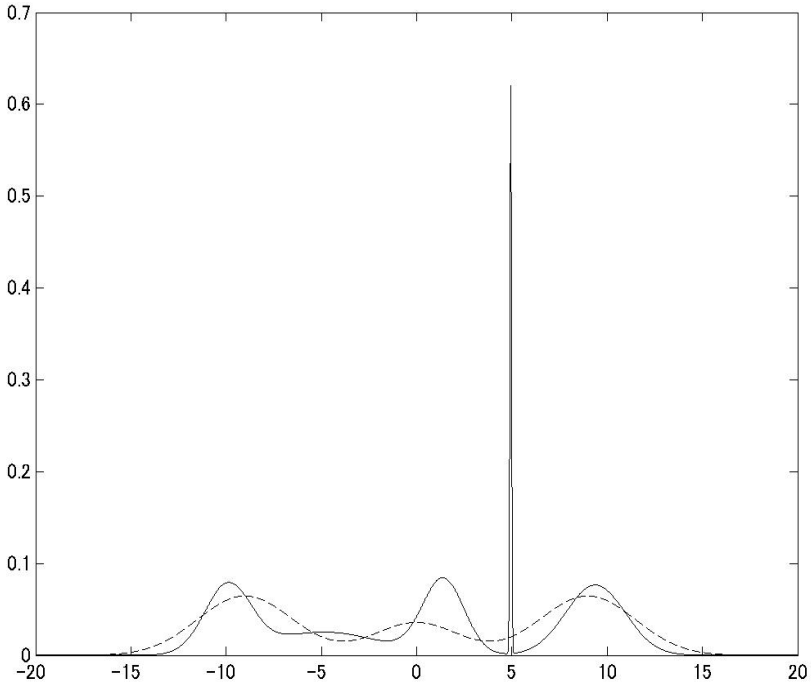

Figure 5.2: Results of No. 3 in Table 5.1

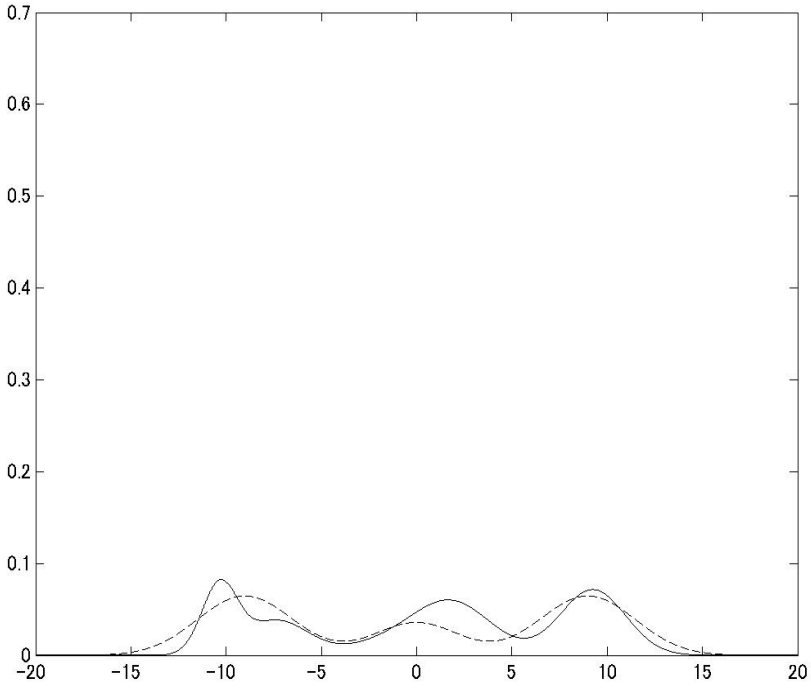

Table 5.3: Results of No. 3 in Table 5.1

	$l_i = 0$ $(i = 1, \ldots, 5)$			$l_i = 0.15$ $(i = 1, \ldots, 5)$		
	α_i	μ_i	Λ_i^{-1}	α_i	μ_i	Λ_i^{-1}
1	0.2233	1.3976	1.1864	0.1500	0.0175	7.8016
2	0.1869	-4.8498	8.7129	0.1526	-7.3207	2.6494
3	0.2232	-9.9129	1.4691	0.1764	-10.3462	0.8697
4	0.3003	9.3869	2.4501	0.3074	9.2639	2.9449
5	0.0663	4.9377	0.0018	0.2136	2.0166	3.9377

Table 5.4: Comparison of log-likelihoods (The dimension is 10.)

		(A) $(\underline{\lambda}_i, \overline{\lambda}_i) = (0, \infty)$		(B) $(\underline{\lambda}_i, \overline{\lambda}_i) = (10^{-3}, 10^3)$	
No.	# of data	Observation	Test	Observation	Test
1	100	$*$	$*$	$*$	$*$
2	200	$*$	$*$	-16.2364	-24.5632
3	300	-17.4153	-22.0561	-17.4153	-22.0561
4	400	-17.7676	-21.1256	-17.7676	-21.1256
5	500	-17.9586	-20.7800	-17.9586	-20.7800
6	600	-18.0046	-20.5797	-18.0046	-20.5797
7	700	-18.1480	-20.2934	-18.1480	-20.2934
8	800	-18.2535	-20.0519	-18.2535	-20.0519
9	900	-18.2848	-19.9797	-18.2848	-19.9797
10	1000	-18.2381	-19.6845	-18.2381	-19.6845

Table 5.5: Comparison of log-likelihoods (The dimension is 30.)

		(A) $(\underline{\lambda}_i, \overline{\lambda}_i) = (0, \infty)$		(B) $(\underline{\lambda}_i, \overline{\lambda}_i) = (10^{-3}, 10^3)$	
No.	# of data	Observation	Test	Observation	Test
1	100	$*$	$*$	$*$	$*$
2	200	$*$	$*$	$*$	$*$
3	300	$*$	$*$	-45.2811	-151.2185
4	400	$*$	$*$	-55.0462	-85.0240
5	500	-58.2170	-76.7472	-58.2170	-76.7472
6	600	-59.6852	-73.3414	-59.6852	-73.3414
7	700	-60.3356	-71.5572	-60.3356	-71.5572
8	800	-60.6195	-70.5857	-60.6195	-70.5857
9	900	-61.0230	-70.0405	-61.0230	-70.0405
10	1000	-61.5174	-69.2069	-61.5174	-69.2069

5.6 Concluding remarks

In this chapter, we presented a BCD method for a general class of maximum likelihood estimation problems for mixture distributions. The general class includes maximum likelihood estimation problems with L_1 regularizations and/or some constraints on parameters. Moreover, we presented efficient implementations of the BCD method for some special problems. In particular, we gave an $O(m)$ solution method for subproblem (5.3.24) when the lower constraints $l_i \leq \alpha_i$ $(i = 1, \ldots, m)$ exist. In addition, we provided an analytical solution for subproblem (5.3.29) with the constraint $\underline{\lambda}_i I \preceq \Lambda_i \preceq \overline{\lambda}_i I$. Finally, we conducted the numerical experiments for the models discussed in Subsections 5.4.1 and 5.4.3. From the experiments, we see that the models with reasonable constraints yield the valid parameter estimations even if the amount of the observational data is small.

As a future work, we are interested in an inexact version of the proposed BCD method. The proposed method requires that subproblems (5.3.24) and (5.3.25) are solved exactly for the globel convergence. It is worth constructing a global convergent BCD method that allows inexact solutions of subproblems (5.3.24) and (5.3.25).

Chapter 6

Conclusion

In this thesis, we studied solution methods for nonlinear SDP. We summarize the results obtained in this thesis.

- In Chapter 3, we proposed a primal-dual interior point method based on the shifted barrier KKT conditions. Since we have to find an approximate shifted barrier KKT point at each iteration of this method, we presented a differentiable merit function F for the shifted barrier KKT point, and proved its three nice properties:

  (i) The merit function F is differentiable;

  (ii) Any stationary point of the merit function F is a shifted barrier KKT point;

  (iii) The level set of the merit function F is bounded under some reasonable assumptions.

  These properties imply that an approximate shifted barrier KKT point can be obtained by minimizing the merit function F. Thus, we also proposed a Newton-type method for minimizing the merit function F, and showed the global convergence of the Newton-type method under some milder assumptions compared with Yamashita, Yabe and Harada [72]. Moreover, we gave some results of numerical experiments for the proposed method, and observed its efficiency.

- In Chapter 4, we presented a two-step primal-dual interior point method based on the generalized shifted barrier KKT conditions. This method has to solve two different Newton equations derived from the generalized shifted barrier KKT conditions at each iteration. However, in order to reduce calculations, we replaced the coefficient matrix in the second equation with that in the first one. Thus, we can solve the second equation more rapidly using some computational results obtained by solving the first equation. Despite this change, we proved the superlinear convergence of the two-step primal-dual interior point method under the same assumptions as Yamashita and Yabe [71]. In addition, we conducted some numerical experiments, and showed that the proposed method can find a solution faster than Yamashita and Yabe's two step method [71].

- In Chapter 5, we studied a maximum likelihood estimation for mixture distributions. In particular, we mainly considered the case where mixture distributions are Gaussian

mixtures. In this case, maximum likelihood estimation problems are expressed as nonlinear SDP. Moreover, we presented a general class of maximum likelihood estimation problems for mixture distributions. It includes maximum likelihood estimation problems with the L_1 regularization and/or some constraints on parameters. We proposed a BCD method for the general class. Since we have to deal with some subproblems generated in the proposed BCD method, we also proposed some efficient solution methods for such subproblems. Finally, we gave some results of numerical experiments for the maximum likelihood estimation problems with some additional constraints on parameters. Then, we observed that such problems yield valid results even if the amount of the observational data is small.

Finally, we discuss some future works.

- We mentioned that the barrier parameter μ_k must satisfy that $\mu_k \to 0$ $(k \to \infty)$ and $\mu_k > 0$, when we proposed Algorithm 3.2.1 in Chapter 3. However, we did not discuss a concrete choice of the barrier parameter μ_k. If we choose μ_k which converges to zero so quickly, it will take a good amount of time to obtain an approximate shifted barrier KKT point by Algorithm 3.3.1, although there are few iteration counts of Algorithm 3.2.1. Conversely, if we choose μ_k which converges to zero so slowly, there will be many iteration counts of Algorithm 3.2.1, although it will not be long before Algorithm 3.3.1 finds an approximate shifted barrier KKT point. Therefore, a future work is to give a reasonable update rule of the barrier parameter μ_k.

- In Chapter 4, we proved the superlinear convergence of Algorithm 4.2.3 which uses scaling and is based on two-step method. However, in the present moment, there is still no proof for a one-step method with scaling. This should be a topic of future research.

- We have some room to improve the BCD method proposed in Chapter 5. The proposed method requires that their subproblems are solved exactly for the global convergence. It is worth constructing a global convergent BCD method that allows inexact solutions of the subproblems.

Bibliography

[1] F. Alizadeh and D. Goldfarb, *Second-order cone programming*, Mathematical Programming, Series B 95 (2003), 3–51.

[2] F. Alizadeh, J. A. Haeberly and M. L. Overton, *Primal-dual interior-point methods for semidefinite programming: convergence rates, stability and numerical results*, SIAM Journal on Optimization, 8 (1998), 746–768.

[3] M. F. Anjos and J. B. Lasserre, *Handbook on semidefinite, conic and polynomial optimization*, Springer Science+Business Media, 2012.

[4] A. Ben-Tal, F. Jarre, M. Kočvara, A. Nemirovski and J. Zowe, *Optimal design of trusses under a nonconveex global buckling constraint*, Optimization and Engineering, 1 (2000), 189–213.

[5] D. S. Bernstein, *Matrix Mathematics*, Princeton University Press, 2009.

[6] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, 1995.

[7] D. P. Bertsekas, A. Nedić and A. E. Ozdaglar, *Convex Analysis and Optimization*, Athena Scientific, 2003.

[8] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer Science+Business Media, 2006.

[9] M. Blum, R. W. Floyd, V. Pratt, R. L. Rivest and R. E. Tarjan, *Time bounds for selection, Journal of Computer and System Sciences*, 7 (1973), 448–461.

[10] S. Boyd, L. E. Ghaoui, E. Feron and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, Society for Industrial and Applied Mathematics, 1994.

[11] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[12] P. Brucker, *An $O(n)$ algorithm for quadratic knapsack problems*, Operations Research Letters, 3 (1984), 163–166.

[13] R. Correa and C. H. Ramírez, *A global algorithm for nonlinear semidefinite programming*, SIAM Journal on Optimization, 15 (2004), 303–318.

[14] D. Crisan and B. Rozovskii, *The Oxford Handbook of Nonlinear Filtering*, Oxford University Press, 2011.

[15] G. B. Dantzig, *Linear Programming and Extensions*, Princeton University Press, 1963.

[16] A. P. Dempster, N. M. Laird and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society, 39 (1977), 1–38.

[17] B. Fares, D. Noll and P. Apkarian, *Robust control via sequential semidefinite programming*, SIAM Journal on Control and Optimization, 40 (2002), 1791–1820.

[18] A. Forsgren and P. E. Gill, *Primal-dual interior methods for nonconvex nonlinear programming*, SIAM Journal on Optimization, 8 (1998), 1132–1152.

[19] R. W. Freund, F. Jarre and C. H. Vogelbusch, *Nonlinear semidefinite programming: sensitivity, convergence, and an application in passive reduced-order modeling*, Mathematical Programming, Series B 109 (2007), 581–611.

[20] J. Friedman, T. Hastie and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso*, Biostatistics, 9 (2008), 432–441.

[21] M. Fukuda and M. Kojima, *Branch-and-cut algorithms for the bilinear matrix inequality eigenvalue problem*, Computational Optimization and Applications, 19 (2001), 79–105.

[22] W. Gómez and C. H. Ramírez, *A filter algorithm for nonlinear semidefinite programming*, Computational and Applied Mathematics, 29 (2010), 297–328.

[23] K. C. Goh, M. G. Safonov and G. P. Papavassilopoulos, *Global optimization for the biaffine matrix inequality problem*, Journal of Global Optimization, 7 (1995), 365–380.

[24] G. Golub, and L. C. Van, *Matrix Computations*, The Johns Hopkins University Press, 1989.

[25] I. Grubišić and R. Pietersz, *Efficient rank reduction of correlation matrices*, Linear Algebra and its Applications, 422 (2007), 629–563.

[26] R. J. Hathaway, *Another interpretation of the EM algorithm for mixture distributions*, Statistics & Probability Letters, 4 (1986), 53–56.

[27] R. T. Haftka and Z. Gürdal, *Elements of Structural Optimization*, Springer Netherlands, 1992.

[28] C. Helmberg, F. Rendl, R. J. Vanderbei and H. Wolkowicz, *An interior-point method for semidefinite programming*, SIAM Journal on Optimization, 6 (1996), 342–361.

[29] C. W. J. Hol, C. W. Scherer, E. G. Van der Meche and O. H. Bosgra, *A nonlinear SDP approach to fixed-order controller synthesis and comparison with two other methods applied to an active suspension system*, European Journal of Control, 9 (2003), 13–28.

[30] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.

[31] X. X. Huang, X. Q. Yang and K. L. Teo, *Lower-order penalization approach to nonlinear semidefinite programming*, Journal of Optimization Theory and Applications, 132 (2007), 1–20.

[32] F. Jarre, *An interior method for nonconvex semidefinite programs*, Optimization and Engineering, 1 (2000). 347–372.

[33] C. Kanzow, C. Nagel, H. Kato and M. Fukushima, *Successive linearization methods for nonlinear semidefinite programs*, Computational Optimization and Applications, 31 (2005), 251–273.

[34] A. Kato, H. Yabe and H. Yamashita, *An interior point method with a primal-dual quadratic barrier penalty function for nonlinear semidefinite programming*, Journal of Computational and Applied Mathematics, 275 (2015), 148–161.

[35] M. Kojima, S. Shindoh and S. Hara, *Interior-point methods for the monotone semidefinite linear complementarity problem in symmetric matrices*, SIAM Journal on Optimization, 7 (1997), 86–125.

[36] F. Leibfritz and E. M. E. Mostafa, *An interior point constrained trust region method for a special class of nonlinear semidefinite programming problems*, SIAM Journal on Optimization, 12 (2002), 1048–1074.

[37] Q. Li and H.-D. Q, *A sequential semismooth Newton method for the nearest low-rank correlation matrix problem*, SIAM journal on optimization, 21 (2011), 1641–1666.

[38] L. Li and K. C. Toh, *An inexact interior point method for L_1-regularized sparse covariance selection*, Mathematical Programming Computation, 2 (2010), 291–315.

[39] Z. Lu, *Adaptive first-order methods for general sparse inverse covariance selection*, SIAM Journal on Matrix Analysis and Applications, 31 (2010), 2000–2016.

[40] Z.-Q. Luo, J. F. Sturm and S. Zhang, *Superlinear convergence of a symmetric primal-dual path following algorithm for semidefinite programming*, SIAM Journal on Optimization, 8 (1998), 59–81.

[41] H. Z. Luo, H. X. Wu and G. T. Chen, *On the convergence of augmented Lagrangian methods for nonlinear semidefinite programming*, Journal of Global Optimization, 54 (2012), 599–618.

[42] R. D. C. Monteiro, *Primal-dual path-following algorithms for semidefinite programming*, SIAM Journal on Optimization, 7 (1997), 663–678.

[43] R. B. Millar, *Maximum Likelihood Estimation and Inference*, John Wiley & Sons, 2011.

[44] Y. E. Nesterov and M. J. Todd, *Self-scaled barriers and interior-point methods for convex programming*, Mathematics of Operations Research, 22 (1997), 1–42 .

[45] Y. E. Nesterov and M. J. Todd, *Primal-dual interior-point methods for self-scaled cones*, SIAM Journal on Optimization, 8 (1998), 324–364.

[46] J. Ortega and W. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, 1970.

[47] P. M. Pardalos, N. Kovoor, *An algorithm for a singly constrained class of quadratic programs subject to upper and lower bounds*, Mathematical Programming, Series B 46 (1990), 321–328.

[48] F. A. Potra and R. Sheng, *A superlinearly convergent primal-dual infeasible-interior-point algorithm for semidefinite programming*, SIAM Journal on Optimization, 8 (1998), 1007–1028.

[49] H. Qi, *Local duality of nonlinear semidefinite programming*, Mathematics of Operations Research, 34 (2009), 124–141.

[50] T. D. Quoc, S. Gumussoy, W. Michiels and M. Diehl, *Combining convex-concave decompositions and linearization approaches for solving BMIs, with application to static output feedback*, IEEE Transactions on Automatic Control, 57 (2012), 1377–1390.

[51] R. A. Redner and H. F. Walker, *Mixture densities, maximum likelihood and the EM algorithm*, SIAM Review, 26 (1984), 195–239.

[52] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.

[53] L. Ruan, M. Yuan and H. Zou, *Regularized parameter estimation in high-dimensional Gaussian mixture models*, Neural Computation, 23 (2011), 1605–1622.

[54] G. Stewart, *Matrix Algorithms*, Society for Industrial and Applied Mathematics, 1998.

[55] M, Stingl, *On the solution of nonlinear semidefinite programs by augmented Lagrangian methods*, Doctoral Thesis, University of Erlangen, (2005).

[56] M. Stingl, M. Kocvara and G. Leugering, *A new non-linear semidefinite programming algorithm with an application to multidisciplinary free material optimization*, International Series of Numerical Mathematics, 158 (2009), 275–295.

[57] D. Sun, *The strong second-order sufficient condition and constraint nondegeneracy in nonlinear semidefinite programming and their implications*, Mathematics of Operations Research, 31 (2006), 761–776.

[58] D. Sun, J. Sun and L. Zhang, *The rate of convergence of the augmented Lagrangian method for nonlinear semidefinite programming*, Mathematical Programming, Series A 114 (2008), 349–391.

[59] J. B. Thevenet, P. Apkarian and D. Noll, *Reduced-order output feedback control design with specSDP, a code for linear/nonlinear SDP problems*, International Conference and Automation, 1 (2005), 465–470.

[60] M. J. Todd, *Semidefinite optimization*, Acta Numerica, 10 (2001), 515–560.

[61] M. J. Todd, K. C. Toh and R. H. Tütüncü, *On the Nesterov-Todd direction in semidefinite programming*, SIAM Journal on Optimization, 8 (1998), 769–796.

[62] P. Tseng, *Convergence of a block coordinate descent method for nondifferentiable minimization, Journal of Optimization Theory and Applications*, 109 (2001), 475-494.

[63] J. G. VanAntwerp and R. D. Braatz, *A tutorial on linear and bilinear matrix inequality*, Journal of Process Control, 10 (2000), 363–385.

[64] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Review, 38 (1996), 49–95.

[65] L. Vandenberghe, S. Boyd and S. P. Wu, *Determinant maximization with linear matrix inequality constraints*, SIAM Journal on Matrix Analysis and Applications, 19 (1998), 499–533.

[66] H.-T. Wai, W.-K. Ma and A. M.-C. So, *Cheap semidefinite relaxation MIMO detection using row-by-row block coordinate descent*, Proceedings of IIIE International Conference on Acoustics, Speech and Signal Processing, (2011), 3256–3259.

[67] L. Xu and M. I. Jordan, *On convergence properties of the EM algorithm for Gaussian mixtures*, Neural Computation, 8 (1996), 129–151.

[68] Y. Yamakawa and N. Yamashita, *Differentiable merit function for shifted perturbed Karush-Kuhn-Tucker conditions of nonlinear semidefinite programming*, Pacific Journal of Optimization, to appear.

[69] Y. Yamakawa and N. Yamashita, *A two-step primal-dual interior point method for nonlinear semidefinite programming problems and its superlinear convergence*, Journal of the Operations Research Society of Japan, 57 (2014), 105–127.

[70] H. Yamashita, *A globally convergent primal-dual interior point method for constrained optimization*, Optimization Methods and Software, 10 (1998), 443–469.

[71] H. Yamashita and H. Yabe, *Local and superlinear convergence of a primal-dual interior point method for nonlinear semidefinite programming*, Mathematical Programming, Series A 132 (2012), 1–30.

[72] H. Yamashita, H. Yabe and K. Harada, *A primal-dual interior point method for nonlinear semidefinite programming*, Mathematical Programming, Series A 135 (2012), 89–121.

[73] X. Yuan, *Alternating direction method for covariance selection models*, Journal of Scientific Computing, 51 (2012), 261–273.

[74] Z. Zhang and L. Wu, *Optimal low-rank approximation to a correlation matrix*, Linear Algebra and its Applications, 364 (2003), 161–187.