On Polynomial-time Path-following Interior-point Methods with Local Superlinear Convergence

by

Shuxin Zhang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Combinatorics & Optimization

Waterloo, Ontario, Canada, 2016

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Interior-point methods provide one of the most popular ways of solving convex optimization problems. Two advantages of modern interior-point methods over other approaches are:

(i) robust global convergence, and

(ii) the ability to obtain high accuracy solutions in theory (and in practice, if the algorithms are properly implemented, and as long as numerical linear system solvers continue to provide high accuracy solutions)

for well-posed problem instances. This second ability is typically demonstrated by asymptotic superlinear convergence properties.
In this thesis, we study superlinear convergence properties of interior-point methods with proven polynomial iteration complexity. Our focus is on linear programming and semidefinite programming special cases. We provide a survey on polynomial iteration complexity interior-point methods which also achieve asymptotic superlinear convergence. We analyze the elements of superlinear convergence proofs for a dual interior-point algorithm of Nesterov and Tunçel and a primal-dual interior-point algorithm of Mizuno, Todd and Ye. We present the results of our computational experiments which observe and track superlinear convergence for a variant of Nesterov and Tunçel's algorithm.

Acknowledgements

I would like to thank all people who helped me with the preparation and the writing of this thesis.

First and foremost, I would like to thank my supervisor, Prof. Levent Tunçel, for his outstanding guidance, patience and support throughout the preparation of this thesis. Without his dedication and encouragement, this thesis would not have been possible.

The material in this thesis is based upon research supported in part by NSERC Discovery Grants, William Tutte Postgraduate Scholarship, U.S. Office of Naval Research under award number: N00014-15-1-2171. This financial support is gratefully acknowledged.

I would like to thank Prof. Stephen Vavasis and Prof. Henry Wolkowicz for taking their time reviewing the thesis and providing precious comments.

Last but not least, I would like to thank my family and my friends for their love and support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Optimization is the area that deals with minimizing or maximizing an objective function subject to some constraints. One part of optimization may been seen as designing tools for finding better solutions by using mathematical analysis. Optimization is very useful in various kinds of decision making problems, such as finding the route that takes least time when you travel between two different locations and constructing a portfolio which maximizes expected return while keeping the risk at a low level. As a result, a large variety of real-world problems can be modelled as continuous optimization problems. In this thesis, we are interested in polynomial-time interior-point algorithms for Linear Optimization and generalizations of these algorithms to Semidefinite Optimization and convex optimization problems. For such algorithms, our main focus is on achieving superlinear and quadratic convergence asymptotically, while maintaining a global polynomial iteration bound.

Convex optimization problems (minimization of a convex function over a convex set) have a huge advantage over nonconvex problems due to their special structure. One such crucial property is, in convex problems, local optimality gives global optimality. In addition, in terms of the existence of faster and more efficient algorithms to find the optimal solutions, convex optimization problems tend to behave better than the nonconvex ones among all continuous optimization problems. In this thesis, we will mainly study two classes of convex optimization problems which admit fast and efficient algorithms: Linear Programs and Semidefinite Programs. Linear programs minimize or maximize a linear objective function over a convex set (called the *feasible region*) which is polyhedral. As for Semidefinite Programs, the objective function is linear and the feasible region is the intersection of the set of positive semidefinite matrices with an affine space.

The Simplex Method was developed by George Dantzig in the 1940's and it is one

of the most popular methods to solve Linear Programs (see [12]). However, no variant of the Simplex Method has been proven to run in polynomial time and many variants of the method have corresponding worst-case examples proving that they are exponential time algorithms in the worst-case. In 1979, Khachian [23] showed that Linear Programs can be solved in polynomial time by applying ellipsoid method. In 1984, Karmarkar [22] showed that an interior-point method solves Linear Programs in polynomial time. There exist software packages to solve Linear Programs (for example, CPLEX[2], Gurobi[6], MOSEK[7] and GLPK[5]) and Semidefinite Programs (for example, SeDuMi[10], SDPA[8], CSDP[3], DSDP[4] and SDPT3[9]), so we are interested in finding faster and more efficient algorithms to solve those two classes of optimization programs. Therefore, this explains why we might be interested in such problems and algorithms.

Newton's method is an iterative method to find roots of a differentiable function, and it is a popular method we use to solve nonlinear optimization problems. A mathematical statement of quadratic convergence of Newton's method was first proved by Kantorovich [20] in 1948. In 1986, Smale [37] proved a theorem on quadratic convergence of Newton's method applied to analytic functions that use only information at the starting point.

Next, we explain the intrinsic challenges of designing and theoretically analyzing an interior-point method that is both polynomial time and is asymptotically superlinearly convergent.

Let the *central path* denote a smooth curve that is in the interior of the feasible region converging to an optimal solution at one end, and converging to a central point (possibly at infinity) at the other end. Let the region between the thick curves in the Figure 1.4 denote a neighbourhood of the central path. The central path is defined as the solution set of a system of nonlinear equations and inequalities. These vague statements are just for demonstrating ideas, and the underlying concepts will be formally defined in Chapter 3.

From Kantorovich's theory presented in Section 2.2, we know that if the objective function is nice and smooth, there is a small neighbourhood of the optimal solution such that Newton's Method converges quadratically in it. However, whether we can always generate iterates in that small neighbourhood in a constrained convex optimization setting is not guaranteed. We will use the following figures to illustrate the underlying difficulties.

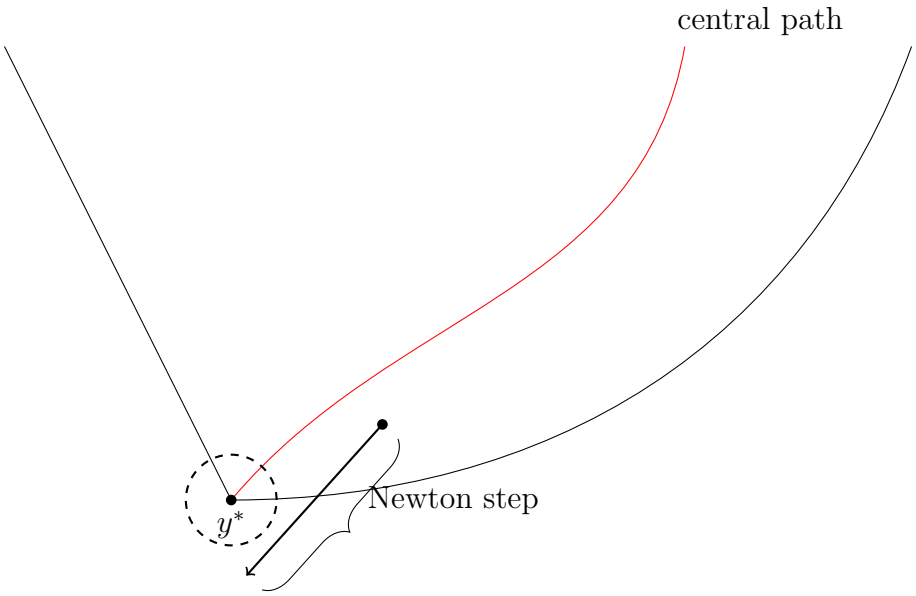


Figure 1.1: The unlucky case where the next iterate is not in the small neighbourhood

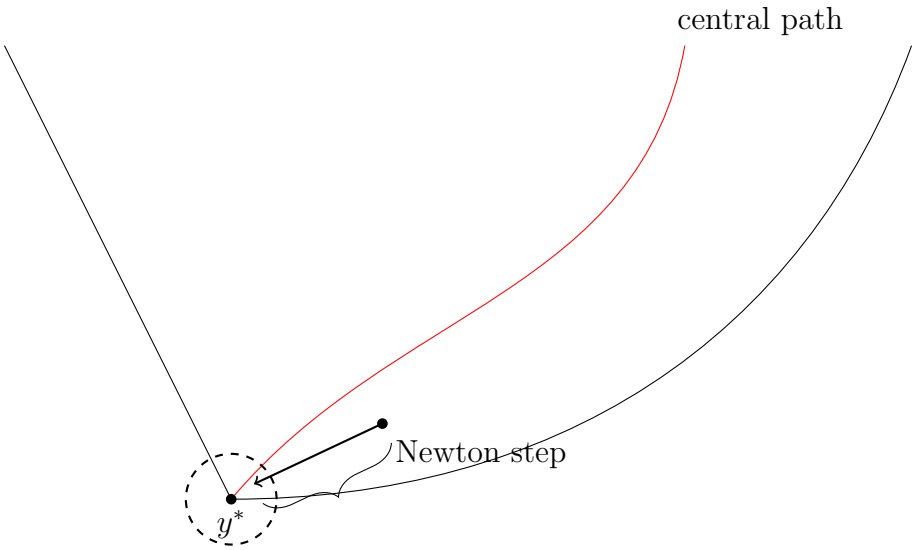


Figure 1.2: The lucky case where the next iterate is in the small neighbourhood

The above Figure 1.1 and Figure 1.2 demonstrate the difficulty of achieving asymptotically superlinear convergence since we do not always have the lucky case where the next

iterate is in the small neighbourhood and in the domain of interior-point methods. Next, we will present the difficulty in maintaining polynomial time interior-point method. The circles representing the small neighbourhood are blown up pictures in the following Figure 1.3 and Figure 1.4.

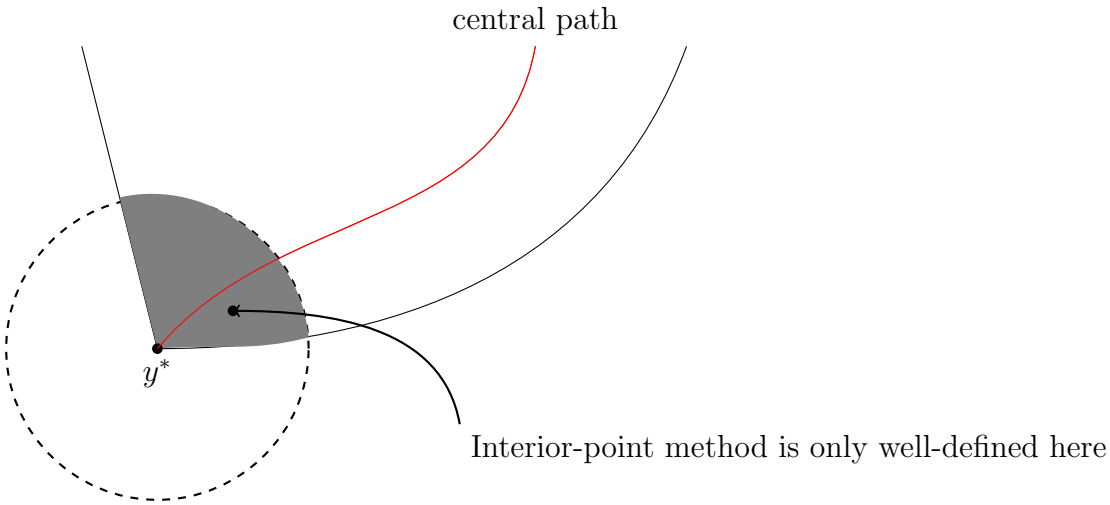


Figure 1.3: The region in the neighbourhood of the optimal solution where interior-point method is well-defined

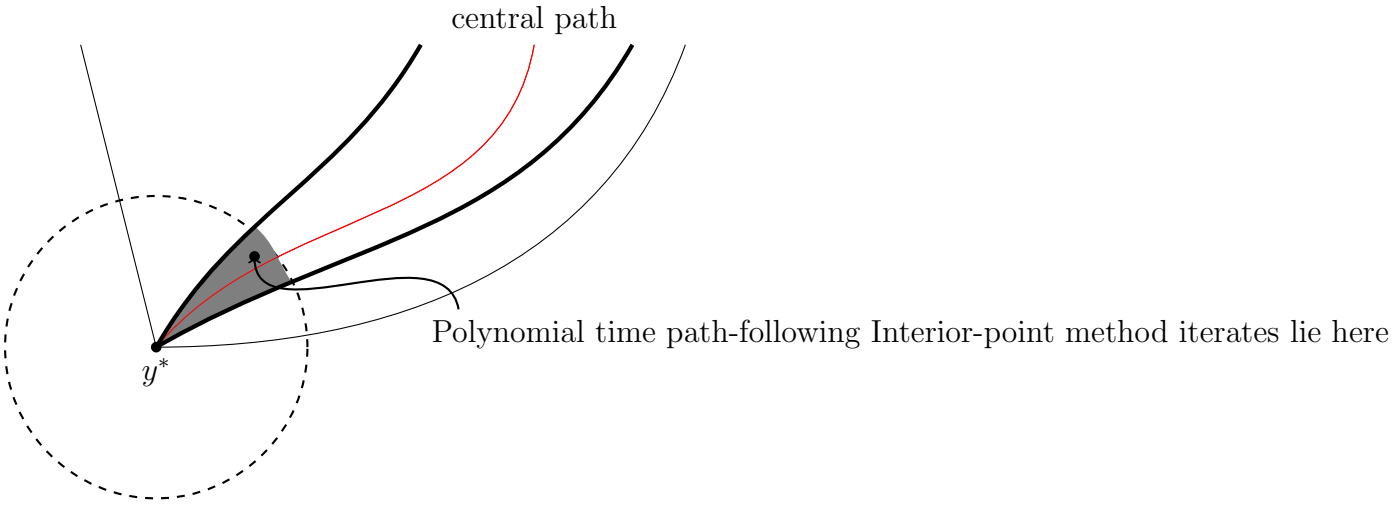


Figure 1.4: The region where polynomial time path-following interior-point method iterates must lie

Nesterov and Tunçel [30] proposed a new polynomial-time path-following predictor-corrector interior-point method for general conic optimization problems and established the local superlinear convergence property of this method. It motivates us to investigate the superlinear and quadratic convergence property of this method in the case of Linear Programming and Semidefinite Programming, since for these special cases, the method may achieve stronger convergence properties, perhaps even under weaker assumptions.

The rest of the thesis will be organized as follows. In Chapter 2 we introduce different notions of the rate of convergence and present Kantorovich's theory and Smale's theorem on the superlinear convergence in continuous optimization. In Chapter 3, we introduce a pair of primal-dual problems in the cases of Linear Programming, Semidefinite Programming and the general convex optimization case. After that, we define the concept of central path and the notion of strict complementarity in Linear Programming case and Semidefinite Programming case, and present a literature survey on superlinear convergence in polynomial iteration complexity interior-point methods. Then, in Chapter 4, we demonstrate the results on superlinear and quadratic convergence in various modern, primal-dual interior-point methods in the cases of Linear Programming and Semidefinite Programming.

In Chapter 5, we analyze the superlinear convergence of an algorithm of Nesterov and Tunçel in the special case of Linear Programming. We explore possible ways of relaxing the assumptions needed in the general convex optimization setting, in the special cases of Linear Programming and Semidefinite Programming. In the Linear Programming case, we also investigate a less conservative variant of the algorithm that does not shrink the size of the neighbourhood of the central path proportionally to the duality gap (when the duality gap is small). We study the fundamental elements of superlinear/quadratic convergence results for Linear Programming with a focus on primal-dual elements. After this theoretical material, in Chapter 6, we show the results of some numerical experiments of this algorithm to justify its superlinear convergence. Finally, in Chapter 7, we draw conclusions from the whole thesis and leave directions for future research.

Chapter 2

Superlinear Convergence in Continuous Optimization

In continuous optimization, the variables are allowed to take on values that are in a continuous range, usually real numbers, and satisfy the given constraints. In contrast, variables are restricted to binary or integer values in discrete optimization. This continuous nature of continuous optimization allows us to define the first and the second derivatives of the functions we want to minimize or maximize in the problem as well as those functions defining the constraints.

Typically, continuous optimization problems are solved using specific types of algorithms which generate iterates, defined as a sequence of values of the variables, that converge to an optimal solution of the problem. These algorithms start from an initial point and apply, and then recursively generate the next iteration based on the information about the current iteration. Therefore, what matters a lot in continuous optimization algorithms are convergence and speed of convergence. We not only need the global convergence of the iteration sequences generated by the algorithms, but also hope to establish the fast local convergence properties, which guarantee the ability to converge fast to a solution whenever the current iterate is close enough to such a point. Newton's Method is one of the most commonly used such iterative methods to solve systems of nonlinear equations.

In this chapter, we will first define some terminologies that we use to talk about convergence speed and rate. Then, we will survey some details of the theory of Newton's method and its convergence properties.

2.1 R-convergence,Q-convergence and superlinear convergence rates

Usually, we need to adopt some iterative numerical methods to solve continuous optimization problems. In order to judge the performance of the algorithm, we need some measures to examine how well suited the algorithm is to the specific optimization problem. One important measure of speed of convergence of an algorithm is its order of convergence.

In this section, we will follow the notations used in [32] and [34]. We will introduce some different notions related to R-order and Q-order of convergence.

First, we will define R-order and Q-order of convergence. Let $\{\epsilon_n\}$ be a sequence of positive real numbers converging to zero.

Definition 2.1.1. *We say that the sequence* $\{\epsilon_n\}$ *converges with* Q-order *at least* $\tau > 1$ *if there is a constant* d *such that*

$$\epsilon_{n+1} \leq d\epsilon_n^\tau, \quad n = 0, 1, 2, \ldots$$

We say that the sequence $\{\epsilon_n\}$ *has the exact* Q-order *of convergence* τ *if there are two positive constants* a, d *such that*

$$a\epsilon_n^\tau \leq \epsilon_{n+1} \leq d\epsilon_n^\tau, \quad n = 0, 1, 2, \ldots$$

From the above definitions, we can see that if $\{\epsilon_n\}$ converges with Q-order at least τ, then it implies that $\{\epsilon_n\}$ converges with Q-order at least τ_1 such that $0 < \tau_1 \leq \tau$. So, the notion of Q-order of convergence does not give a unique characterization of the speed of convergence. In contrast, it is easily seen that the notion of exact Q-order of convergence gives a unique characterization, if it exists. Moreover, if we have a sequence $\{\epsilon_n\}$ such that it converges with Q-order at least $\tau > 1$, then we can attach a unique quantity to it to represent its order of convergence.

Definition 2.1.2. *The* Q-order *of the sequence* $\{\epsilon_n\}$, $Q_{\{\epsilon_n\}}$, *is defined as follows:*

$$Q_{\{\epsilon_n\}} := \sup\{\tau > 1 : \{\epsilon_n\} \text{ converges with } Q\text{-order at least } \tau\}.$$

Note that if $\{\epsilon_n\}$ has an exact Q-order of convergence τ, then we have that $Q_{\{\epsilon_n\}} = \tau$. We use Q in Q-order of convergence since Q stands for quotient and we compute the limit of quotient $\left\{\frac{\epsilon_{n+1}}{\epsilon_n}\right\}$ in the definition. Similarly, we will define another way to measure the speed of convergence: the following notion R-order of convergence, where R stands for roots.

Definition 2.1.3. *We say that the sequence* $\{\epsilon_n\}$ *converges with* R-order *at least* $\tau > 1$ *if there are constants* $d > 0$, $\theta \in (0,1)$ *such that*

$$\epsilon_n \leq d\theta^{\tau^n}, \;\; n = 0, 1, 2, \ldots$$

We say that the sequence $\{\epsilon_n\}$ *has the exact* R-order *of convergence* τ *if there are constants* a, $d > 0$, $\theta, \eta \in (0,1)$*such that*

$$a\eta^{\tau^n} \leq \epsilon_n \leq d\theta^{\tau^n}, \;\; n = 0, 1, 2, \ldots$$

Definition 2.1.4. *The* R-order *of the sequence* $\{\epsilon_n\}$, $R_{\{\epsilon_n\}}$, *is defined as follows:*

$$R_{\{\epsilon_n\}} := \sup\{\tau > 1 : \{\epsilon_n\} \text{ converges with } R\text{-order at least } \tau\}.$$

Similar to Q-order of convergence, statements can be derived for R-order of convergence. R-order of convergence does not uniquely characterize the speed of convergence, while the exact R-order of convergence is unique. Moreover, if $\{\epsilon_n\}$ has an exact R-order of convergence τ, then we have that $R_{\{\epsilon_n\}} = \tau$.

The following Proposition gives us the differences between Q-order and R-order of convergence.

Proposition 2.1.5. *(Proposition 1.3 in [32])*

1. *If* $\{\epsilon_n\}$ *converges with* Q*-order at least* τ*, then* $\{\epsilon_n\}$ *converges with* R*-order at least* τ*.*
2. *If* $\{\epsilon_n\}$ *has exact* Q*-order of convergence* τ*, then* $\{\epsilon_n\}$ *has exact* R*-order of convergence* τ*.*
3. $R_{\{\epsilon_n\}} \geq Q_{\{\epsilon_n\}}$

Moreover, note that none of the reverse statements are true. For example ([32]), suppose that we are given $\theta \in (0,1)$ and s, τ such that $1 < s < \tau$. Take $c = \theta^q$ with $q > 1$ such that $qs > \tau$. Define

$$\epsilon_n = \begin{cases} \theta^{\tau^n} & \text{if } n \text{ is odd,} \\ c^{\tau^n} & \text{if } n \text{ is even.} \end{cases}$$

Then, by construction, $\{\epsilon_n\}$ has the exact R-order of convergence τ. However, for Q-order of convergence, we know that, for n even

$$\frac{\epsilon_{n+1}}{\epsilon_n^s} = \frac{(\theta^\tau)^{\tau^n}}{(c^s)^{\tau^n}} = \left(\frac{\theta^\tau}{\theta^{qs}}\right) \to \infty \;\text{ as }\; n \;\to \infty.$$

Therefore, $Q_{\{\epsilon_n\}} \leq s < r$.

In predictor-corrector type interior-point methods, choosing different notions of order of convergence may make a big difference. For instance, let us consider the sequence of μ generated by the dual path-following algorithm in [30].

$$\mu_{seq1} := \{\mu_k : \ \mu_k \text{ obtained from a predictor step or a corrector step of the algorithm}\}.$$

On the other hand, consider another sequence

$$\mu_{seq2} := \{\mu_k : \ \mu_k \text{ obtained from a predictor step of the algorithm}\}.$$

Notice that from each predictor step to its following corrector step, the value of μ_k does not change. If we want to use Q-order of convergence, then μ_{seq1} converges to zero Q-linearly. However, if we use R-order of convergence, we can say that μ_{seq1} converges to zero R-superlinearly. Instead, if we use μ_{seq2}, then we can conclude Q-superlinear and R-superlinear convergence result for it.

Now, we are interested in finding a full characterization of Q-order and R-order of convergence. According to [32], by using the definitions of $Q_{\{\epsilon_n\}}$ and properties of $\liminf_{n\to\infty}\left(\frac{\log \epsilon_{n+1}}{\log \epsilon_n}\right)$, we are able to show the two following propositions.

Proposition 2.1.6. *(Proposition 1.1 in [32]) Let $\{\epsilon_n\}$ be a sequence of positive real numbers converging to zero and $\tau > 1$. Then $Q_{\{\epsilon_n\}} = \tau$ if and only if*

$$\tau = \lim_{n\to\infty} \inf \left(\frac{\log \epsilon_{n+1}}{\log \epsilon_n}\right)$$

Proposition 2.1.7. *(Proposition 1.2 in [32]) Let $\{\epsilon_n\}$ be a sequence of positive real numbers converging to zero and $\tau > 1$. Then $R_{\{\epsilon_n\}} = \tau$ if and only if*

$$\tau = \lim_{n\to\infty} \inf |\log \epsilon_n|^{\frac{1}{n}}$$

Then, we are ready to give definitions of Q-linear and R-linear convergence.

Definition 2.1.8. *Let $\{\epsilon_n\}$ be a sequence of positive real numbers. We say that $\{\epsilon_n\}$ converges Q-*linearly *if*

$$\lim_{n\to\infty} \sup \left(\frac{\epsilon_{n+1}}{\epsilon_n}\right) < 1.$$

Definition 2.1.9. *Let $\{\epsilon_n\}$ be a sequence of positive real numbers. We say that $\{\epsilon_n\}$ converges* R-linearly *if*

$$\limsup_{n\to\infty}(\epsilon_n)^{\frac{1}{n}} < 1.$$

Here, note that if the sequence $\{\epsilon_n\}$ satisfied any of the above conditions, we can derive that $\{\epsilon_n\}$ converges to zero. Moreover, we know that the Q-linear convergence of $\{\epsilon_n\}$ implies the R-linear convergence of $\{\epsilon_n\}$ because

$$\limsup_{n\to\infty}(\epsilon_n)^{\frac{1}{n}} \leq \limsup_{n\to\infty}\left(\frac{\epsilon_{n+1}}{\epsilon_n}\right).$$

Similarly, we define the Q-superlinear and R-superlinear convergence.

Definition 2.1.10. *Let $\{\epsilon_n\}$ be a sequence of positive real numbers. We say that $\{\epsilon_n\}$ converges* Q-superlinearly *if*

$$\lim_{n\to\infty}\left(\frac{\epsilon_{n+1}}{\epsilon_n}\right) = 0.$$

Definition 2.1.11. *Let $\{\epsilon_n\}$ be a sequence of positive real numbers. We say that $\{\epsilon_n\}$ converges* R-superlinearly *if*

$$\lim_{n\to\infty}(\epsilon_n)^{\frac{1}{n}} = 0.$$

Notice that from the above four definitions, it is easily seen that Q-superlinear convergence implies Q-linear convergence and R-superlinear convergence implies R-linear convergence. Furthermore, R-superlinear convergence is immediately obtained if we have Q-superlinear convergence.

Next, we will give the definitions of Q-quadratic and R-quadratic convergence.

Definition 2.1.12. *Let $\{\epsilon_n\}$ be a sequence of positive real numbers converging to zero. We say that $\{\epsilon_n\}$ converges* Q-quadratically *if $\{\epsilon_n\}$ converges with Q-order at least 2.*

Definition 2.1.13. *Let $\{\epsilon_n\}$ be a sequence of positive real numbers converging to zero. We say that $\{\epsilon_n\}$ converges* R-quadratically *if $\{\epsilon_n\}$ converges with R-order at least 2.*

In this thesis, we may be concerned with sequences $\left\{x^{(n)}\right\}$ in $\mathbb{R}^n$ that converges to a point x^*. Then, by definition, the convergence property of the sequence $\{x^{(n)}\}$ is the same as the convergence property of the sequence $\left\{\left|\left|x^{(n)} - x^*\right|\right|\right\}$. For example, we say that $\left\{x^{(n)}\right\}$ converges Q-quadratically if there is a constant q_2 such that

$$\left|\left|x^{(n+1)} - x^*\right|\right| \leq q_2 \cdot \left|\left|x^{(n)} - x^*\right|\right|^2, \forall n \in \mathbb{N}.$$

2.2 Details of Kantorovich's theory and Smale's Theorem

The Kantorovich Theorem is a theorem on the convergence of the Newton's method and it is part of a fundamental theory in optimization, as well as in numerical analysis. By Newton's method, we can generate a sequence of points which under certain conditions will converge to a solution x of the equation $f(x) = 0$, where $f : \mathbb{R} \to \mathbb{R}$ is a scalar function. Also, we can use Newton's method to find the zeroes of a continuously differentiable function: $F : \mathbb{R}^n \to \mathbb{R}^n$, which is equivalent to solve a system of n (nonlinear) equations.

Sufficient conditions for the existence and uniqueness of the solutions of system of nonlinear equations in Banach spaces were provided by Kantorovich Theorem in [20]. He also showed that under those conditions, the sequences generated by Newton's Method converge to the solution which is close to the initial point. This theorem also has a lot of theoretical and practical applications, such as, finding optimal bounds for iterative methods and constructing a path-following algorithm for linear complementarity problems. Before stating the theorem, we will introduce some useful notations.

Let X be a Banach space. Then the open and closed ball at $x \in X$ are denoted by

$$B(x, r) = \{y \in X : ||x - u|| < r\} \text{ and } B[x, r] = \{y \in X : ||x - u|| \leq r\}$$

respectively. We will use F' to denote the Frechet derivative of a mapping F.

Theorem 2.2.1. *(Kantorovich Theorem (page 170 in [20])) Let X, Y be Banach spaces, $C \subseteq X$ and let $F : C \to Y$ be a continuous function that is continuously differentiable on* $\text{int}(C)$. *Take $x^{(0)} \in \text{int}(C), L, b > 0$ and suppose that*

1. *$F'\left(x^{(0)}\right)$ is invertible,*
2. *$\left|\left|F'\left(x^{(0)}\right)^{-1}\left[F'(y) - F'(x)\right]\right|\right| \leq L\,||x - y||\,,$ for any $x, y \in C$,*
3. *$\left|\left|F'\left(x^{(0)}\right)^{-1} F\left(x^{(0)}\right)\right|\right| \leq b$,*
4. *$2bL \leq 1$.*

Define

$$t_* := \frac{1 - \sqrt{1 - 2bL}}{L}, \quad t_{**} := \frac{1 + \sqrt{1 - 2bL}}{L}.$$

If

$$B\left[x_0, t_*\right] \subset C,$$

then the sequence $\{x_k\}$ *generated by Newton's Method for solving* $F(x) = 0$ *with starting point* $x^{(0)}$,

$$x^{(k+1)} = x^{(k)} - F'\left(x^{(k)}\right)^{-1} F\left(x^{(k)}\right), \quad k = 0, 1, 2, \ldots.$$

is well-defined, is contained in $B(x^{(0)}, t_*)$, *converges to a point* $x_* \in B\left[x_0, t_*\right]$ *which is the unique zero of* F *in* $B\left[x_0, t_*\right]$ *and*

$$\left|\left|x_* - x^{(k+1)}\right|\right| \leq \frac{1}{2}\left|\left|x_* - x^{(k)}\right|\right|, \quad k = 0, 1, 2, \ldots.$$

Moreover, if assumption 4 holds with a strict inequality, i.e. $2bL < 1$, *then*

$$\left|\left|x_* - x^{(k+1)}\right|\right| \leq \frac{1 - \theta^{2^k}}{1 + \theta^{2^k}} \frac{L}{2\sqrt{1 - 2bL}}\left|\left|x_* - x^{(k)}\right|\right|^2 \leq \frac{L}{2\sqrt{1 - 2bL}}\left|\left|x_* - x^{(k)}\right|\right|^2, \quad k = 0, 1, 2, \ldots.$$

where $\theta := \frac{t_*}{t_{**}} < 1$, *and* x_* *is the unique zero of* F *in* $B[x^{(0)}, \rho]$ *for any* ρ *such that*

$$t_* \leq \rho \leq t_{**}, \quad B[x^{(0)}, \rho] \subset C.$$

Notice that if we only have assumptions 1-4, then the sequence $\left\{x^{(k)}\right\}$ is Q-linearly convergent to x_*. If we additionally have $2bL < 1$, then we can guarantee the Q-quadratic convergence of $\left\{x^{(k)}\right\}$.

Note that t_* and t_{**} are actually the roots of the polynomial $p(t) = \frac{1}{2}Lt^2 - t + b$. Kantorovich has given two different proofs of this theorem using recurrence relations (page 170 in [20]) and majorant functions (page 564 in [21]).

Smale [37] pointed out that Kantorovich's approach requires weak differentiability hypotheses on the system, for example, the function is C^2 on some domain in a Banach space, and it also requires the derivative bounds to exist over the whole of this domain. Smale adopted a different point of view and derived results from data at a single point and in contrast he needs strong hypotheses on differentiability and analyticity of the function. Moreover, Smale's theorem does not involve any bound on the second derivative of the function F on some neighbourhood of an approximate zero z.

Let X and Y be Banach spaces and let $f : X \to Y$ be an analytic map from X to Y.

The derivative of $f : X \to Y$ at $z \in X$ is a linear map $Df(z) : X \to Y$. If $Df(z)$ is invertible, Newton's Method gives a new iterate z' from z by

$$z' = z - Df(z)^{-1} f(z) =: N_f(z).$$

Let β denote the norm of this Newton step $z' - z$, i.e.,

$$\beta(z, f) = \beta(z) := \left|\left|Df(z)^{-1} f(z)\right|\right|.$$

If $Df(z)$ is not invertible, let $\beta(z) = \infty$.

Definition 2.2.2. *For a point $z_0 \in \mathcal{E}$, inductively define the sequence $z_n = z_{n-1} - Df(z_{n-1})^{-1} f(z_{n-1})$. We say that z_0 is an* approximate zero *of f if z_n is defined for all n and satisfies:*

$$||z_n - z_{n-1}|| \leq \left(\frac{1}{2}\right)^{2^{n-1}-1} ||z_1 - z_0||, \forall n.$$

Smale mentioned that for an approximate zero, Newton's method converges faster starting with the first iteration than generally expected. He also defined

$$\gamma(z, f) := \sup_{k>1} \left|\left| Df(z)^{-1} \frac{D^k f(z)}{k!} \right|\right|^{\frac{1}{k-1}}$$

where $D^k f(z)$ is the k-th derivative of of f at z as a k-linear map. In addition, he defined

$$\alpha(z, f) := \beta(z, f)\gamma(z, f)$$

Then, he showed the following theorem.

Theorem 2.2.3. *(Theorem A in [37]) If there is a naturally defined number α_0 approximately equal to 0.130707 such that if $\alpha(z, f) < \alpha_0$, then z is an approximate zero of f.*

This naturally defined number α_0 is a zero of the real quartic polynomial:

$$(2r^2 - 4r + 1)^2 - 2r.$$

By using Newton's method, one can compute that $\alpha_0 \approx 0.130707$.

Chapter 3

Linear Programming, Semidefinite Programming and Central Path

In the analysis of [30], the problem that Nesterov and Tunçel introduce is a standard convex optimization problem in conic form and the algorithm they propose is aimed for general convex optimization. For this thesis, we mainly focus on analyzing the rate of convergence of the algorithm in the case of Linear Programming and Semidefinite Programming. Hence, we need to introduce the backgrounds and preliminaries of Linear Programming and Semidefinite Programming and some notions that will be useful in understanding and analyzing the algorithm.

3.1 Linear Programming

We will define some important terms in order to state our primal-dual problem. First, we define the notions of **convex cones** and **dual cones**.

Definition 3.1.1. *Let $K \subseteq \mathbb{R}^n$. We say that K is a* convex cone *if for any $x, y \in K$ and any scalar $a, b \geq 0$, we have $ax + by \in K$.*

Definition 3.1.2. *Let $K \subseteq \mathbb{R}^n$. The* dual cone *of K is defined as*

$$K^* := \{s \in \mathbb{R}^n : \langle x, s \rangle \geq 0, \forall x \in K\}.$$

Now, we consider the standard convex optimization problem in an important special case, where the problem is a linear program. In this case, the cone K we are referring

to is the cone of all non-negative vectors in $\mathbb{R}^n$, denoted by $\mathbb{R}^n_+$. Indeed, $\mathbb{R}^n_+$ is self-dual: $\left(\mathbb{R}^n_+\right)^* = \mathbb{R}^n_+$.

Let (LP) and (LD) be a pair of primal- dual Linear Programming problems.

$$\begin{array}{lllll} (LP) & \min & c^\top x & & \\ & & Ax & = & b, \\ & & x & \geq & 0. \end{array}$$

$$\begin{array}{lllllll} (LD) & \max & b^\top y & & & & \\ & & A^\top y & + & s & = & c, \\ & & & & s & \geq & 0. \end{array}$$

For the above primal problem (LP), since we want to apply interior-point methods, we are interested in the feasible solutions which are strictly positive, i.e., the ones that are in the interior of the cone $\mathbb{R}^n_+$. Therefore, we want to remove the nonnegativity constraints and add a penalizing term in the objective function to force the feasible solutions to stay in the interior of the cone $\mathbb{R}^n_+$.

Hence, for $\mu > 0$, consider the following parameterized problem:

$$\begin{array}{lllll} (LP_\mu) & \min & \frac{1}{\mu} c^\top x - \sum_{i=1}^n \ln(x_i) & & \\ & & Ax & = & b. \end{array}$$

Since $x \geq 0$ is equivalent to $x \in \mathbb{R}^n_+$, we can see that for all feasible x, x is in the cone $\mathbb{R}^n_+$. Since $A \in \mathbb{R}^{m \times n}$, $A : \mathbb{R}^n \to \mathbb{R}^m$ can be seen as a linear map.

Next, we need the following notions of *Legendre-Fenchel conjugate* and *self-concordant barrier* to analyze the parameterized part we added to the objective function of (LP_μ).

Let $G \subseteq \mathbb{R}^n$. In this thesis, we use $\mathrm{bd}(G)$ to denote the boundary of the set G and $\mathrm{int}(G)$ to denote the interior of the set G.

Definition 3.1.3. *Let $f : \mathbb{R}^n \to [-\infty, +\infty]$. Then its* Legendre-Fenchel conjugate *is $f_* : \mathbb{R}^n \to [-\infty, +\infty]$ with $f_*(s) = \sup_{x \in \mathbb{R}^n} \{\langle s, x \rangle - f(x)\}$.*

Definition 3.1.4. *Let $G \subseteq \mathbb{R}^n$ be a closed convex set with nonempty interior. Then $f : \mathrm{int}(G) \to \mathbb{R}$ is called a* self-concordant barrier *for G with barrier parameter ν if the following conditions are satisfied:*

- *$f \in \mathcal{C}^3$, strictly convex on $\mathrm{int}(G)$;*

- *for every sequence* $\left\{x^{(k)}\right\} \subset \text{int}(G)$ *such that* $x^{(k)} \to \bar{x} \in \text{bd}(G)$, $f(x^{(k)}) \to +\infty$*;*
- $|D^3 f(x)[h,h,h]| \leq 2\left[D^2 f(x)[h,h]\right]^{\frac{3}{2}}$, $\forall x \in \text{int}(G)$, $\forall h \in \mathbb{R}^n$*;*
- $(Df(x)[h])^2 \leq \nu D^2 f(x)[h,h]$*, for every* $x \in \text{int}(G), h \in \mathbb{R}^n$.

Definition 3.1.5. *Suppose* G *is a closed convex cone with nonempty interior. Then a self-concordant barrier* f *of* G *with barrier parameter* ν *is called* logarithmically homogeneous *if*

$$f(tx) = f(x) - \nu \ln t, \ \ \forall x \in \text{int}(G), \ \ \forall t > 0.$$

Then, from the setting of (LP_μ), we define a n-Logarithmically Homogeneous Self-concordant Barrier (LHSCB) for $\mathbb{R}^n_+$: $F(x) := -\sum_{j=1}^n \ln(x_j)$. Moreover, we have $(\mathbb{R}^n_+)^* = \mathbb{R}^n_+$ and $F_*(s) = -\sum_{j=1}^n \ln(s_j) + \text{constant}$.

We can derive that

$$F'_*(s) = -S^{-1}e,$$
$$F''_*(s) = S^{-2},$$

where S is the $n-$by$-n$ diagonal matrix with diagonal entries s_j, for $j \in \{1, 2, \ldots, n\}$ and e is a vector in $\mathbb{R}^n$ with all ones.

Then, we have

$$\begin{aligned} f(y) &= F_*(c - A^\top y), \\ f'(y) &= AS^{-1}e, \\ f''(y) &= AS^{-2}A^\top. \end{aligned}$$

Next, we will define the central path of problem (LP_μ). In order to guarantee the uniqueness of the solution when we define the central path, we make the following assumption.

Assumption 3.1.6. *(Linear Programming Case)* $\text{rank}(A) = m$.

We can justify this assumption using Gaussian Elimination to solve $Ax = b$. If the resulting system is inconsistent, then we know that $Ax = b$ does not have a solution and (LP) is infeasible, and we are done. So, we may assume, there exists $x \in \mathbb{R}^n$ such that $Ax = b$. Now, if $\text{rank}(A) < m$, then $Ax = b$ has redundant row(s) and the redundant row(s) can be expressed as a linear combination of other rows. So, we can just remove the redundant row(s) from A, b and redefine A, b. This theoretical justification works when all computations are performed in exact arithmetic or when the data (A, b) are rational and Gaussian Elimination with suitable pivoting is employed. However, in general, in practice there can be serious numerical challenges.

Definition 3.1.7. *For the Linear Programming case, for each $\mu > 0$, let (x_μ, s_μ, y_μ) denote the unique solution to the following system:*

$$\begin{aligned} Ax &= b \\ A^\top y + s &= c \\ s &= \mu X^{-1} e, \;\; x > 0, \;\; s > 0. \end{aligned}$$

Then the primal-dual central path *is defined as* $\{(x_\mu, s_\mu, y_\mu) : \mu > 0\}$.

Let $v \in \mathbb{R}^m$ and we define $||v||_y := \left\{v^\top f''(y)^{-1} v\right\}^2$. Then, we use this local norm to define a neighbourhood of the central path as follows. Note that for every pair of primal-dual interior-points (x, s) (i.e. $Ax = b, x > 0, A^\top y + s = c, s > 0$), we have

$$s = \mu X^{-1} e \text{ iff } x = \mu S^{-1} e,$$

and the last equation implies $\mu A S^{-1} e = Ax = b$. This is equivalent to $f'(y) = \frac{1}{\mu} b$. This leads to the following notion of neighbourhood for the central path, for a given $\beta \geq 0$, in the dual space:

$$\begin{aligned} \mathcal{N}(\mu, \beta) &:= \left\{ y \in \mathbb{R}^m : \left|\left| f'(y) - \frac{1}{\mu} b \right|\right|_y \leq \beta \right\} \qquad (3.1) \\ &= \left\{ y \in \mathbb{R}^m : \left\{ \left(f'(y) - \frac{1}{\mu} b \right)^\top f''(y)^{-1} \left(f'(y) - \frac{1}{\mu} b \right) \right\}^{1/2} \leq \beta \right\}. \end{aligned}$$

3.2 Semidefinite Programming

Now, let us generalize a bit and look at another significant case, where the problem is a semidefinite programming problem. Let $\mathbb{S}^n_+$ denote the set of all $n \times n$ symmetric positive semidefinite matrices. For any pair of $n \times n$ symmetric matrices X, Y, we write $X \succeq Y$ to mean $X - Y \in \mathbb{S}^n_+$.

Let (SP) and (SD) be a pair of primal-dual Semidefinite Programming problems defined as follows:

$$\begin{aligned} (SP) \quad \min \quad & \langle C, X \rangle \\ & \langle A_i, X \rangle = b_i, \quad i \in \{1, 2, \ldots, m\} \\ & X \succeq 0, \end{aligned}$$

$$
\begin{array}{lll}
(SD) \quad \max & b^\top y & \\
& \sum_{i=1}^m A_i y_i \;+\; S \;=\; C, \\
& S \;\succeq\; 0,
\end{array}
$$

where $A_i \in \mathbb{S}^n, i \in \{1, 2, \ldots, m\}$, $C \in \mathbb{S}^n$ and $b \in \mathbb{R}^m$.

Since $X \succeq 0$ is equivalent to $X \in \mathbb{S}^n_+$, we can see that for all feasible X, X is in the cone $\mathbb{S}^n_+$. Let $X \in \mathbb{S}^n_+$, and let $\lambda_1(X), \lambda_2(X), \ldots, \lambda_n(X)$ denote the n eigenvalues of X. Then, we can characterize the interior of $\mathbb{S}^n_+$ as follows:

$$
\operatorname{int}(\mathbb{S}^n_+) \;=\; \left\{ X \in \mathbb{S}^n_+ : \lambda_1(X) > 0, \ldots, \lambda_n(X) > 0 \right\}.
$$

Hence, a natural barrier function to use is:

$$
F(X) := \; -\sum_{j=1}^n \ln(\lambda_j(X)) \;=\; -\ln(\textstyle\prod_{j=1}^n \lambda_i(X)) \;=\; -\ln(\det(X)).
$$

Moreover, we have $(\mathbb{S}^n_+)^* = \mathbb{S}^n_+$ since $\mathbb{S}^n_+$ is self- dual. Also,

$$
F_*(S) = -\ln(\det(S)) + \text{constant}.
$$

Since $F : \operatorname{int}(\mathbb{S}^n_+) \to \mathbb{R}$ is a $\mathcal{C}^2$ function, we can derive that:

$$
\begin{array}{c}
F'(X) = -X^{-1}, \\
F'_*(S) = -S^{-1}.
\end{array}
$$

Then, $F''(X)$ is the linear operator defined by $F''(X)W = X^{-1}WX^{-1}$. Similarly, $F''_*(S)$ is the linear operator defined by $F''_*(S)T = S^{-1}TS^{-1}$.

Let $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ be a linear map where $[\mathcal{A}(X)]_i = \langle A_i, X \rangle$. Then, we have

$$
f(y) \;=\; F_* \left(C - \sum_{i=1}^m y_i A_i \right),
$$

$$
\begin{array}{rcl}
\nabla f(y) & = & \mathcal{A} \nabla F_*(C - \mathcal{A}^* y) \\
& = & \mathcal{A} S^{-1},
\end{array}
$$

$$
\begin{array}{rcl}
\nabla^2 f(y) y & = & \mathcal{A} \nabla^2 F_*(C - \mathcal{A}^* y) \mathcal{A}^* y \\
& = & \mathcal{A} S^{-1} (\mathcal{A}^* y) S^{-1}.
\end{array}
$$

Similarly, we will define the central path in the case of Semidefinite Programming and the corresponding neighbourhood of the central path. Similar to the Linear Programming case, in order to guarantee the uniqueness of the solution when we define central path, we make the following assumption.

Assumption 3.2.1. *(Semidefinite Programming Case)* $\{A_1, A_2, \ldots, A_m\}$ *is a linearly independent set.*

We can justify this assumption using the similar idea. We use Gaussian Elimination to solve $\langle A_i, X\rangle = b_i, i \in \{1, \ldots, m\}$. If the resulting system is inconsistent, then we know that the system $\langle A_i, X\rangle = b_i, i \in \{1, \ldots, m\}$ does not have a solution. Therefore, (SP) is infeasible, and we are done. Otherwise, we may assume, there exists $X \in \mathbb{S}^n$ such that $\langle A_i, X\rangle = b_i, i \in \{1, \ldots, m\}$. Then, if the set $\{A_i : i \in \{1, 2, \ldots, m\}\}$ is linearly dependent, then the system of equations $\langle A_i, X\rangle = b_i, i \in \{1, \ldots, m\}$ has redundant equation(s). We can just remove the redundant equation(s) and redefine the set $\{A_i\}$ and vector b so that $\{A_i\}$ is a linearly independent set.

Definition 3.2.2. *In the case of Semidefinite Programming, the* primal-dual central path (X_μ, S_μ, y_μ) *is defined as the set of unique solutions to the following system:*

$$\begin{aligned} \langle A_i, X\rangle &= b_i, i \in \{1, \ldots, m\} \\ \sum_{i=1}^{m} y_i A_i + S &= C \\ S &= \mu X^{-1}, \;\; X \succ 0, \;\; S \succ 0 \end{aligned}$$

for all $\mu > 0$.

Moreover, going through a similar reasoning as in the LP case, the neighbourhood of the central path for a given $\beta \geq 0$ is defined as follows:

$$\begin{aligned} \mathcal{N}(\mu, \beta) &:= \left\{ y \in \mathbb{R}^m : \left\| f'(y) - \frac{1}{\mu} b \right\|_y \leq \beta \right\} \\ &= \left\{ y \in \mathbb{R}^m : \left\{ \left(f'(y) - \frac{1}{\mu} b \right)^\top f''(y)^{-1} \left(f'(y) - \frac{1}{\mu} b \right) \right\}^{1/2} \leq \beta \right\} \end{aligned}$$

Semidefinite Optimization can be seen as a generalization of Linear Optimization since in Semidefinite Optimization, we replace each $\mathbb{R}_+$ by $\mathbb{S}_+^{n_i}$, where $\mathbb{S}_+^{n_i}$ is the set of n_i-by-n_i symmetric positive semidefinite matrices.

3.3 General Convex Optimization

Now, let us generalize the previous two classes and consider the standard conic optimization problem:

$$\begin{array}{lllll} (P) & \inf & \langle c, x\rangle & & \\ & & \mathcal{A}x & = & b, \\ & & x & \in & K, \end{array}$$

where $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$, $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation and $K \subset \mathbb{R}^n$ is a pointed, closed, convex cone with nonempty interior. Then the dual problem of (P) is:

$$\begin{array}{lllllll} (D) & \sup & \langle b, y\rangle_D & & & & \\ & & \mathcal{A}^*y & + & s & = & c, \\ & & & & s & \in & K^*. \end{array}$$

For the above primal problem (P), since we want to apply interior-point methods, we are only interested in the feasible solutions which are in the interior of the cone K. Therefore, we want to remove the constraint that requires the points to stay in the cone K and add a penalizing term in the objective function to force the feasible solutions to stay in the interior of the cone K.

Hence, for $\mu > 0$, consider the following parameterized problem:

$$\begin{array}{lllll} (P_\mu) & \inf & \frac{1}{\mu}\langle c, x\rangle + F(x) & & \\ & & \mathcal{A}x & = & b, \end{array}$$

where $F : \text{int}(K) \to \mathbb{R}$ is a ν-Logarithmically Homogeneous Self-concordant Barrier (LHSCB) for K. Every convex cone K admits a ϑ-LHSCB where $\nu = O\left(dim(k)\right)$ (see [29]).

Throughout this thesis, we make the following assumption.

Assumption 3.3.1. *There exists $x^{(0)} \in \text{int}(K)$ such that $\mathcal{A}x^{(0)}) = b$ and there exists $y^{(0)} \in \mathbb{R}^m, s^{(0)} \in \text{int}(K^*)$ such that $\mathcal{A}^*y^{(0)} + s^{(0)} = c$. Moreover, $\mathcal{A}$ is surjective.*

In the case of Linear Programming, $\mathcal{A}$ is surjective is equivalent to Assumption 3.1.6 and in the case of Semidefinite Programming, $\mathcal{A}$ is surjective is equivalent to Assumption 3.2.1. We know that the above assumption can be justified in both cases.

Under Assumption 3.3.1, (P_μ) has a unique optimal solution (x_μ, s_μ, y_μ) for every $\mu > 0$ determined by

$$\begin{array}{rcll} \mathcal{A}x & = & b, & x \in \operatorname{int}(K) \\ \mathcal{A}^* y + s & = & c, & s \in \operatorname{int}(K^*) \\ s & = & -\mu F'(x). & \end{array}$$

Note that the equation $s = -\mu F'(x)$ can be replaced by $x = -\mu F'_*(s)$. This is due to the fact that

$$\begin{array}{rcccl} s = -\mu F'(x) & \Leftrightarrow & F'_*(s) & = & F'_*(-\mu F'(x)) \\ & \Leftrightarrow & F'_*(s) & = & \frac{1}{\mu} F'_*(-F(x)) \\ & \Leftrightarrow & F'_*(s) & = & -\frac{1}{\mu} x, \end{array}$$

as we saw in the special case of Linear Programming.

Another way of arriving at $x = -\mu F'_*(s)$ is by starting with

$$\begin{array}{lll} (D_\mu) \quad \inf & -\frac{1}{\mu}\langle b, y\rangle_D + F_*(s) & \\ & \mathcal{A}^* y + s \;=\; c, & \end{array}$$

instead of (P_μ) as we did above.

3.4 Strict Complementarity

In order to use some special properties of Linear Programming and Semidefinite Programming to analyze the asymptotic behaviour and the rate of convergence of the algorithms, we need the notions of **strict complementarity** and **analytic centre**. In this section, we will define the related terminologies in the cases of Linear Programming and Semidefinite Programming.

3.4.1 Linear Programming case

Let (LP) and (LD) be a pair of primal-dual Linear Programming problems defined in Section 3.1. The algorithms we are analyzing in this thesis is a primal-dual path-following

predictor-corrector interior-point method and a dual, path-following predictor-corrector interior-point method.

A *path-following* algorithm is a class of algorithms where we restrict all iterates to a neighbourhood of the central path and we find a solution of the problem by following the central path. A *predictor-corrector* algorithm is a class of algorithms where we use two types of steps: predictor step and corrector step alternatively to find a solution of the problem. Predictor directions are used to reduce μ, equivalently, the duality gap, and the corrector directions are used to improve centrality, which is, staying closer to the central path. In addition, there is a pair of neighbourhoods nested one inside the other. In a *predictor-corrector* algorithm, we start with a point in the smaller neighbourhood and take a predictor step. We move along the prediction direction so that the iterate is in the larger neighbourhood and μ is reduced as much as possible (this typically means, after a prediction step, the iterate ends up on the boundary of the larger neighbourhood). Then, we apply a corrector step to the current iterate to take the iterate back into the smaller neighbourhood and leave μ unchanged. Next, we repeat the above iterations.

In this thesis, the algorithm we adopt is a variant of the algorithm proposed in [30] in the cases of LP and SDP.

Let $(y^{(k)}, s^{(k)})$ be the current iterate. Then, the predictor direction is defined as:

$$d_y^{(k)} := \left[\nabla^2 f(y^{(k)})\right]^{-1} \nabla f(y^{(k)}).$$

In the case of LP, we can derive the following

$$d_y = \left[AS^{-2}A^\top\right]^{-1} AS^{-1}e.$$

On the other hand, for the derivation of corrector steps we consider various approaches (see Appendix A). The one that we use in our implementation of the algorithm is the first corrector step in Appendix A, i.e.,

$$\Delta_y^{(k)} := \frac{1}{\mu}\left[f''(y^{(k)})\right]^{-1} b - \left[f''(y^{(k)})\right]^{-1} f'(y^{(k)}).$$

Now, we define the notion of *complementarity* and *strict complementarity* for Linear Programming and some notations we will use in the following analysis.

Definition 3.4.1. *(Complementarity for LP) Let x^* and (y^*, s^*) be feasible solutions to (LP) and (LD) respectively. Then, (x^*, s^*) is called a* complementary *solution if*

$$x_i^* s_i^* = 0, \ \ \forall \ \ i \in \{1, 2, \ldots, n\}.$$

Moreover, x^* and (y^*, s^*) are optimal in (LP) and (LD) respectively, if and only if, (x^*, s^*) is a complementary solution.

Definition 3.4.2. *(Strict complementarity for LP) Let x^* and (y^*, s^*) be feasible solutions to (LP) and (LD) respectively. Let*

$$B := \{i : x_i^* > 0\}.$$

$$N := \{1, 2, \ldots, n\} \setminus B.$$

If we have

$$N = \{i : s_i^* > 0\},$$

then we say that (x^*, s^*) is a strictly complementary *solution.*

Theorem 3.4.3. *(see [36] for a proof) Among all optimal solutions for (LP) and (LD), there exists at least one optimal solution pair (x^*, s^*), which is strictly complementary.*

Let (LP) and (LD) be a pair of primal- dual Linear Programming problems defined in Section 3.1. Then, the optimal solutions can be characterized by using the $[B, N]$ partition.

Let P^* and D^* denote the optimal sets of (LP) and (LD) respectively. Then, we have

$$P^* = \{x : A_B x_B = b, \;\; x_B \geq 0, x_N = 0\},$$

$$D^* = \{(y, s) : A^\top y + s = c, \;\; s_N \geq 0, s_B = 0\}.$$

Now, we define the *analytic centre* of the optimal face for LP case. The following definition gives an analytic characterization as minimizers of a strictly convex function. Furthermore, the analytic centre of P^* and the analytic centre of D^* is a strict complementary solution pair.

Definition 3.4.4. *$x^a \in P^*$ is the* analytic centre *of P^* if*

$$(x^a)_B = \arg\min_{x_B > 0} \left\{ -\sum_{i \in B} \ln x_i : A_B x_B = b \right\},$$

$(y^a, s^a) \in D^*$ is the analytic centre *of D^* if*

$$(y^a, (S^a)_N) = \arg\min_{y \in \mathbb{R}^m, s_N > 0} \left\{ -\sum_{i \in N} \ln s_i : A_N^\top y + s_N = c_N, \;\; A_B^\top y = c_B \right\}.$$

We denote by $X^* \subset \mathbb{R}^n$ the set of limit points of the primal central path and by $S^* \subset \mathbb{R}^n$ the set of limit points of the dual central path. Then, X^* is the set of analytic centre of P^* and S^* is the set of analytic centre of D^*.

We make the following two remarks based on the uniqueness of the analytic centre and the relationship between analytic centre and Newton's Method.

Remark 3.4.5. *Let x^a and (y^a, s^a) be defined as above. Then, x^a and (y^a, s^a) are unique.*

The objective function: $-\sum_{i\in B} \ln x_i$ is strictly convex, so the minimizer x^a is unique. Similarly, s^a is also unique because it is the minimizer of the objective function $-\sum_{i\in N} \ln s_i$. Moreover, y^a is uniquely determined since $\operatorname{rank}(A) = m$ and hence the rows of A are linearly independent.

Remark 3.4.6. *Analytic centre is in the relative interior of the solution set. In the constrained problem setting, we can derive the optimality conditions using KKT Theorem and then solve the resulting system of (nonlinear) equations using Newton's method. By Kantorovich's theory ([20]), we know that in a small neighbourhood of the analytic centre x^a, i.e. the minimizer of the objective function $-\sum_{i\in B} \ln x_i$, Newton's Method admits quadratic convergence.*

3.4.2 Semidefinite Programming Case

Consider the central path $\{(X_\mu, S_\mu, y_\mu) : \mu > 0\}$ defined in 3.2.2. Since we can vectorize symmetric matrices, we define the following linear map that maps a matrix to a vector.

Let $A \in \mathbb{R}^{m\times n}$, $\operatorname{vec}(A) := [a_{1,1}, \ldots, a_{m,1}, a_{1,2}, \ldots, a_{m,2}, \ldots, a_{1,n}, \ldots, a_{m,n}]^\top$. Then, the above system is equivalent to:

$$\begin{aligned} \operatorname{vec}(A_i)^\top \operatorname{vec}(X) &= b_i, i \in \{1, \ldots, m\} \\ \sum_{i=1}^{m} y_i \operatorname{vec}(A_i) + \operatorname{vec}(S) &= \operatorname{vec}(C) \\ S &= \mu X^{-1},\ X \succeq 0,\ S \succeq 0 \end{aligned}$$

for all $\mu > 0$.

Let (SP) and (SD) be a pair of primal- dual Semidefinite Programming problems as defined in Section 3.2.

Let $\mathcal{P}$ and $\mathcal{D}$ denote the feasible sets of (SP) and (SD) respectively and $\mathcal{P}^*$ and $\mathcal{D}^*$ denote the optimal sets of (SP) and (SD) respectively.

Definition 3.4.7. *A pair of optimal solutions* $(X, S) \in \mathcal{P}^* \times \mathcal{D}^*$ *is called a* maximally complementary solution pair *to the pair of problems (SP) and (SD) if it maximizes* $\text{rank}(X) + \text{rank}(S)$ *over all optimal solution pairs.*

In the field of Semidefinite Programming, we use the following notation: let $X \in \mathbb{S}^n$, we say that $X \succ 0$ if $X \in \mathbb{S}^n_{++}$, i.e. we write $X \succ 0$ if X is a n-by-n positive definite matrix.

Let X^* and S^* be a pair of optimal solutions. Under our assumptions, by the optimality conditions applied to (SP) and (SD), we know that $X^*S^* = S^*X^* = 0$, and hence the matrices X^* and S^* commute. Then X^* and S^* can be diagonalized simultaneously. Therefore, without loss of generality, we may assume that X^*, S^* are both diagonal and of the form:

$$X^* = \begin{bmatrix} \bar{X} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad S^* = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \bar{S} \end{bmatrix}, \quad \text{where } \bar{X} \succ 0, \bar{S} \succ 0.$$

Then, we define index sets B and N as the subsets of $\{1, 2, \ldots, n\}$ consisting of the indices of the rows of X^* and S^* containing the rows of $\bar{X}$ and $\bar{S}$ respectively. Clearly, $|B| + |N| \leq n$.

Definition 3.4.8. *Let* T:= $\{1, 2, \ldots, n\} \setminus (B \cup N)$. *We say that* (X^*, S^*) *is a* strictly complementary solution *if* $T = \emptyset$.

From the above definition, we know that each optimal solution pair $(\hat{X}, \hat{S})$ is (under an orthogonal similarity transformation) of the form

$$\hat{X} = \begin{bmatrix} \hat{X}_B & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \hat{S} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \hat{S}_N \end{bmatrix},$$

where $\hat{X}_B \in \mathbb{S}^B_+$ and $\hat{S}_N \in \mathbb{S}^N_+$.

Moreover, if $(\hat{X}, \hat{S})$ is a maximally complementary solution pair of (SP) and (SD) respectively, then

$$\hat{X} = \begin{bmatrix} \hat{X}_B & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \hat{S} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \hat{S}_N \end{bmatrix}.$$

where $\bar{X}_B \in \mathbb{S}^B_{++}$ and $\bar{S}_N \in \mathbb{S}^N_{++}$.

Now, we can characterize the optimal solutions by using the block partition.

$$\mathcal{P}^* = \left\{ X \in \mathcal{P} : X = \begin{bmatrix} X_B & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \right\}$$

$$\mathcal{D}^* = \left\{ (y, S) \in \mathcal{D} : S = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & S_N \end{bmatrix} \right\}.$$

Note that the A_i's appearing in the next definition are not necessarily the same as in the original data. Let $Q \cdot Q^\top$ be the orthogonal similarity transformation exposing the above block diagonal structure of X^* and S^*, i,e. $X^* = Q \begin{bmatrix} \bar{X} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} Q^\top$. Then considering the change of variables $\tilde{X} := Q^\top X Q$, $\tilde{S} := Q^\top S Q$, leads to (using $\langle A_i, X \rangle = \langle A_i, Q\tilde{X}Q^\top \rangle = \langle Q^\top A_i Q, \tilde{X} \rangle$) replacing A_i by $Q^\top A_i Q$ for every i. For simplicity of notation, we continue to use A_i below.

Definition 3.4.9. *Let $\mathcal{P}^*$ and $\mathcal{D}^*$ be defined as above.*

$X^a \in \mathcal{P}^*$ is the analytic centre *of P^* if*

$$(X^a)_B = \arg \min_{X_B \in \mathbb{S}^{|B|}_{++}} \left\{ -\ln \det X_B : \langle (A_i)_B, X_B \rangle = b_i, \;\; i \in \{1, 2, \ldots, m\} \right\}.$$

$(y^a, S^a) \in \mathcal{D}^*$ is the analytic centre *of D^* if*

$$(y^a, (S^a)_N) = \arg \min_{y \in \mathbb{R}^m, S_N \in \mathbb{S}^{|N|}_{++}} \left\{ -\ln \det S_N : \sum_{i=1}^m (A_i)_N + S_N = C_N, \;\; \sum_{i=1}^m (A_i)_k = C_k, \;\; k \notin N \right\}.$$

3.5 Literature survey on superlinear convergence in polynomial iteration complexity interior-point methods

In 1989, Potra [32] proved some properties of the R-order convergence and Q-order convergence and gave sufficient conditions for a sequence to admit the Q-order and the R-order of convergence greater than 1. Additionally, he gave an extra condition where it will imply

the sequence having exact Q-order of convergence and compared the results of the Q-order convergence with previous results from Feldstein-Firestone [14].

Later in 1994, El-Bakry, Tapia and Zhang [13] studied various indicators proposed in the literature to identify the zero variables in the case of Linear Programming where the only inequality constraints of the problem are the non-negativity constraints. They used the term *indicator* to denote a function which indicates the set of constraints that are active at a solution of a constrained optimization problem. The main focus is on an indicator that can be used in primal-dual interior-point methods. They defined indicator function and analyzed its properties as well as those that a good indicator should have. Comparison on advantages and disadvantages of numerical and theoretical behaviour of the variables used as indicators, primal-dual indicators, which use information on both primal and dual problem such as $[S^{(k)}]^{-1}X^{(k)}e$ and the Tapia indicators [41], which use quotient of successive slack variables and quotient of successive primal iterates, are investigated. At last, they also presented the rate of convergence of some indicator functions and demonstrated their numerical performance.

One year later, Tsuchiya [44] showed the quadratic convergence property of the Iri-Imai Algorithm, which is a polynomial-time interior-point potential reduction algorithm where the Newton's method is applied to a multiplicative barrier function. This multiplicative barrier function is first introduced in [19] and it is defined as follows:

Definition 3.5.1. *We consider a linear program in the standard form as defined in 3.1. Besides Assumption 3.3.1, we further assume that the optimal value of the objective function is known a priori to be equal to zero. Now, we define the following function $f_\rho(x)$ which is made up of the objective function and a monomial representing the inequalities:*

$$f_\rho(x) := \frac{(c^\top x)^\rho}{\prod_{i=1}^n x_i}.$$

The domain of f_ρ is $\{x \in \mathbb{R}^n : Ax = b, x > 0\}$.

Note that $f_\rho(x)$ is strictly convex in the relative interior of the primal feasible region if the parameter $\rho \geq n+1$ and the function value goes to 0 only if x approaches to the set of optimal solutions. On the other hand, Karmarkar [22] introduced a potential function to measure the quality of different feasible points of the linear programs in 1984 and its adaptation to the current formulation is:

$$\phi_\rho(x) := \rho \ln(c^\top x) - \sum_{i=1}^n \ln x_i.$$

Comparing this potential function and the multiplicative barrier function, we can easily see that

$$\phi_\rho(x) = \ln(f_\rho(x)).$$

Since the Hessian matrix of the multiplicative barrier function is not defined when the current iterates are on the boundary of the feasible region, the proof of quadratic convergence of this variant Newton's method is not trivial (also recall our discussion in Chapter 1). In [44], Tsuchiya showed that the limit point of the iterates generated by the Iri-Imai algorithm exists in the relative interior of a unique face of the optimal solution set without assuming the nondegeneracy of the problem. Moreover, he showed that the iterates generated by the Iri-Imai algorithm converges to the relative analytic centre of that unique face.

In 2001, Potra [34] studied a class of infeasible interior point methods to solve the horizontal linear complementarity problem (HLCP) and presented sufficient conditions for the Q-superlinear convergence of the iteration sequences generated by primal-dual interior-point methods for linear complementarity problems. The horizontal linear complementarity problem (HLCP) is defined as follows:

$$\begin{aligned} x^\top s &= 0, \\ Qx + Rs &= b, \\ x, s &\geq 0, \end{aligned} \tag{3.2}$$

where $b \in \mathbb{R}^n$ and $Q, R \in \mathbb{R}^{n\times n}$ with $\text{rank}[Q, R] = n$. Moreover, he assumed that there is a constant $\kappa \geq 0$ such that for any $u, v \in \mathbb{R}^n$,

$$Qu + Rv = 0 \text{ implies } (1 - 4\kappa) \sum_{i\in I_+(u,v)} u_i v_i + \sum_{i\in I_-(u,v)} u_i v_i \geq 0$$

where

$$I_+(u, v) := \{i : u_i v_i > 0\}, \;\; I_+(u, v) := \{i : u_i v_i < 0\}.$$

If a pair (Q, R) satisfies the above condition, we say that (Q, R) is a $P_*(\kappa)$ - pair.

Moreover, he applied those sufficient conditions to demonstrate Q-superlinear convergence results of the iterates generated by some particular well-known primal-dual interior-point methods, for example, simplified largest step method, simplified Mizuno-Todd-Ye method and the LPF+ algorithm of Wright [48].

In the same year, Tütüncü [46] analyzed an algorithm which is a primal-dual variant of the Iri-Imai method and minimize the Tanabe-Todd-Ye (TTY) potential function using

modified Newton's search directions. For a primal-dual pair of linear programs defined in Section 3.1, the Tanabe-Todd-Ye (TTY) potential function ([40] and [43]) is defined as follows ($\rho \geq n$):

$$\Phi_\rho(x,s) := \rho \ln(x^\top s) - \sum_{i=1}^{n} \ln(x_i s_i), \text{ for every } (x,s) > 0.$$

Compared to Karmarkar's potential function, TTY potential function is a primal-dual variant of the Karmarkar's potential function.

Tütüncü focused on the degenerate problems and showed both the global and local convergence properties as well as polynomial iteration complexity of this primal-dual interior-point pure potential-reduction algorithm in the case of Linear Programming. This work improved and generalized three previous papers, analyzed the asymptotic behaviour of the search directions and the iteration sequences, and the uniqueness of the limit points of the algorithm, and then proved the quadratic convergence results of the iterates generated.

Potra [35] presented three affine scaling methods that produce iteration sequences in a wide neighbourhood of the central path to solve monotone linear complementarity problems in 2008. One is a first order affine-scaling method and the other two are m^{th} order affine-scaling methods. For the first order affine scaling method, if the linear complementarity problem (LCP) admits a strictly complementary solution, then both the duality gap and the iteration sequences converge Q-superlinearly to zero and Q-superlinearly to a strictly complementary solution. Linear complementarity problem (LCP) is a horizontal linear complementarity problem (HLCP) as defined in (3.2) where $R = -I$ and Q is a positive semidefinite matrix. Throughout his paper, he assumed that the HLCP is monotone, where

$$Qu + Rv = 0 \text{ imples } u^\top v \geq 0 \text{ for any } u, v \in \mathbb{R}^n.$$

It is the first affine-scaling method generating sequences in the $\mathcal{N}_\infty^-(\alpha)$ neighbourhood of the central path which obtain $O(\sqrt{n}L)$ iteration complexity and Q-superlinear convergence without assuming strict complementarity.

Let $\mathcal{F}$ denote the set of all feasible solution pairs of HLCP and $\mathcal{F}^0$ be the set of all strictly feasible points. Then, we define the following proximity measure that is used to measure the distance of a point $z \in \mathcal{F}$ to the central path.

$$\delta_\infty^-(z) := \left\| \left[\frac{x^\top s}{\mu(z)} - e \right]^- \right\|_\infty$$

where $[v]^-$ denotes the negative part of a vector, i.e., $[v]^- = -\max\{-v, 0\}$. The $\mathcal{N}_\infty^-(\alpha)$ neighbourhood of the central path, also called the *wide neighbourhood*, is defined as

$$\mathcal{N}_\infty^-(\alpha) := \left\{ z \in \mathcal{F}^0 : \delta_\infty^-(z) \leq \alpha \right\}.$$

In 2009, Potra and Stoer [33] proposed a class of infeasible interior-point methods which has the same Q-order and computational cost per iteration as the methods Potra presented in [35] for sufficient LCPs. *Sufficient HLCP* and *Sufficient LCP* are defined using the following notations.

Consider an HLCP as defined in (3.2). Let Φ denote the null space of the matrix $[Q, R] \in \mathbb{R}^{n \times 2n}$, i.e.

$$\Phi := \left\{ \begin{bmatrix} u \\ v \end{bmatrix} : Qu + Rv = 0 \right\}$$

and $\Phi^\perp$ denote its orthogonal space.

Then, they gave the definition of *column sufficient* and *row sufficient*.

Definition 3.5.2. *We say that the pair (Q, R) is* column sufficient *if*

$$\begin{bmatrix} u \\ v \end{bmatrix} \in \Phi, \;\; u^\top v \leq 0 \;\; \textit{implies} \;\; u^\top v = 0,$$

and row sufficient *if*

$$\begin{bmatrix} u \\ v \end{bmatrix} \in \Phi^\perp, \;\; u^\top v \geq 0 \;\; \textit{implies} \;\; u^\top v = 0.$$

Moreover, (Q, R) is called a sufficient pair *if it is both column and row sufficient, and then the HLCP is called a* sufficient HLCP*. Then, a* sufficient LCP *is defined as a sufficient HLCP where $R = -I$ and $Q \succeq 0$.*

Furthermore, they used the following definition of weighted central path:

Definition 3.5.3. *For any vector $\rho \in \mathbb{R}^n_{++}$ and any parameter $\tau > 0, \bar{b}$, the curve $z = z(\tau, \rho)$ defined by the solutions of the following nonlinear system is called the* weighted infeasible central path pinned on $\bar{b}$ with weight vector ρ*.*

$$\begin{aligned} x^\top s &= \tau\rho, \\ Qx + Rs &= b - \tau\bar{b}, \\ x, s &> 0. \end{aligned}$$

If we set $\rho = e$, then $z(\tau, e)$ is called the infeasible central path pinned on $\bar{b}$*.*

The algorithm they proposed in this paper does not depend on κ of the complementarity problem. It only uses one matrix factorization and m back-solves for each step and it generates iterates that lie in the wide neighbourhood of the central path, where the neighbourhood is given by $\mathcal{N}_{\infty}^{-}(1-\beta)$.

This paper generalized the result of [35] to sufficient LCPs where the starting points are not necessary feasible and proposed an algorithm that has both polynomial complexity and Q-superlinear convergence. It used the weighted infeasible central path in the analysis and also helps us better understand the behaviour of interior point methods in the wide neighbourhood of the central path.

Chapter 4

Superlinear and Quadratic Convergence in modern, primal-dual Interior Point Methods

In the past few decades, superlinear and quadratic convergence results for polynomial-time primal-dual interior-point methods were proven for the Linear Programming and for Semidefinite Programming problems. In 1993, several primal-dual interior-point methods for Linear Programming were presented by Mizuno, Todd and Ye [28] and this type of algorithms are widely studied afterwards. We survey two important papers which investigate and analyze the asymptotic quadratic convergence of Mizuno-Todd-Ye $O(\sqrt{n}L)$ iteration predictor-corrector primal-dual interior-point algorithm. Moreover, for the Semidefinite Programming case, we survey several papers that focus on the superlinear convergence result of some primal-dual predictor-corrector interior-point methods.

4.1 Linear Programming Case

In both analysis of [26] by Mehrotra and [50] by Ye, Güler, Tapia, and Zhang, they focus on the asymptotic properties of the primal-dual affine-scaling direction and use these properties to show that Mizuno-Todd-Ye $O(\sqrt{n}L)$ iteration predictor-corrector primal-dual interior-point algorithm admits Q-quadratic convergence for the case of linear programming.

Unlike a lot of previous results, neither paper assumes that the iteration sequence

generated by the algorithm is convergent. Moreover, the usual linear programming nondegeneracy is not required in these analyses. They assume the existence of the strict feasible solution(s), which gives the existence of the primal-dual central path in the (x, s)-space and the boundedness of the primal and dual optimal faces, and assume that A has full row rank. In both papers, they use a unique partition $A = [B, N]$ according to the strictly complementary solution pair and this partition characterizes the primal optimal face, and then they use this idea to prove the boundedness of the norm of the directions respectively.

For feasible iterates (x, s) and $\beta > 0$, we use the 2-norm neighbourhood of the primal-dual central path: $\mathcal{N}(\mu, \beta) := \left\{(x, s) : \left|\left|\frac{Xs}{\mu} - e\right|\right| \leq \beta\right\}$. For each iteration of the Mizuno-Todd-Ye $O(\sqrt{n}L)$ iteration predictor-corrector primal-dual interior-point algorithm, we are given a feasible point $\left(x^{(k)}, s^{(k)}\right) \in \mathcal{N}(\mu_k, \beta)$, where $\mu_k := \frac{{x^{(k)}}^\top s^{(k)}}{n}$. Then, we compute the primal-dual affine-scaling direction (d_x, d_s, d_y), which is defined as the solution of the following system:

$$\begin{bmatrix} S & X & 0 \\ 0 & I & A^\top \\ A & 0 & 0 \end{bmatrix} \begin{bmatrix} d_x \\ d_s \\ d_y \end{bmatrix} = \begin{bmatrix} Xs \\ 0 \\ 0 \end{bmatrix}. \tag{4.1}$$

Next, we generate $\hat{x}^{(k)} = x^{(k)} - \alpha_k d_x$, $\hat{y}^{(k)} = y^{(k)} - \alpha_k d_y$, and $\hat{s}^{(k)} = s^{(k)} - \alpha_k d_s$ for some step size parameter α_k such that $\left(\hat{x}^{(k)}, \hat{s}^{(k)}\right) \in \mathcal{N}(\hat{\mu_k}, 2\beta)$, where $\hat{\mu_k} = \frac{(\hat{x}^{(k)})^\top \hat{s}^{(k)}}{n}$. This is called a predictor step.

In both papers, they proved the following main theorem on the Q-quadratic convergence of Mizuno-Todd-Ye $O(\sqrt{n}L)$ iteration predictor-corrector primal-dual interior-point algorithm.

Theorem 4.1.1. *Let $\left\{\left(x^k, s^k\right)\right\}$ be the sequence generated by Mizuno-Todd-Ye $O(\sqrt{n}L)$ iteration predictor-corrector primal-dual interior-point algorithm. Then, with constants $0 < \beta \leq \frac{1}{4}$,*

1. *the algorithm has iteration complexity $O(\sqrt{n}L)$;*
2. *$1 - \alpha^k = O((x^k)^\top s^k)$; and*
3. *$(x^k)^\top s^k \to 0$, Q-quadratically.*

For Mehrotra's result [26], he mentioned the Mehrotra-Ye approach to find a solution on the primal and dual optimal faces, and the method finds solutions for problems that

use integral data after a polynomial number of iterations. In this approach, a direction $(\Delta x, \Delta s, \Delta y)$ is defined such that $x^k - \Delta x = x^*,\ y^k - \Delta y = y^*$ and $s^k - \Delta s = s^*$, where $x^*, (y^*,\ s^*)$ is a pair of optimal solution to (LP) and (LD) respectively. Then, he derived that

$$A\Delta x = 0,\ \ A^\top \Delta y + \Delta s = 0,\ \text{ and }\ (\Delta x)^\top \Delta s = 0.$$

Moreover, $(\Delta x)_i = x_i^k,\ \forall i \in N$ and $(\Delta s)_i = s_i^k,\ \forall i \in B$.

In order to show the main result on the Q-quadratic convergence, he first showed the following theorem.

Theorem 4.1.2. *For the current iterate $\left(x^k, s^k\right)$, if it satisfies the condition*

$$\frac{{x^k}^\top s^k}{\min_i\{x_i^j s_i^k\}} \leq O(n), \tag{4.2}$$

then we have

$$\left|\left|\begin{bmatrix} X^{-1}(\Delta x - d_x) \\ S^{-1}(\Delta s - s_x) \end{bmatrix}\right|\right|^2 \leq O(n) \max\left\{\max_{i\in B}\left(\left(\frac{(\Delta x)_i}{x_i^k}\right)^2\right), \max_{i\in N}\left(\left(\frac{(\Delta s)_i}{s_i^k}\right)^2\right)\right\}.$$

By using the above theorem and the formula of the affine-scaling directions defined in (4.1), he also showed the following theorem:

Theorem 4.1.3. *Let α_{max} denote the maximum step size along the primal-dual affine-scaling direction (d_x, d_s, d_y) defined in (4.1). Then, the feasible step size along the primal-dual affine scaling direction approaches 1 as ${x^k}^\top s^k \to 0$. Moreover,*

$$(x - \alpha_{max} d_x)^\top (s - \alpha_{max} d_s) = (1 - \alpha_{max}) x^\top s \leq K_2 (x^\top s)^2,$$

where $K_2 = O(n^{\frac{1}{2}})K_1$ and K_1 is a large data-dependent constant. Hence, the duality gap decreases quadratically for sufficiently small ${x^k}^\top s^k$.

The maintaining of the condition (4.2) can be performed by taking a step along the primal-dual affine-scaling direction d_x, d_y and d_s with the appropriate length and a corrector step by Mizuno-Todd-Ye $O(\sqrt{n}L)$ iteration predictor-corrector algorithm.
In the paper [50] by Ye, Güler, Tapia, and Zhang, they focus on bounding $\left|\left|\delta^k\right|\right|$, where $\delta^k := \frac{D_x^k D_s^k}{\mu^k}$ by first bounding $||d_x||$ and $||d_s||$. They show the following theorem that bounds the norms of the predictor directions:

Theorem 4.1.4. *Let d_x^k and d_s^k are obtained from the above linear system for the kth predictor step and $\mu_k = \frac{(x^k)^\top s^k}{n}$. Then, d_x^k and d_s^k satisfy*

$$\left|\left|d_x^k\right|\right| = O(\mu_k) \text{ and } \left|\left|d_s^k\right|\right| = O(\mu_k),$$

for all $k > 0$.

In the proof of the above theorem, they write d_x and d_s in terms of some orthogonal projection matrices and relate $(d_x)_B$ and $(d_s)_N$ to the minimizers of some least square problems.

Then, they use $\left|\left|\delta^k\right|\right|$ to bound $1 - \alpha^k$, where α^k is the largest step size they choose according to Mizuno-Todd-Ye $O(\sqrt{n}L)$ iteration predictor-corrector primal-dual interior-point algorithm. Moreover, they make some observations about the asymptotic behaviour of $(X^k)^{-1}d_x^k$ and $(S^k)^{-1}d_s^k$ at the predictor steps.

In both papers, they use this unique partition $\{1, 2, \ldots, n\} = [B, N]$ that defines the primal and dual optimal faces based on the complementarity properties. Note that every iteration in this algorithm requires solving two linear systems, one for predictor direction and one for corrector direction. Also, Mizuno-Todd-Ye $O(\sqrt{n}L)$ predictor-corrector interior-point algorithm is a primal-dual symmetric algorithm, while in [30] and our analysis, the path-following algorithm we adopt works only in the dual space. Moreover, in [30] and our analysis, for the linear programming case, the only assumption we need is that the primal and dual problems admit Slater points. So, the superlinear convergence result can be driven without assuming the convergence of the iterations generated by the algorithm or the nondegeneracy of the linear program.

4.2 Semidefinite Programming Case

In the paper [25] by Luo, Sturm and Zhang, they demonstrate the superlinear convergence of a primal-dual symmetric path-following algorithm for semidefinite programming, only assuming the following assumptions:

1. the existence of positive definite solutions for primal and dual problem (P) and (D) respectively;

2. the existence of a pair of strictly complementary primal-dual optimal solutions; specifically, it means that there exists a feasible primal-dual solution pair (X^*, S^*) such that $X^*S^* = 0$ and $X^* + S^* \succ 0$; and

3. the tangential convergence of the iterates to the central path. In this paper, the assumption of tangential convergence is embedded in the primal-dual path-following predictor-corrector interior-point algorithm they proposed. In this algorithm, the parameter β_k, which is size of a centrality measure, is decreasing in every predictor step of the main iteration instead of being fixed like in [26] and [50] for the Linear Programming case.

In their analysis, they do not assume that the semidefinite program is nondegenerate, which means that their result is established in the absence of the assumption that the Jacobian matrix of its Karush-Kuhn-Tucker (KKT) system is nonsingular. They show that the duality gap is reduced superlinearly after every predictor step if the iterates are sufficiently close to the central path. Especially, the predictor step reduces the duality gap superlinearly with order $\frac{2}{(1+2^{-r})}$, provided that every predictor step is succeeded by r consecutive corrector steps. Therefore, they showed the following convergence result:

Theorem 4.2.1. *Let (X^a, S^a) denote the analytic centre of the primal and dual optimal solution sets. The iterates (X^k, S^k) generated by the algorithm converge to (X^a, S^a) superlinearly with order $\frac{2}{(1+2^{-r})}$. The duality gap μ^k converges to 0 at the same rate.*

The symmetric path-following algorithm they adopt is the primal-dual path following algorithm of Sturm and Zhang [39] using a V-space framework. In this paper, they also adopt the method of having "basic" and "nonbasic" subspaces B and N where every matrix can be written as a block of four submatrices. They characterize the limiting behaviour of the primal-dual central path as μ approaches to 0 and bound the distance from any point on the primal-dual central path to the optimal solution set. Specifically, in Section 3 of [25], they proved that

Theorem 4.2.2. *Let a feasible pair (X_μ, S_μ) denote a point on the central path for some $\mu > 0$, which gives that $X_\mu S_\mu = \mu I$. Let $\mu \in (0, 1)$. Then,*

$$||X_\mu - X^a|| + ||S_\mu - S^a|| = O(\mu),$$

where (X^a, S^a) is the analytic centre of the primal and dual optimal solution sets.

This result can be seen as an error bound result along the central path, while in the case of linear programming, a Hoffman's error bound result [18] is required to prove the quadratic convergence of predictor-corrector interior-point algorithms. Also, they make some connections to the similar results on limiting behaviour of the central path in the case of linear programming and monotone horizontal linear complementarity problem. Hence, this observation leads to the following universal and algorithm-independent property of the central path.

Remark 4.2.3. *Assuming the existence of a pair of strictly complementary optimal solution, it is shown that the primal-dual central path converges to the analytic centre of the optimal solution set. Moreover, the duality gap can bound the distance from any point on the central path to the analytic centre from above.*

On the other hand, in the paper [24] by Kojima, Shida and Shindoh, they establish the superlinear convergence of a Mizuno-Todd-Ye type predictor-corrector infeasible-interior-point algorithm for the monotone semidefinite linear complementarity problems. Indeed, their algorithm starts with a point that is not necessarily strictly feasible. Moreover, this Mizuno-Todd-Ye type predictor-corrector infeasible-interior-point algorithm forces the generated iterates to converge to a solution tangentially to the central surface.

Definition 4.2.4. *The central surface is defined as*

$$\left\{(S,X) \in \mathbb{S}^n_{++} \times \mathbb{S}^n_{++} : XS = \mu I, \text{ for some } \mu > 0\right\}.$$

In the discussion of this paper, they convert every semidefinite program (SDP) into a semidefinite linear complementarity problem (SDLCP) using the following definition.

Definition 4.2.5. *Consider the following set:*

$$\mathcal{F} := \left\{ (S,X) \in \mathbb{S}^n \times \mathbb{S}^n : S = \sum_{i=1}^m A_i y_i - C, \text{ for some } y \in \mathbb{R}^m, \mathrm{Tr}(A_i, X) = b_i, i \in \{1,2,\cdots,m\} \right\}.$$

Notice that each $(S,X) \in \mathcal{F}$ with $X \succeq 0$ and $S \succeq 0$ gives a feasible solution (y,S,X) of the SDP

$$\min \sum_{i=1}^m b_i y_i$$

$$\text{subject to } \sum_{i=1}^m A_i y_i - S = C, \;\; S \succeq 0,$$

and its dual. Then, the corresponding SDLCP is defined as:

$$(S,X) \in \mathcal{F}, \;\; X, S \succeq 0, \;\; \text{and } \langle S, X \rangle = 0.$$

They also define $\mathcal{H} := \{(S,X) \in \mathbb{S}^n \times \mathbb{S}^n : SX = 0\}$. Before we state the assumptions they made, we define the *square root* of a symmetric positive definite matrix.

Definition 4.2.6. *Let $X \in \mathbb{S}^n_{++}$. A* square root *of X is defined as a matrix $S \in \mathbb{S}^n$ such that $S^2 = SS = X$. If X is diagonal, then S is diagonal.*

For $X \in \mathbb{S}^n_{++}$, the square root of X can be computed by using the spectral decomposition $X = Q\Lambda Q^\top$, where Λ is a diagonal matrix whose diagonal entries are the positive eigenvalues of X and Q is an orthogonal matrix. Then, $\sqrt{X} = Q\Lambda^{\frac{1}{2}}Q^\top$.

They impose the following three assumptions to guarantee the superlinear convergence result:

1. a strict complementarity condition, where they assume the existence of a solution (S^*, X^*) of the monotone SDLCP such that $X^* + S^* \succ 0$;

2. a nondegeneracy condition, which means that $\mathcal{F}$ and $\mathcal{H}$ are transversal at some feasible and strict complementary solution (S^*, X^*), i.e,

$$\mathcal{H}_{(S^*,X^*)} \cap \mathcal{F} = \{(S^*, X^*)\}; \;\; and$$

3. the generated sequence $\{(S^k, X^k)\}$ converges to the solution (S^*, X^*) tangentially to the central surface in the sense that

$$\lim_{r\to+\infty} \left|\left| \sqrt{S^k} X^k \sqrt{S^k} - \left(\frac{\langle S^k, X^k\rangle}{n}\right) I \right|\right|_F \Big/ \left(\frac{\langle S^k, X^k\rangle}{n}\right) = 0.$$

They present an example of semidefinite program to illustrate the substantial difficulty in analyzing local convergence of a direct extension of the Mizuno-Todd-Ye type predictor-corrector primal-dual interior-point algorithm. Then, this example suggests an additional assumption that assumes the iteration sequence generated by the algorithm converges tangentially to a solution on the central surface. Similar to [25], during the analysis of local convergence, all matrices were written as a block of four submatrices.

Although the two papers above adopt different types of predictor-corrector primal-dual interior-point algorithms, they are both primal-dual symmetric algorithms. In addition, the algorithm-independent property of the central path mentioned in [25] is connected to the Lemma 3.2 in [30], where they analyzed the upper bound of distance of two points on the central path.

In the paper [15] by Goldfarb and Scheinberg, the papers [16] and [17] by Halická, de Klerk and Roos and the paper [11] by da Cruz Neto, Ferreira and Monteiro, they all investigate the limiting behaviour of the central path and its connections to the analytic centre of the optimal set in the case of Semidefinite Programming.

It is known that the central path always converges to the analytic centre of the optimal set in the case of Linear Programming. Halická, de Klerk and Roos [16] show that the central path does not converge to the analytic centre in general for the SDP case by providing counterexamples in SDP case and Second Order Cone case. In each counterexample, the strict complementarity does not hold and the limit point of the primal central path is different from the analytic centre of the primal optimal face. Moreover, they provide a proof of the convergence of the central path in the case of SDP using results from algebraic geometry.

In addition, they summarize the following common properties that the central path for SDP share with the central path for LP:

1. The central path restricted to $0 < \mu < \bar{\mu}$ for some $\bar{\mu} > 0$ is bounded, and thus it has limit points as $\mu \to 0$ in the optimal set; and
2. The limit points are in the relative interior of the optimal set.

In the paper [17] by Halická, de Klerk and Roos, they analyze the limiting behaviour of the central path in semidefinite optimization. They give the following characterization of strict complementarity and use it to prove the following convergence result of the central path to the analytic centre of a certain subset of the optimal set. Let us first define

$$\left[\tilde{S}_\mu\right]_B := \left(\frac{1}{\mu}\right)[S_\mu]_B \quad \text{and} \quad \left[\tilde{X}_\mu\right]_N := \left(\frac{1}{\mu}\right)[X_\mu]_N .$$

Theorem 4.2.7. *Let* $\left[\tilde{S}_\mu\right]_B$ *and* $\left[\tilde{X}_\mu\right]_N$ *be defined as above. Then, both* $\left[\tilde{S}_\mu\right]_B$ *and* $\left[\tilde{X}_\mu\right]_N$ *converge as* $\mu \to 0$*, and the limit matrices* $\tilde{S}^*_B := \lim_{\mu\to 0}\left[\tilde{S}_\mu\right]_B$ *and* $\tilde{X}^*_N := \lim_{\mu\to 0}\left[\tilde{X}_\mu\right]_N$ *are positive definite. Moreover,* $|B| + |N| = n$ *(strict complementarity holds) if and only if*

$$(\tilde{S}^*_B)^{-1} = X^*_B \ \text{ and } \ (\tilde{X}^*_N)^{-1} = S^*_N$$

Moreover, they provide sufficient conditions for the central path to converge to the analytic centre of the optimal set by describing a class of SDP problems where those conditions are satisfied.

They make the Assumption 3.3.1 throughout this article. They also make an assumption that the data matrices are in the following block diagonal form:

$$A^i = \begin{bmatrix} A^i_U & 0 \\ 0 & A^i_V \end{bmatrix}, \ \ i \in \{1, 2, \ldots, m\}, \ \ C = \begin{bmatrix} C_U & 0 \\ 0 & C_V \end{bmatrix}$$

where $A_U^i, C_U \in \mathbb{S}^s$, for some $s \leq n$.

Denote $S_U(y) := C_U - \sum_{i=1}^m A_U^i y_i$ and $S_V(y) := C_V - \sum_{i=1}^m A_V^i y_i$. Besides Assumption 3.3.1, they assume that:

- There exists $S_U^* \succeq 0$ such that each dual optimal solution (y, S) satisfies $S_U(y) = S_U^*$. Moreover, there exists an optimal solution (y, S) for which $S_V(y) \succ 0$.

Then, they show the following theorem using the above notations.

Theorem 4.2.8. *Let the SDP problem be of the above form and satisfy all three assumptions. Then, the dual central path (y_μ, S_μ) converges to the analytic centre of $\mathcal{D}^*$.*

Last but not least, they show that the convex quadratically constrained quadratic optimization problems ($QCQP$) satisfy these sufficient conditions.

In the paper [11], da Cruz Neto, Ferreira and Monteiro investigate the asymptotic behaviour of the central path (X_μ, S_μ, y_μ) as $\mu \to 0$ for a class of degenerate semidefinite programming (SDP) problems. They study the problems which do not have a strictly complementary primal-dual optimal solutions and whose "degenerate diagonal blocks" $[X_\mu]_\mathcal{T}$ and $[S_\mu]_\mathcal{T}$ of the central path satisfy $\max\{||[X_\mu]_\mathcal{T}||, ||[S_\mu]_\mathcal{T}||\} = O(\sqrt{\mu})$.

Let (X^*, S^*, y^*) be a maximally complementary solution pair and we may assume that

$$X^* = \begin{bmatrix} X_\mathcal{B}^* & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad S^* = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & S_\mathcal{N}^* \end{bmatrix},$$

where $X_\mathcal{B}^* \in \mathbb{S}_{++}^B$ and $S_\mathcal{N}^* \in \mathbb{S}_{++}^N$.

Define $T := \{1, 2, \ldots, n\} \backslash (B \cup N)$. They make the following two assumptions throughout the article:

1. Assumption 3.3.1, and
2. $T \neq \emptyset$, i.e., there exists no strictly complementary primal-dual optimal solution.

They derive estimates on the off-diagonal blocks of the central path and show the convergence of the central path to a primal-dual optimal solution pair, which can be seen as the unique optimal solution of a log-barrier problem. Moreover, they present a characterization of the class of SDP problems which satisfies their initial assumptions on the

degenerate diagonal blocks of the central path. Also, a re-parameterization of the central path is provided and they use it to analyze the limiting behaviour of the derivatives of the central path. The re-parameterization of the central path is defined as follows:

For $t > 0$, let P_t and D_t denote the block diagonal matrices given by

$$P_t := \mathrm{Diag}(I_B, t^{-1}I_T, t^{-2}I_N), \text{ and } D_t := \mathrm{Diag}(t^{-2}I_B, t^{-1}I_T, I_N)$$

Then the re-parameterized central path is given by

$$\left(\tilde{X}(t), \tilde{S}(t)\right) := \left(P_t X(t^4) P_t, D_t S(t^4) D_t\right).$$

Finally, they apply their results to the convex quadratically constrained convex programming problem.

In the above analysis, they also make the following two assumptions and show that they are actually equivalent:

$$[X_\mu]_{\mathcal{T}} = O(\sqrt{\mu}), \text{ and } [S_\mu]_{\mathcal{T}} = O(\sqrt{\mu});$$

$$\left|\left|[X_\mu]_{\mathcal{T}}\right|\right| \left|\left|[S_\mu]_{\mathcal{T}}\right|\right| = O(\mu).$$

Chapter 5

Superlinear Convergence of an algorithm of Nesterov and Tunçel in Linear Programming

From previous work mentioned in Chapter 4, we see that the primal-dual interior-point methods behave well in the case of Linear Programming and Semidefinite Programming. Primal-dual interior-point methods we considered are primal-dual symmetric and these methods are proven to admit quadratic convergence in the case of Linear Programming and superlinear convergence in the case of Semidefinite Programming under mild assumptions. Then, why do we need to investigate and study a new algorithm that works only in the dual space? In those primal-dual symmetric interior-point methods, we need to compute both primal and dual iterates in each iteration, while in the algorithm proposed in [30], we only need to compute the iterates in the dual space for every iteration. Notice that asymmetry may exist between our primal problem (P) and dual problem (D), then computing both primal and dual iterates for every iteration may be inefficient. For instance, if we know that $m << n$, then working only in the dual space may be much easier and much more efficient than working in both primal and dual spaces.

The Algorithm we are studying is a variant in the case of LP of the algorithm proposed in [30]. It is a polynomial-time path-following interior-point algorithm. Unlike other primal-dual symmetric path-following algorithms analyzed in [26] and [50], where in each iteration, the algorithm generates both primal and dual iterates, the algorithm proposed in [30] only generates iterates in the dual space. Moreover, in each iteration of the algorithm we first take a predictor step to reduce duality gap and stay in the larger neighbourhood,

and then take a corrector step to get back to the smaller neighbourhood to get closer to the central path.

First, we define

$$\xi_{\bar{\alpha}}(\alpha) \;:=\; 1+\tfrac{\alpha\bar{\alpha}}{\bar{\alpha}-\alpha}, \quad \alpha\in[0,\bar{\alpha}).$$

Let $(y^{(k)}, s^{(k)})$ be the current iterate. In the case of Linear Programming, the predictor direction is derived as:

$$d_y = \left[AS^{-2}A^\top\right]^{-1} AS^{-1}e.$$

On the other hand, for the corrector direction, we explore various approaches (see Appendix A). The one that we use in our implementation of the algorithm is the first corrector step in Appendix A, i.e.,

$$\Delta_y^{(k)} := \frac{1}{\mu}\left[f''(y^{(k)})\right]^{-1} b - \left[f''(y^{(k)})\right]^{-1} f'(y^{(k)}).$$

In the following algorithm, we use the notion of the neighbourhood for the central path $\mathcal{N}(\mu,\beta)$ defined in (3.1).

Algorithm 1 Algorithm of Nesterov and Tunçel [30] in Linear Programming

Input: A, b, c and $y_0 \in \mathcal{N}(1, \frac{1}{25})$.
Set $\mu_0 := 1$.
For $k \geq 0$, apply the following iterations:

1. Compute $d_{y^{(k)}}$ and $\bar{\alpha}_k := \bar{\alpha}(y^{(k)}) := \max\left\{\alpha \geq 0 : A^\top\left(y^{(k)} + \alpha d_{y^{(k)}}\right) \leq c\right\}$.
2. Use a line search method, find the largest α such that $y(\alpha) = y^{(k)} + \alpha d_{y^{(k)}} \in \mathcal{N}(\frac{\mu_k}{\xi_{\bar{\alpha_k}}(\alpha_k)}, \frac{1}{6})$
3. Set $p^{(k)} = y^{(k)} + \alpha_k d_{y^{(k)}}$, and $\mu_{k+1} = \frac{\mu_k}{\xi_{\bar{\alpha_k}}(\alpha_k)}$.
4. Apply one corrector step to $p^{(k)}$ to find $y^{(k+1)} \in \mathcal{N}(\mu_{k+1}, \frac{1}{25})$.

One of the differences between the above algorithm and the algorithm proposed in [30] is that in this variant, we do not shrink the neighbourhoods in each iteration, i.e., the larger neighbourhood and the smaller neighbourhood stay unchanged throughout the algorithm. Another difference is that we just apply a simple line search in this variant, while the original algorithm uses another proximity measure Γ, which will be defined at the end of this chapter.

The following pictures sketch how this algorithm works. The first figure shows the predictor step and the two following graphs illustrate the corrector step. The red curve represents the central path, the green and the blue curves represent the boundaries of the small and larger neighbourhood respectively. $y^{(k)}$ is the current point, y' denotes the point after taking a predictor step and $y(\mu_{k+1})$ corresponds to the point on the central path with the parameter μ_{k+1}.

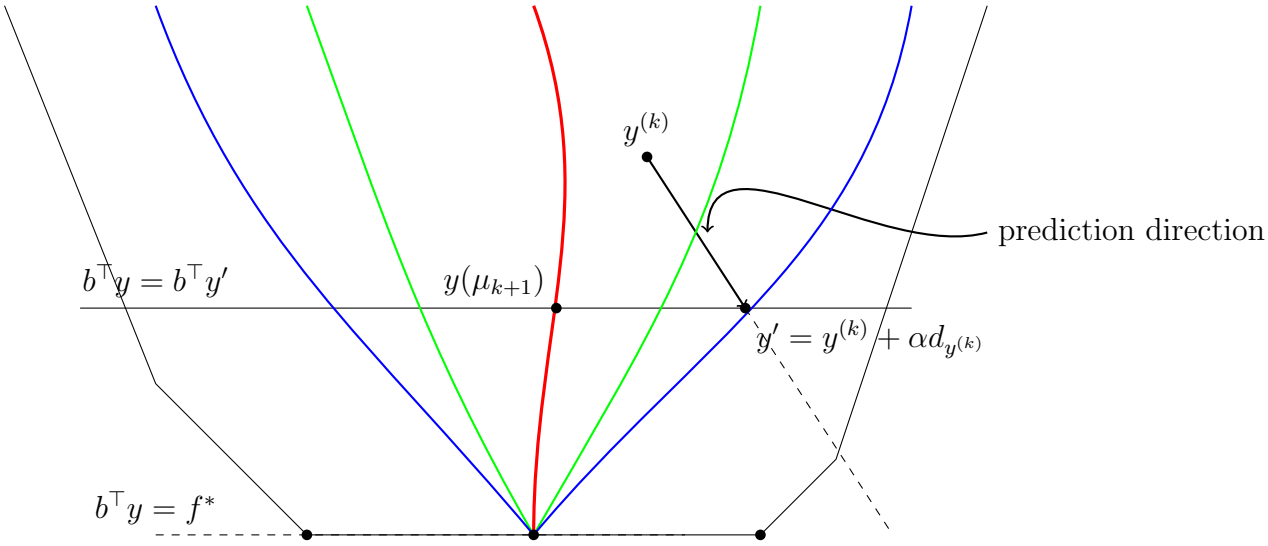


Figure 5.1: Predictor step of the algorithm

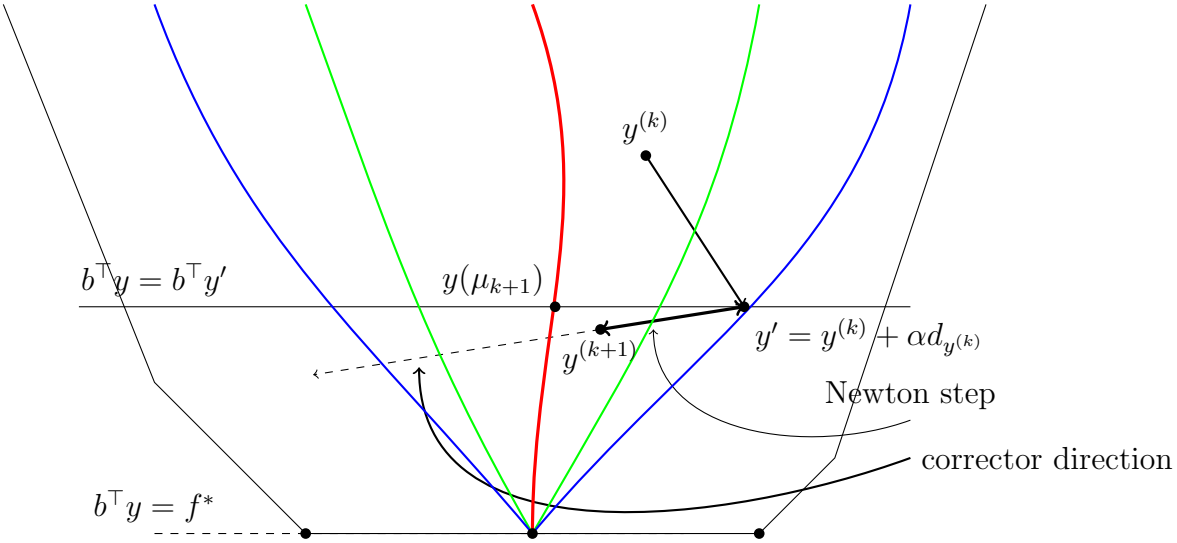


Figure 5.2: Corrector step in the choice of A.1 or A.2 in Appendix A

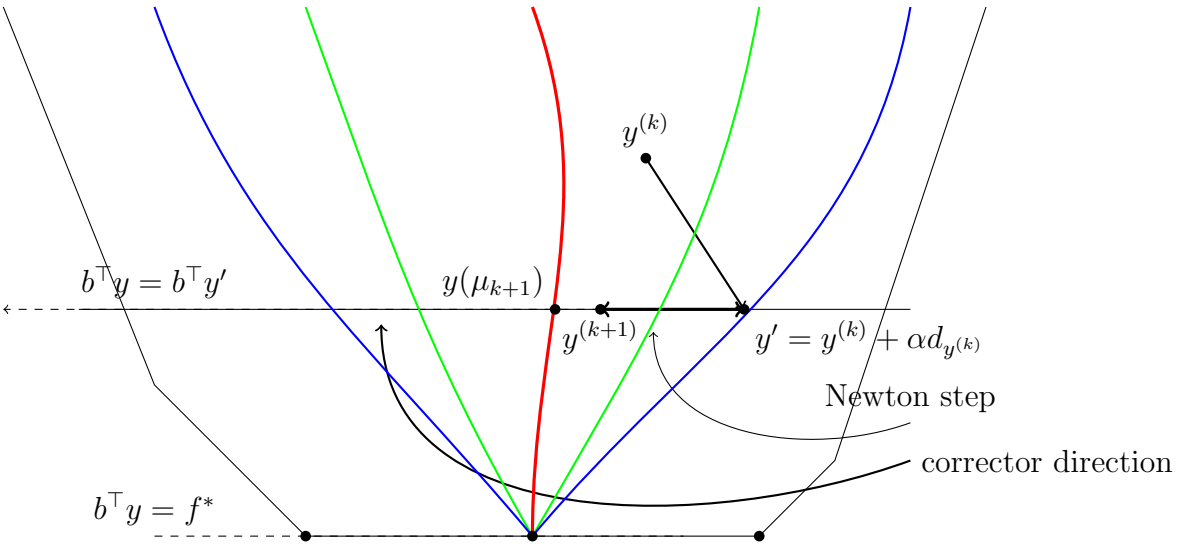


Figure 5.3: Corrector step in the choice of A.3, A.4 or A.5 in Appendix A

According to different choices of corrector step we adopt, the demonstration of the corrector step slightly changes. For the choice of A.1 or A.2, we do not require the new iterate $y^{(k+1)}$ to lie in the affine subspace $b^\top y = b^\top y'$. However, if we adopt A.3, A.4 or

A.5, then we restrict the new iterate $y^{(k+1)}$ to lie in the affine subspace $b^\top y = b^\top y'$. So, in Figure 5.2, $y^{(k+1)}$ could be above, below or on the affine subspace $b^\top y = b^\top y'$. In Figure 5.3, $y^{(k+1)}$ must lie in the affine subspace $b^\top y = b^\top y'$. Note however that in general, the Newton step in Figure 5.3 does not point to $y(\mu_{k+1})$; the corrector directions from A.3, A.4, A.5 only guarantee that $y^{(k+1)}$ lies on the hyperplane $\left\{y : b^\top y = b^\top y'\right\}$.

We will go through the general construction and analysis of the above Algorithm 1 in the special case of LP and wish to prove the following conjecture.

Conjecture. *Assuming the existence of the Slater point(s) of the primal problem (P) and the dual problem (D), Algorithm 1 (a relaxed variant of Nesterov-Tunçel dual interior-point algorithm) converges Q-quadratically.*

Below we propose an approach to prove this Conjecture or a weaker variant of it (e.g. Q-superlinear convergence).

5.1 Proposed approach towards a proof of the conjecture

In order to establish the superlinear (quadratic) convergence property of the Algorithm 1, we need to show the following result

$$f^* - b^\top y^{(k+1)} \leq M \cdot \left(f^* - b^\top y^{(k)}\right)^2$$

where f^* is the optimal objective value of the primal problem (P) and dual problem (D), and M is a constant, and k is large enough.

From Algorithm 1, we have that

$$\begin{aligned}
f^* - b^\top y^{(k+1)} &= f^* - b^\top (y^{(k)} + \alpha_k d_{y^{(k)}}) \\
&= f^* - b^\top y^{(k)} - \alpha_k b^\top d_{y^{(k)}}) + \alpha_k b^\top y^{(k)} - \alpha_k b^\top y^{(k)} \\
&= f^* - (1-\alpha_k) b^\top y^{(k)} - \alpha_k b^\top (d_{y^{(k)}}) + y^{(k)}) \\
&= (1-\alpha_k)\left[f^* - b^\top y^{(k)}\right] + \alpha_k \left[f^* - b^\top (d_{y^{(k)}}) + y^{(k)})\right] \\
&= (1-\alpha_k)\left[f^* - b^\top y^{(k)}\right] + \alpha_k b^\top \left[y_* - (d_{y^{(k)}}) + y^{(k)})\right].
\end{aligned}$$

If $(1-\alpha_k) = O(f^* - b^\top y^{(k)})$ and $\left|\left|d_{y^{(k)}}\right|\right| = O(\mu)$, we can expect Q-superlinear convergence. In order to show that $(1-\alpha_k) = O(f^* - b^\top y^{(k)})$, we first need to prove that if we

only assume the existence of the Slater points of the primal problem (P) and dual problem (D), we can still prove a similar result to the *Assumption NT1* (will be defined in Section 5.2). This part corresponds to the Proposition 5.2.1 in Section 5.2. Then, we want to apply the proof techniques in Theorem 5.1 of [30] to show that $(1-\alpha_k) = O(f^* - b^\top y^{(k)})$, and in this step there is still a gap.

In Section 5.3, we analyze the predictor step by applying the similar proof techniques of Nesterov and Tunçel in [30].

Next, we need to show that $\left|\left|d_{y^{(k)}}\right|\right| = O(\mu)$ and $||d_{s^{(k)}}|| = O(\mu)$. We want to use a primal-dual approach (similar to the proof techniques of Ye, Güler, Tapia, Zhang [50] and Mehrotra [26]) to prove that $\left|\left|d_{y^{(k)}}\right|\right| = O(\mu)$ and $||d_{s^{(k)}}|| = O(\mu)$. In this primal-dual approach, we use a primal-dual proximity measure $\left|\left|\frac{1}{\mu}S^{(k)}x^{(k)} - e\right|\right|_2^2$ to measure the distance from the current iterate $(x^{(k)}, s^{(k)})$ to the central path. However, Algorithm 1 does not generate primal iterates, so we want to associate to each pair $y^{(k)}, \mu_k$ (generated by the algorithm) a primal iterate $x^{(k)}$. In order to use the primal-dual proximity measure $\left|\left|\frac{1}{\mu}S^{(k)}x^{(k)} - e\right|\right|_2^2$ to show $\left|\left|d_{y^{(k)}}\right|\right| = O(\mu)$ and $||d_{s^{(k)}}|| = O(\mu)$, we need to compare the primal-dual proximity measure with the dual proximity measure $\left|\left|\nabla f(y) - \frac{b}{\mu}\right|\right|_y^2$ we used in the Algorithm 1 to see their differences for varying choices of x. This part will be discussed in Section 5.4 and we will prove a related proposition to discuss the correspondence between these two norms. In establishing the explicit relationship between these two norms there is still a gap. In Section 5.4.1, we present some computational experiments to illustrate the differences geometrically and computationally, and to test the validity of the missing step in finding the explicit relationship between these two norms. In Section 5.5, we prove Lemma 5.5.1, 5.5.2 and 5.5.3 and therefore use these lemmas to show that $\left|\left|d_{y^{(k)}}\right|\right| = O(\mu)$ and $||d_{s^{(k)}}|| = O(\mu)$.

We will use the notion of *strict complementarity* and *analytic centre* for Linear Programming we defined in 3.4.2 and 3.4.4 in the following analysis.

5.2 Towards weaker assumptions

In Section 4 of [30], there are two main assumptions (Assumption NT1 & NT2) established for the general conic programming case.

In the case of LP, we only assume that the primal problem (P) and dual problem (D)

have Slater points.

In the analysis of [30], they have the following assumption (Assumption NT1) about the uniqueness and sharpness of the optimal solution of the dual problem.

Assumption NT1. *The dual problem has a unique optimal solution y_* and there exists a constant $\gamma_d > 0$ such that*

$$f^* - \langle b, y\rangle = \langle s, x_*\rangle \geq \gamma_d \, ||y - y_*||$$

for every y feasible for dual problem.

However, we cannot apply this Assumption NT1 in our analysis in the special case of LP because we only assume the existence of Slater points of the primal and dual problems for the LP case. So, we need to prove a similar result only using the existence of Slater points of the primal and dual problems.

Let $\mathbb{Y}^*$ denote the set of optimal solution(s) of the dual problem (D). From our assumption, we know that $\mathbb{Y}^* \neq \emptyset$.

Proposition 5.2.1. *There exists a constant $\gamma_d' > 0$, depending on the data (A, b, c) such that*

$$f^* - \langle b, y\rangle \geq \gamma_d' \cdot \text{dist}(y, \mathbb{Y}^*),$$

for every feasible solution y of the dual problem (D).

Proof. If $y \in \mathbb{Y}^*$, then both sides of the inequality evaluate to zero. Therefore, we are done. So, we may assume that y is not an optimal solution of (D). Let y_* denote the closest point to y in $\mathbb{Y}^*$. Such a $y_* \in \mathbb{Y}^*$ exists, and is unique, since $\mathbb{Y}^*$ is nonempty, closed and convex. For any vector a, let $\angle(a, b)$ denote the angle from vector a to the vector b. Here, we may assume that $\angle(a, b) \in [0, \pi]$. Then, we have that

$$\begin{aligned} f^* - \langle b, y\rangle &= \langle b, y_*\rangle - \langle b, y\rangle \\ &= \langle b, y_* - y\rangle \\ &= ||b||_2 \, ||y_* - y||_2 \cos(\angle(y_* - y, b)) \\ &= [||b||_2 \cos(\angle(y_* - y, b))] \, ||y_* - y||_2 \, . \end{aligned}$$

We claim that $\cos(\angle(y_* - y, b)) > 0$. We know that b is orthogonal to the affine subspace where $\mathbb{Y}^*$ lies in and b acts as a normal vector of that subspace. If $\angle(y_* - y, b) \in [\frac{\pi}{2}, \pi]$, then it means that y is not in the open half-space defined by $b^\top y < f^*$. This contradicts

the fact that y is feasible but not optimal. Hence, $\angle(y_* - y, b) \in (0, \frac{\pi}{2})$. Therefore, $\cos(\angle(y_* - y, b)) > 0$.

By the definition of y_*, $||y_* - y||_2 = \text{dist}(y, \mathbb{Y}^*)$, and then we can define

$$\gamma'_d := ||b||_2 \cdot \inf_{y \text{ is feasible}} \cos(\angle(y_* - y, b)).$$

If $b = 0$, then $\langle b, y\rangle = 0$ for all y. Then, all feasible y's are optimal solutions for the dual problem (D). Hence, $f^* - \langle b, y\rangle = 0 = \text{dist}(y, \mathbb{Y}^*)$ and then any positive γ'_d will work. So, we may assume that $b \neq 0$. Therefore, $||b||_2 > 0$. In order to show that γ'_d is always positive, we need to show that $\inf_{y \text{ is feasible}} \cos(\angle(y_* - y, b)) > 0$.

Define $P := \left\{y \in \mathbb{R}^m : A^\top y \leq c\right\}$. We say that a facet F of P is *incident on* a proper face G of P, if $F \cap G \neq \emptyset$.

Let G be the optimal face of P (corresponding to the objective function "$\max b^\top y$"). Let $\mathcal{F}$ denote the set of facets of P that are incident on G and let $\tilde{A}^\top y \leq \tilde{c}$ denote the subset of inequalities in the system $A^\top y \leq c$, representing the facets in $\mathcal{F}$. Then, we define

$$\tilde{P} := \left\{y \in \mathbb{R}^m : \tilde{A}^\top y \leq \tilde{c}, b^\top y \leq f^*\right\} \supseteq P,$$

and we have

$$\text{argmax}\left\{b^\top y : y \in \tilde{P}\right\} = \text{argmax}\left\{b^\top y : y \in P\right\} = G.$$

Therefore, it suffices to prove the claim for $\tilde{P}$.

Note that $\tilde{P} = G + K$, where K is a pointed polyhedral cone. Every extreme ray of K is a solution to a linear system of equations $\tilde{A}_T^\top y = \tilde{c}_T$, where $|T| = m - 1$. Let $y^{(1)} \in \tilde{P}$. Then, every vector of the form $(y^{(1)} - y_*)$ is a nonnegative linear combination of extreme rays of K. Thus, the worst angle between $(y^{(1)} - y_*)$ and b, where $y^{(1)}$ varies over all points in $\tilde{P}$, is attained by an extreme ray of K. Since there are finitely many extreme rays of K and for every ray $\frac{y - y_*}{||y - y_*||}$ of K, we have $b^\top \frac{y - y_*}{||y - y_*||} < 0$, we are done.

Then, $\inf_{y \text{ is feasible}} \cos(\angle(y_* - y, b)) > 0$. Therefore, the desired γ'_d exists. □

Some elements of the proof are shown on the following picture.

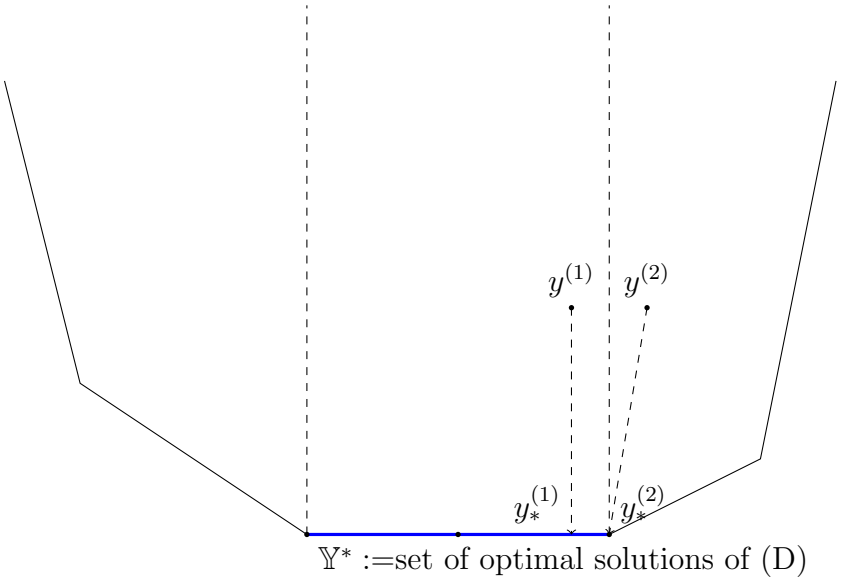


Figure 5.4: An illustration for the proof of Proposition 5.2.1

5.3 Analysis of the predictor step

Now, I analyze the norm of the predictor step in the case of Linear programming where we only assume the existence of the Slater points for primal and dual problem.

Let $y^*(s^*)$ be the analytic centre of the optimal face in the y-space (and s-space respectively). From [30], we know that

$$p(y) = y + v(y) = y^* + \left[\nabla^2 f(y)\right]^{-1} A\nabla^2 F_*(s(y))s^*.$$

So, we start by bounding the norm of $\nabla^2 F_*(s(y))s^*$.

Recall that we denote by $x^* \in \mathbb{R}^n$ the limit point of the primal central path and by $s^* \in \mathbb{R}^n$ the limit point of the dual central path. Moreover, x^* is the analytic centre of P^* and s^* is the analytic centre of D^*. We define $T := \nabla^2 F(x_\mu)$ with $\mu = 1$.

Lemma 5.3.1. *(Lemma 3.2 in [30]) If $\mu_1 \in (0, \mu_0]$, then*

$$||x_{\mu_1}||_{x_{\mu_0}} \leq n, \quad ||s_{\mu_1}||_{s_{\mu_0}} \leq n.$$

In particular, for every $x^* \in X^*$ and every $s^* \in S^*$ we have:

$$||x^*||_T \leq n, \quad ||s^*||_T \leq n.$$

Moreover, if $\mu \in (0,1]$, *then*

$$\frac{1}{4n^2}T \preceq \nabla^2 F(x_\mu) \preceq \frac{4n^2}{\mu^2}T$$

$$\frac{1}{4n^2}T^{-1} \preceq \nabla^2 F(s_\mu) \preceq \frac{4n^2}{\mu^2}T^{-1}.$$

The first part of the above lemma, in the LP case, is due to Vavasis and Ye [47].

Lemma 5.3.2. *(Lemma 3.3 in [30]) We define* $G := AT^{-1}A^\top : \mathbb{R}^m \to \mathbb{R}^m$. *We have*

$$||A||_{G,T} := \max_{h\in\mathbb{R}^n}\{||Ah||_G : ||h||_T = 1\} \leq 1$$

$$\left|\left|A^\top\right|\right|_{T,G} := \max_{y\in\mathbb{R}^m}\{\left|\left|A^\top y\right|\right|_T : ||y||_G = 1\} \leq 1$$

$$||b||_G \leq n^{\frac{1}{2}}.$$

Lemma 5.3.3. *Let* s^* *be the analytic centre of the optimal face in the s-space and* $N := \{j : s^*_j > 0\}$. *Then,*

$$\left|\left|\nabla^2 F_*(s(y))s^*\right|\right|_2 \leq \max_{i\in N}\frac{n}{\sqrt{s^*_{(i)}}}\max_i\left\{\frac{\sqrt{s^*_{(i)}}}{[s^{(i)}_\mu]}\right\}.$$

Proof.

$$\begin{aligned}
\langle e, \nabla^2 F_*(s(y))s^*\rangle &= \sum_{i\in N}\frac{s^*_{(i)}}{[s^{(i)}_\mu]^2} \\
&= \sum_{i\in N}\frac{1}{s^*_{(i)}}[\frac{s^*_{(i)}}{s^{(i)}_\mu}]^2 \\
&\leq ||s^*||^2_{s_\mu}\max_{i\in N}\frac{1}{s^*_{(i)}} \\
&\leq \max_{i\in N}\frac{n^2}{s^*_{(i)}}.
\end{aligned}$$

Last inequality holds because of the Lemma 5.3.1, which does not require Assumption NT1 or NT2 in [30].

Therefore,

$$\begin{aligned}\left|\left|\nabla^2 F_*(s(y))s^*\right|\right|_2^2 &\leq \max_{i\in N}\left\{\frac{s^*_{(i)}}{[s^{(i)}_\mu]^2}\right\}\langle e, \nabla^2 F_*(s(y))s^*\rangle \\ &\leq \max_{i\in N}\frac{n^2}{s^*_{(i)}}\max_{i\in N}\left\{\frac{s^*_{(i)}}{[s^{(i)}_\mu]^2}\right\}\end{aligned}$$

It gives

$$\left|\left|\nabla^2 F_*(s(y))s^*\right|\right|_2 \leq \max_{i\in N}\frac{n}{\sqrt{s^*_{(i)}}}\max_{i\in N}\left\{\frac{\sqrt{s^*_{(i)}}}{[s^{(i)}_\mu]}\right\} =: \sigma'_d$$

□

By Lemma 5.3.2, we know that

$$||A||_{G,T} \leq 1,$$

and this is proved without assuming Assumption NT1 or NT2 in [30]. Now, I try to find a bound for the norm of $[\nabla^2 f(y)]^{-1}$.

Lemma 5.3.4. *For $\mu \in (0, 1]$,*

$$\left[\nabla^2 f(y_\mu)\right]^{-1} \preceq 4n^2 G^{-1}.$$

Hence,

$$\left|\left|\left[\nabla^2 f(y_\mu)\right]^{-1}\right|\right|_G \leq 4n^2.$$

Proof. By Lemma 5.3.1 in [30], we know that

$$\frac{1}{4n^2}T^{-1} \preceq \nabla^2 F_*(s_\mu) \preceq \frac{4n^2}{\mu^2}T^{-1}.$$

Then, we can have

$$\frac{1}{4n^2}AT^{-1}A^\top \preceq A\nabla^2 F_*(s_\mu)A^\top \preceq \frac{4n^2}{\mu^2}AT^{-1}A^\top.$$

It gives us

$$\frac{1}{4n^2}AT^{-1}A^\top \preceq \nabla^2 f(y_\mu) \preceq \frac{4n^2}{\mu^2}AT^{-1}A^\top.$$

Hence,

$$\left[\nabla^2 f(y_\mu)\right]^{-1} \preceq 4n^2 \left(AT^{-1}A^\top\right)^{-1} = 4n^2 G^{-1}.$$

According to the definition,

$$\left|\left|\left[\nabla^2 f(y_\mu)\right]^{-1}\right|\right|_G = \max_{h\in\mathbb{R}^m} \left\{\left|\left|\left[\nabla^2 f(y_\mu)\right]^{-1} h\right|\right|_G : ||h||_G = 1\right\}.$$

Then, we have that

$$\begin{aligned}
\left|\left|\left[\nabla^2 f(y_\mu)\right]^{-1} h\right|\right|_G^2 &= \langle h, \left[\nabla^2 f(y_\mu)\right]^{-1} G \left[\nabla^2 f(y_\mu)\right]^{-1} h\rangle \\
&\leq 4n^2 \langle h, \left[\nabla^2 f(y_\mu)\right]^{-1}\rangle \\
&\leq (4n^2)^2 \langle h, G^{-1} h\rangle \\
&= (4n^2)^2.
\end{aligned}$$

□

Hence, we can bound the norm of $\left[\nabla^2 f(y)\right]^{-1} A\nabla^2 F_*(s(y))s_*$, which is an important part of the superlinear convergence proof in [30].

5.4 Auxiliary primal sequence and comparing proximity measures for centrality

Algorithm 1 does not explicitly generate a primal sequence $x^{(k)}$. Hence, in theory, we may associate to each $y^{(k)}, \mu_k$ (generated by the algorithm) an x, namely $x(\mu_k)$. For each μ_k generated by the algorithm, we find the corresponding $x(\mu_k)$ by considering the problem (P_{μ_k}) defined by this specific μ_k and obtaining the minimizer $x(\mu_k)$ of (P_{μ_k}) , i.e.,

$$x(\mu_k) := \text{argmin}(P_{\mu_k}) = \text{argmin}\left\{\frac{1}{\mu_k} c^\top x - \sum_{j=1}^n \ln x_j : Ax = b\right\}$$

For simplicity, we drop the index k.

We denote by Π_L the orthogonal projection onto the linear subspace L of $\mathbb{R}^n$, and denote by $R(A^\top)$ the range of $A^\top$.

The iterates generated by our algorithm satisfy:

$$\left|\left|\nabla f(y) - \frac{1}{\mu}b\right|\right|_y \leq \beta$$

Then, we have

$$\begin{aligned}\left|\left|\nabla f(y) - \frac{1}{\mu}b\right|\right|_y^2 &= \left(\nabla f(y) - \frac{1}{\mu}b\right)^\top \left[AS^{-2}A^\top\right]^{-1}\left(\nabla f(y) - \frac{1}{\mu}b\right) \\ &= \left(AS^{-1}e - \frac{1}{\mu}b\right)^\top \left[AS^{-2}A^\top\right]^{-1}\left(AS^{-1}e - \frac{1}{\mu}b\right) \leq \beta^2. \quad (5.1)\end{aligned}$$

If we have $S = S(\mu)$, then

$$A\left[S(\mu)\right]^{-1}e - \frac{1}{\mu}b = 0.$$

Then, (5.1) is equivalent to

$$\begin{aligned}&\left(A\left(S^{-1} - S_\mu^{-1}\right)e\right)^\top \left[AS^{-2}A^\top\right]^{-1}\left(A\left(S^{-1} - S_\mu^{-1}\right)e\right) \\ =\ &\left(A\left(S^{-1} - \frac{1}{\mu}X_\mu\right)e\right)^\top \left[AS^{-2}A^\top\right]^{-1}\left(A\left(S^{-1} - \frac{1}{\mu}X_\mu\right)e\right) \\ =\ &e^\top\left(I - \frac{1}{\mu}SX_\mu\right)(AS^{-1})^\top\left[AS^{-2}A^\top\right]^{-1}AS^{-1}\left(I - \frac{1}{\mu}SX_\mu\right)e \\ =\ &e^\top\left(I - \frac{1}{\mu}SX_\mu\right)\Pi_{R(S^{-1}A^\top)}\left(I - \frac{1}{\mu}SX_\mu\right)e \leq \beta^2. \quad (5.2)\end{aligned}$$

Let $h := \frac{1}{\mu}Sx_\mu - e$. Then, we have

$$h^\top \Pi_{R(S^{-1}A^\top)} h \leq \beta^2.$$

Note that $||h|| = \left|\left|\frac{1}{\mu}Sx_\mu - e\right|\right|$ measures the deviation from the central point (x_μ, s_μ). On the other hand, $\left|\left|\nabla f(y) - \frac{1}{\mu}b\right|\right|$ also measures the deviation from the central path since

$$\nabla f(y) - \frac{1}{\mu}b = 0 \ \Leftrightarrow\ AS^{-1}e = \frac{1}{\mu}b \ \Leftrightarrow\ A(\mu S^{-1}e) = b.$$

As a result,

$$\left|\left|\nabla f(y) - \frac{1}{\mu}b\right|\right| = 0 \ \Leftrightarrow\ ||h|| = 0.$$

In particular, we proved

Proposition 5.4.1. *Let $y \in \mathbb{R}^m$ be an interior point, i.e. $A^\top y < c$ and $\mu > 0$. Then with $s := c - A^\top y$, we have*

$$\left|\left|\frac{1}{\mu}Sx_\mu - e\right|\right|_2^2 = \left|\left|\nabla f(y) - \frac{1}{\mu}b\right|\right|_2^2 + \left|\left|\Pi_{\mathrm{Null}(AS^{-1})}\left(\frac{1}{\mu}Sx_\mu - e\right)\right|\right|_2^2.$$

We want to investigate the relationship between the primal-dual proximity measure $\left|\left|\frac{Sx}{\mu} - e\right|\right|_2$ and the proximity measure $\left|\left|\nabla f(y) - \frac{1}{\mu}b\right|\right|_y$ in the dual space. By continuity and the above results, we know that when $\left|\left|\nabla f(y) - \frac{1}{\mu}b\right|\right|_y$ is very close to 0, then $\left|\left|\frac{Sx_\mu}{\mu} - e\right|\right|_2$ is also very close to 0. However, to justify the choice of x_{μ_k} as the auxiliary primal iterate (corresponding to $y^{(k)}, \mu_k$), we also want to see if there may be better choices of primal feasible solutions x that might act as such an auxiliary iterate. So, in the next proposition, we consider the problem of finding best x in the affine subspace $Ax = b$ minimizing the primal-dual proximity measure for a given pair of s and μ.

Proposition 5.4.2. *For every $s \in \mathbb{R}^n_{++}$ and every $\mu > 0$,*

$$\min\left\{\left|\left|\frac{Sx}{\mu} - e\right|\right|_2 : Ax = b\right\} = \left|\left|\Pi_{R\left((AS^{-1})^\top\right)}\left(\frac{Sx_\mu}{\mu} - e\right)\right|\right|_2.$$

Moreover, the minimizer is unique and it is given by

$$\bar{x} = \mu S^{-2}A^\top\left(AS^{-2}A^\top\right)^{-1}\left[\frac{1}{\mu}b - AS^{-1}e\right] + \mu S^{-1}e,$$

and with this $\bar{x}$,

$$\frac{1}{\mu}S\bar{x} - e \in R((AS^{-1})^\top).$$

Proof. Consider the following minimization problem:

$$(P_1) \quad \min \quad \left|\left|\frac{Sx}{\mu} - e\right|\right|_2^2 \\ Ax = b.$$

Since we assume that the primal problem (P) has a Slater point, (P_1) is always feasible. Moreover, since the hessian of the objective function is $S^2 \succ 0$, we are minimizing a strictly

convex function over a nonempty affine space, the unique minimizer always exists. Indeed (P_1) is equivalent to a closest point problem:

$$\begin{aligned} \min \quad & ||\eta - e||_2^2 \\ & AS^{-1}\eta = \frac{1}{\mu} b. \end{aligned}$$

When we expand the objective function, we get

$$\left|\left| \frac{Sx}{\mu} - e \right|\right|_2^2 = \frac{1}{\mu^2} x^\top S^2 x - \frac{2}{\mu}\langle s, x\rangle + n.$$

By using KKT conditions, we obtain the following optimality conditions:

$$\text{There exist } \bar{x} \in \mathbb{R}^n, z \in \mathbb{R}^m \text{ such that}$$

$$\begin{aligned} \frac{2}{\mu^2} S^2 \bar{x} - \frac{2}{\mu} s &= A^\top z \\ A\bar{x} &= b. \end{aligned} \tag{5.3}$$

The above optimality conditions (5.3) are equivalent to

$$A\bar{x} = b, \text{ and } \frac{1}{\mu} S\bar{x} - e = S^{-1}A^\top \left(\frac{\mu}{2} z\right) \in R((AS^{-1})^\top). \tag{5.4}$$

Let us define $y := \frac{\mu}{2} z$. Then, we want to solve $\bar{x}$ and y from the optimality conditions (5.3). The optimality conditions (5.3) are equivalent to

$$\begin{aligned} \frac{1}{\mu} A\bar{x} - AS^{-1}e &= \left(AS^{-2}A^\top\right) y \\ A\bar{x} &= b. \end{aligned}$$

By substituting the second equation into the first, we get

$$\frac{1}{\mu} b - AS^{-1}e = \left(AS^{-2}A^\top\right) y.$$

Therefore, we solve for y :

$$y = \left(AS^{-2}A^\top\right)^{-1} \left[\frac{1}{\mu} b - AS^{-1}e\right].$$

Substituting y back gives

$$\bar{x} = \mu S^{-2} A^\top \left(AS^{-2}A^\top\right)^{-1} \left[\frac{1}{\mu} b - AS^{-1}e\right] + \mu S^{-1} e.$$

So, we showed that such $\bar{x}$ always exists and we gave a formula for it.

Moreover, from (5.4) we know that

$$\frac{1}{\mu} S\bar{x} - e \in R((AS^{-1})^\top).$$

Then, it means that, for this special choice of $\bar{x}$

$$\left\| \frac{S\bar{x}}{\mu} - e \right\|_2 = \left\| \Pi_{R\left((AS^{-1})^\top\right)} \left(\frac{S\bar{x}}{\mu} - e \right) \right\|_2.$$

Note that $A(x_\mu - \bar{x}) = 0$. Therefore,

$$\left(AS^{-1}\right) \left(\frac{1}{\mu} S x_\mu - \frac{1}{\mu} S\bar{x} \right) = 0.$$

Thus,

$$\begin{aligned} \Pi_{R\left((AS^{-1})^\top\right)} \left(\frac{S\bar{x}}{\mu} - e \right) &= \Pi_{R\left((AS^{-1})^\top\right)} \left(\frac{S\bar{x}}{\mu} + \frac{1}{\mu} S x_\mu - \frac{1}{\mu} S\bar{x} - e \right) \\ &= \Pi_{R\left((AS^{-1})^\top\right)} \left(\frac{S x_\mu}{\mu} - e \right). \end{aligned}$$

□

Notice that from the formula for $\bar{x}$, we know that $\mu S^{-1} e$ is always positive and $\bar{x}$ is the unique minimizer of (P_1). If $S = S_\mu$, then $\bar{x} = x_\mu$.

If y is very close to y_μ, i.e. $\left[\frac{1}{\mu} b - AS^{-1}e\right]$ is very small and s is very close to s_μ, then $\bar{x}$ given by the above formula is very close to $\mu S^{-1} e$ (hence, $\bar{x}$ is very close to x_μ).

5.4.1 Experiments with two proximity measures

Besides theoretical analysis of the relationship between these two norms, we can run different instances using the following algorithm and compute the two proximity measures to see the difference. In these experiments, we adopt the following algorithm:

Algorithm 2

1. Pick a randomly generated matrix A using a uniform distribution $rand(m, n)$ in MATLAB.
2. Let $b := Ae$ and $c := e$. Then, we set $y^{(0)} := 0$ and $s^{(0)} := c - A^\top y^{(0)} = e$.
3. Compute $\bar{\alpha} = \bar{\alpha}(y^{(0)}) = \max\left\{\alpha \geq 0 : A^\top \left(y^{(0)} + \alpha v^{(0)}\right) \leq c\right\}$.
4. Use bi-section searching method to find the largest α such that $y(\alpha) = y^{(0)} + \alpha d_{y^{(0)}} \in \mathcal{N}(\frac{1}{\xi_{\bar{\alpha}}(\alpha)}, \frac{1}{6})$, where $\xi_{\bar{\alpha}}(\alpha) := 1 + \frac{\alpha\bar{\alpha}}{\bar{\alpha}-\alpha}$, $\alpha \in [0, \bar{\alpha})$ and $d_{y^{(0)}} = \left[AS^{-2}A^\top\right]^{-1} AS^{-1}e$.
5. Set $y^{(1)} = y^{(0)} + \alpha d_{y^{(0)}}$, and update $\mu = \frac{1}{\xi_{\bar{\alpha}}(\alpha)}$.
6. Apply corrector steps to $y^{(k)}$ for $k = 1, 2, \ldots$ until its local norm, i.e., $\left|\left|\nabla f(y^{(K)}) - \frac{1}{\mu}b\right|\right|^2_{y^{(K)}} \leq 10^{-10}$. Compute the corresponding $x_\mu := \mu[S^{(K)}]^{-1}e$
7. Compute the primal-dual proximity measure $\left|\left|\frac{S^{(1)}x_\mu}{\mu} - e\right|\right|^2_2$ and the difference between two norms' squares, i.e., $\left|\left|\Pi_{\text{Null}(A[S^{(1)}]^{-1})}(\frac{S^{(1)}x_\mu}{\mu} - e)\right|\right|^2_2$.

Note that in this algorithm, we always have $\left|\left|\nabla f(y^{(1)}) - \frac{1}{\mu}b\right|\right|^2_{y^{(1)}} \approx \frac{1}{6} \approx 0.0277777777777778$. For each size of the matrix A, we run 100 instances and record the average value of the quantity $\left|\left|\frac{S^{(1)}x_\mu}{\mu} - e\right|\right|^2_2$ and the maximum value of $\left|\left|\Pi_{\text{Null}(A[S^{(1)}]^{-1})}(\frac{S^{(1)}x_\mu}{\mu} - e)\right|\right|^2_2$ over these instances.

m	n	average $\left\|\left\|\frac{S^{(1)}x_\mu}{\mu} - e\right\|\right\|_2^2$	maximum value of $\left\|\left\|\Pi_{\text{Null}(A[S^{(1)}]^{-1})}(\frac{S^{(1)}x_\mu}{\mu} - e)\right\|\right\|_2^2$ over 100 instances
100	400	0.0277778180665973	4.12280280991106e-008
100	400	0.027777817916397	4.23151819556322e-008
100	800	0.0277777931297674	1.62283292359788e-008
100	800	0.0277777931313703	1.61378171832649e-008
200	400	0.0277778130403595	3.62532813560912e-008
200	400	0.0277778130904213	3.64600384571145e-008
200	800	0.0277777909218174	1.36622731461422e-008
200	800	0.0277777909008845	1.37325309179648e-008

Moreover, we want to use a tiny example and then plot the corresponding region of these two norms to visualize the difference.

For example, we fix $m = 2$ and $n = 4$. Choose $A := \begin{bmatrix} a_1 & a_2 & 1 & 0 \\ a_3 & a_4 & 0 & 1 \end{bmatrix}$, where $a_i, i = 1, 2, 3, 4$ are randomly generated using randn(1,1) in MATLAB. Set $b := Ae$ and $c := e$. When $\mu = 1$, we have $s_\mu = x_\mu = e$ and then $y = 0$. In the dual space, the point $y = 0$ is obviously feasible and it is an interior point since it is on the central path. Then, we want to find all feasible $\bar{y}$ such that $||\nabla f(\bar{y}) - Ae||_{\bar{y}} \leq \beta$. On the other hand, we want to find all feasible $\hat{y}$ such that $\left\|\left\|\hat{S}x_1 - e\right\|\right\|_2 \leq \beta$, where $\hat{s}(y) = c - A^\top \hat{y}$.

For the following figures, we try different combinations of a_1, a_2, a_3 and a_4 and set $\beta = \frac{1}{6}$. The intersection of all $\bar{y}$ and $\hat{y}$ is in dark shade, the collection of all points that are in $\hat{y}$ but not in $\bar{y}$ is in lighter shade and all other points will not be shaded.

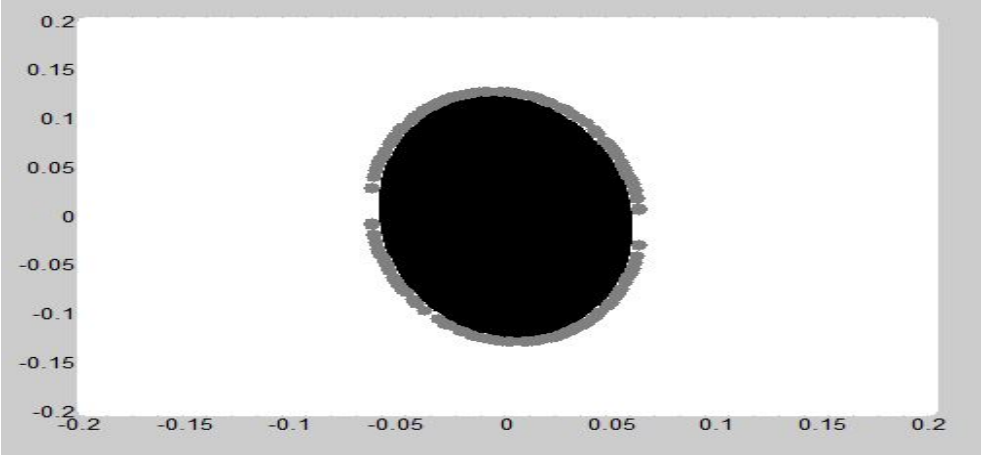


Figure 5.5: The collection of all $\bar{y}$ and $\hat{y}$ when $\beta = \frac{1}{6}$

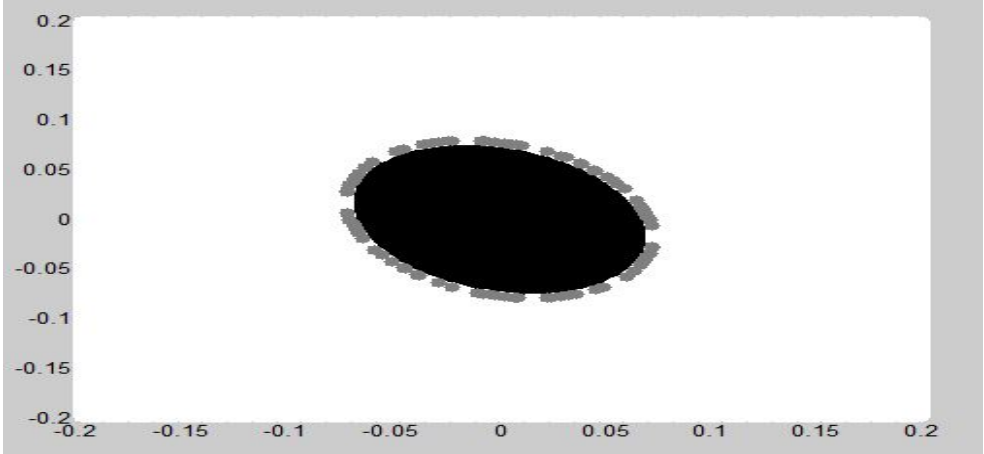


Figure 5.6: The collection of all $\bar{y}$ and $\hat{y}$ when $\beta = \frac{1}{6}$

Similarly, we reduce β from $\frac{1}{6}$ to $\frac{1}{10}$ and run the experiments with different combinations of a_1, a_2, a_3 and a_4.The intersection of all $\bar{y}$ and $\hat{y}$ is in dark shade, the collection of all points that are in $\hat{y}$ but not in $\bar{y}$ is in lighter shade and all other points will not be shaded.

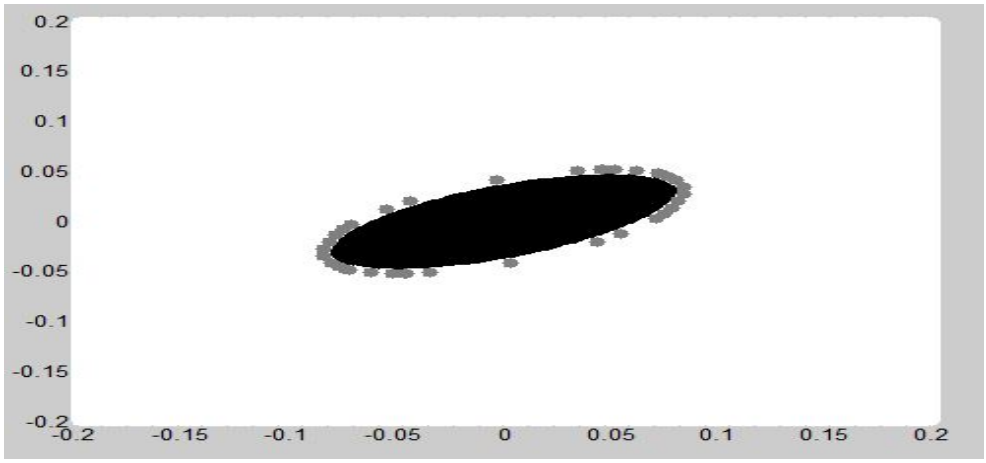


Figure 5.7: The collection of all $\bar{y}$ and $\hat{y}$ when $\beta = \frac{1}{10}$

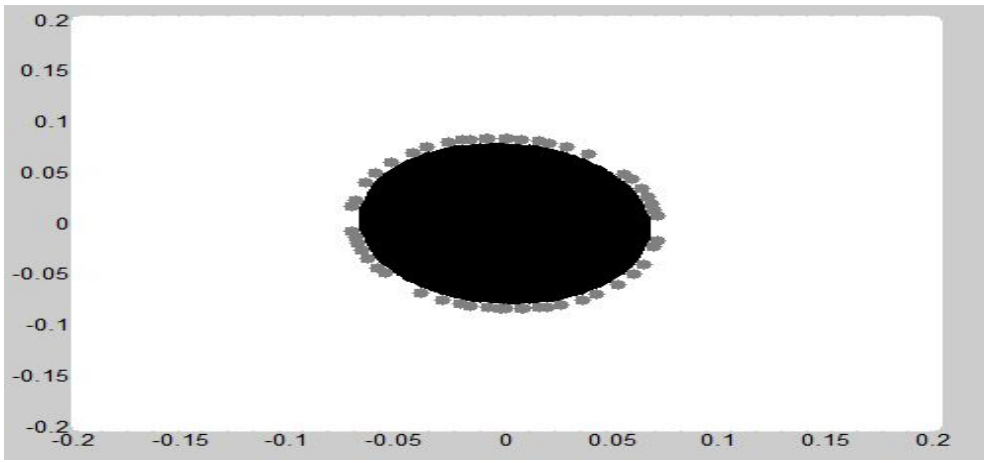


Figure 5.8: The collection of all $\bar{y}$ and $\hat{y}$ when $\beta = \frac{1}{10}$

From the above figures, we can see that, in this instance with tiny size, there is a small region of points y such that it is in $\hat{y}$ but not $\bar{y}$. Compared to the area of the intersection of all $\bar{y}$ and $\hat{y}$, the area of the symmetric difference of $\bar{y}$ and $\hat{y}$ is relatively small.

Also, we change β to $\frac{1}{2}$ and $\frac{2}{3}$, and draw the collection of $\bar{y}$ and $\hat{y}$. The intersection of all $\bar{y}$ and $\hat{y}$ is in dark shade, the collection of all points that are in $\hat{y}$ but not in $\bar{y}$ is in lighter shade and all other points will not be shaded.

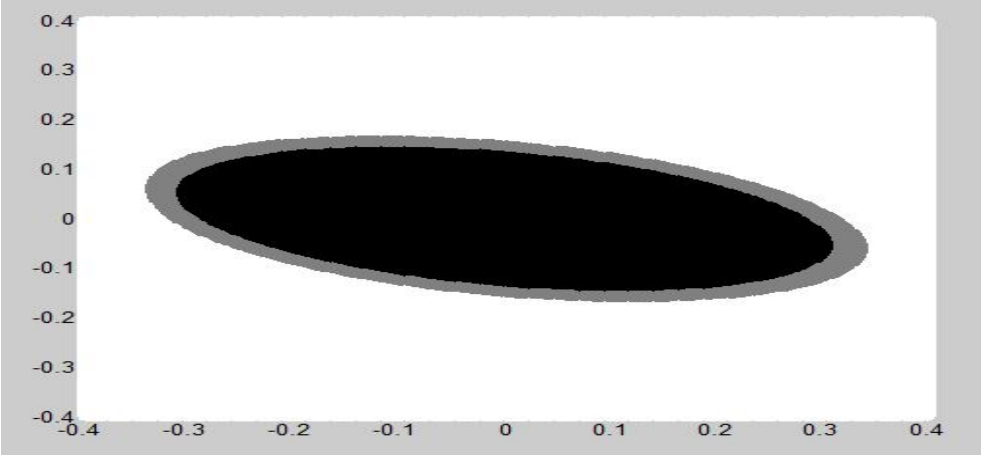


Figure 5.9: The collection of all $\bar{y}$ and $\hat{y}$ when $\beta = \frac{1}{2}$

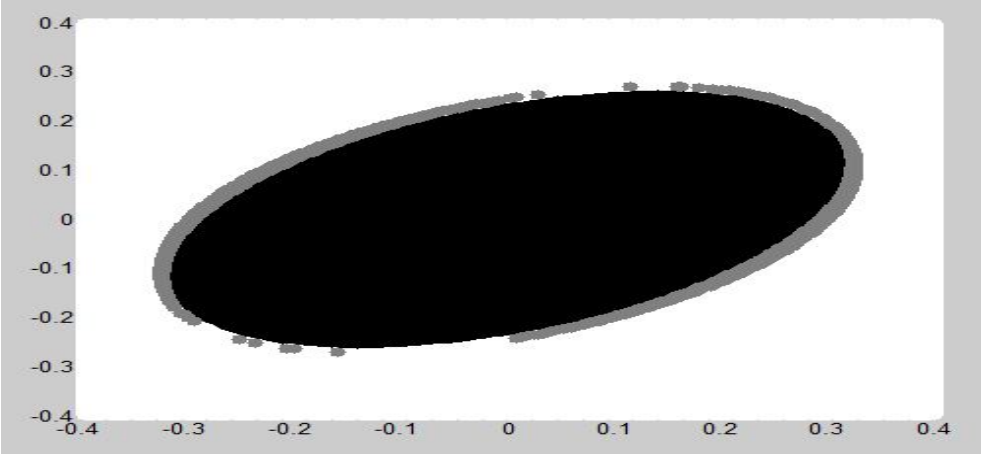


Figure 5.10: The collection of all $\bar{y}$ and $\hat{y}$ when $\beta = \frac{1}{2}$

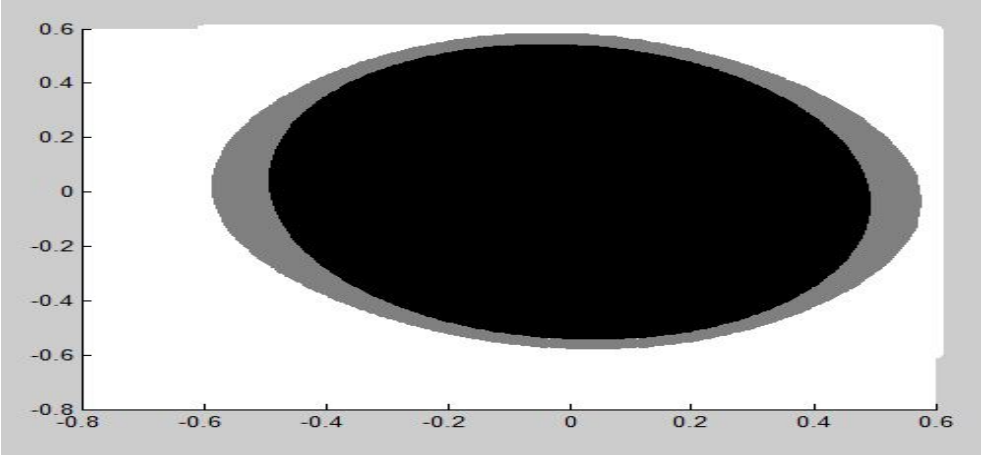


Figure 5.11: The collection of all $\bar{y}$ and $\hat{y}$ when $\beta = \frac{2}{3}$

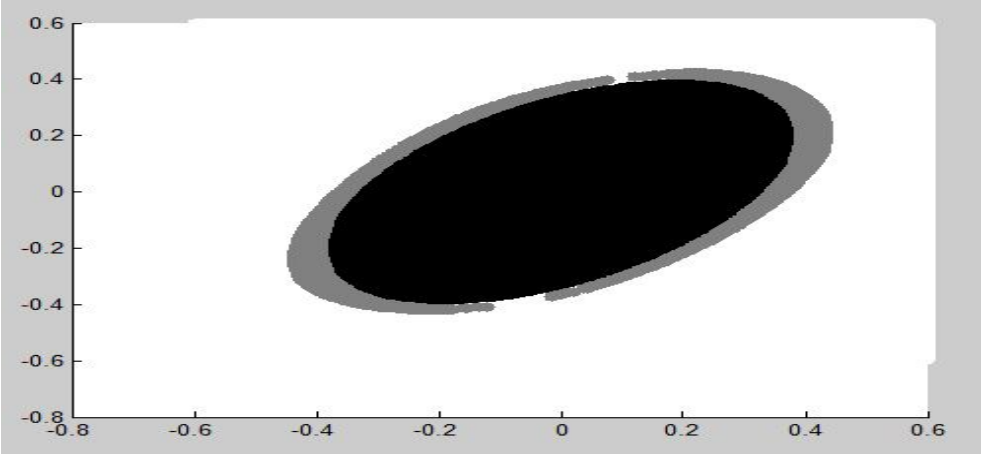


Figure 5.12: The collection of all $\bar{y}$ and $\hat{y}$ when $\beta = \frac{2}{3}$

From the above figures, we can see that, in this experiment, there is a small region of points y such that it is in $\hat{y}$ but not $\bar{y}$. Compared to the area of the intersection of all $\bar{y}$ and $\hat{y}$, the area of the symmetric difference of $\bar{y}$ and $\hat{y}$ is relatively small and it increases as β increases.

Now, we want to consider

$$\left|\left|\Pi_{\text{Null}(AS^{-1})}\left(\frac{Sx_\mu}{\mu} - e\right)\right|\right|_2^2 = \left|\left|\left(\frac{Sx_\mu}{\mu} - e\right) - \Pi_{R((AS^{-1})^\top)}\left(\frac{Sx_\mu}{\mu} - e\right)\right|\right|_2^2.$$

Note that

$$\begin{aligned}\left\|\left(\frac{Sx_\mu}{\mu}-e\right)\right\|_2^2 &= \frac{1}{\mu^2}x_\mu^\top S^2 x_\mu - \frac{2}{\mu}\langle s, x_\mu\rangle + n \\ &= e^\top S^2 S_\mu^{-2} e - 2\langle s, S_\mu^{-1} e\rangle + n\end{aligned}$$

If S is very close to S_μ, equivalently y is very close to y_μ, then we have

$$\left\|\left(\frac{Sx_\mu}{\mu}-e\right)\right\|_2^2 \approx 0.$$

In this case, it is impossible that $\left\|\Pi_{(R(AS^{-1})^\top)}\left(\frac{Sx_\mu}{\mu}-e\right)\right\|_2^2$ is small, but $\left\|\Pi_{\mathrm{Null}(AS^{-1})}\left(\frac{Sx_\mu}{\mu}-e\right)\right\|_2^2$ is large.

5.5 Analysis of the prediction direction via a primal-dual approach

For iterates $y^{(k)}$ very close to the central path, we have that

$$1-\tilde{\beta} \leq \frac{(x_\mu)_j s_j}{\mu} \leq 1+\tilde{\beta}, \ \ \forall j,$$

in a small neighbourhood of primal-dual central path for some small $\tilde{\beta} > 0$.

For $j \in B$, $(x_\mu)_j \to x_j^{(a)} > 0$, $\forall j \in B$, as $\mu \to 0^+$, and $x_j^{(a)} = \Omega(1), \ \forall j \in B$. Since we have

$$\frac{(x_\mu)_j s_j}{\mu} \leq 1+\tilde{\beta}, \ \ \forall j \in B,$$

we must also have

$$s_j \leq (1+\tilde{\beta})\Omega(1)\mu = O(\mu), \ \ \forall j \in B,$$

Therefore,

$$s_j = O(\mu), \ \ \forall j \in B. \tag{5.5}$$

The next three lemmas use the proof techniques of Ye, Güler, Tapia, Zhang [50] and Mehrotra [26].

Lemma 5.5.1. *Let d_y be the predictor direction for the y-component of a dual iterate (y, s). Then,*

$$d_s = -S\Pi_{R(S^{-1}A^\top)}e.$$

Moreover, we have,

$$||(d_s)_B|| = O(\mu),$$

Proof. Since d_y is the unique solution of

$$\nabla^2 f(y) d_y = \nabla f(y),$$

We have,

$$d_y = \left[\nabla^2 f(y)\right]^{-1} \nabla f(y) = [AS^{-2}A^\top]^{-1}AS^{-1}e.$$

Hence,

$$\begin{aligned} ds &= -A^\top d_y \\ &= -A^\top [AS^{-2}A^\top]^{-1}AS^{-1}e \\ &= -S(S^{-1}A^\top)[(AS^{-1})(AS^{-1})^\top]^{-1}(AS^{-1})e \\ &= -S\Pi_{R(S^{-1}A^\top)}e. \end{aligned}$$

For the second part, consider

$$\left|\left|S^{-1}d_s\right|\right| \leq \left|\left|\Pi_{R(S^{-1}A^\top)}\right|\right| ||e|| = \sqrt{n}. \tag{5.6}$$

Then,

$$\begin{aligned} ||(d_s)_B|| &= \left|\left|S_B S_B^{-1}(d_s)_B\right|\right| \\ &= \left|\left|S_B (S^{-1}d_s)_B\right|\right| \\ &\leq ||S_B|| \left|\left|(S^{-1}d_s)_B\right|\right| \\ &\leq ||S_B|| \left|\left|S^{-1}d_s\right|\right| \\ &\leq ||S_B|| \cdot \sqrt{n} \\ &= O(\mu), \end{aligned}$$

where the fifth line uses (5.6) and the last line uses (5.5). □

Now, let us bound the norm of $(d_s)_N$.

Lemma 5.5.2. *Let d_y and d_s be the predictor directions in y-space and s-space respectively. Then, when the current iterate is on the central path, $\bar{v} := d_y$ is a solution to the following least-squares problem*

$$\begin{aligned} \min_v \quad & \tfrac{1}{2} \left|\left| S_N^{-1} A_N^\top v \right|\right|^2 \\ s.t. \quad & A_B^\top v = -(d_s)_B. \end{aligned}$$

Proof. First, we check that $\bar{v}$ satisfies the constraint.

$$\begin{aligned} A_B^\top \bar{v} + (d_s)_B &= A_B^\top d_y + (d_s)_B \\ &= A_B^\top d_y + A_N^\top d_y + (d_s)_B - A_N^\top d_y \\ &= A_B^\top d_y + A_N^\top d_y + (d_s)_B + (d_S)_N \\ &= A^\top d_y + d_s \\ &= 0. \end{aligned}$$

Next, notice that

$$\begin{aligned} A_B S_B^{-2} (d_s)_B - A_N S_N^{-2} A_N^\top \bar{v} &= A_B S_B^{-2} (d_s)_B + A_N S_N^{-2} (d_s)_N \\ &= A S^{-2} d_s \\ &= -\underbrace{(AS^{-1})(S^{-1}A^\top)[(AS^{-1})(AS^{-1})^\top]^{-1}}_{=I} (AS^{-1})e \\ &= -AS^{-1}e. \end{aligned}$$

When the iterate is on the central path,

$$\begin{aligned} & [\mathrm{Diag}\,(S(\mu))]\, x(\mu) = \mu e, \\ \iff & [S(\mu)] = \frac{1}{\mu} x(\mu), \\ \Rightarrow \quad & A\,[S(\mu)] = \frac{1}{\mu} A x(\mu) = \frac{1}{\mu} b. \end{aligned} \tag{5.7}$$

Note that $x_N^* = 0$ for all optimal x^*, so we must have $A_B x_B^* = b$. Therefore, we have $b \in R(A_B)$. Then, $-AS^{-1}e \in R(A_B)$. Then, using (5.7), we deduce

$$A_N S_N^{-2} A_N^\top \bar{v} \in R(A_B).$$

Since $\bar{v}$ satisfies the constraints, we know that $\bar{v}$ satisfies the feasibility condition of the Karush-Kuhn-Tucker (KKT) conditions for the given problem. Moreover,

$$A_N S_N^{-2} A_N^\top \bar{v} \in R(A_B)$$

implies that

$$\exists \lambda \in \mathbb{R}^m \text{ such that } A_B\lambda = A_N S_N^{-2} A_N^\top \bar{v}.$$

Then, we have that

$$-\nabla\left(g(v)\right) = \lambda \nabla\left(h(v)\right),$$

where $g(v) := \frac{1}{2}\left|\left|S_N^{-1}A_N^\top v\right|\right|^2$ and $h(v) := A_B^\top v + (d_s)_B$. This gives the stationary conditions of the Karush-Kuhn-Tucker (KKT) conditions. Hence, $\bar{v}$ satisfies the KKT conditions of the given problem. □

Lemma 5.5.3. *Let d_s be the predictor direction in the s-space, then*

$$||(d_s)_N|| = O(\mu).$$

Proof. Because the least-squares problem defined in Lemma 5.5.2 is always feasible, there must be a feasible v such that

$$A_B^\top v = -(d_s)_B.$$

$|B|$ may not equal m, so A_B is not necessary a square matrix. Now, we consider the augmented system $\left[A_B^\top \;\; | -(d_s)_B\right]$.

Applying a row permutation to the augmented matrix to bring a maximal linearly independent set of rows of $A_B^\top$ to the first $\text{rank}(A_B^\top)$ rows. Denote the indices of those first $\text{rank}(A_B^\top)$ rows by $\bar{B}$. If necessary, we then do a column permutation so that $A_B^\top$ is r-by-r, where $rank(A_B^\top) = r$. Then, we apply elementary row operations to the submatrix of the augmented system which consisting row $r+1$ to row $|B|$ so that this submatrix is 0. We can always do it since the augmented system is feasible.

The found augmented matrix may have the empty matrix in any of the blocks: $(1,2),(2,1),(2,2),(2,3)$. The found augmented matrix should look like:

$$\left[\begin{array}{c|c|c} A_{\bar{B}}^\top & * & -(d_s)_{\bar{B}} \\ 0 & 0 & 0 \end{array}\right].$$

Hence, we can derive the subvector $(v)_{\bar{B}}$ of v as $(v)_{\bar{B}} = -A_{\bar{B}}^{-\top}(d_s)_{\bar{B}}$.
Therefore,

$$||v|| = O(||(d_s)_B||),$$

By Lemma 5.5.1, we know that

$$||v|| = O(||(d_s)_B||) = O(\mu).$$

Furthermore, by Lemma 5.5.1 and Lemma 5.5.2,

$$\begin{aligned}
||(d_s)_N|| &= \left|\left|S_N S_N^{-1}(d_s)_N\right|\right| \\
&\leq ||S_N|| \left|\left|S_N^{-1}(d_s)_N\right|\right| \\
&= ||S_N|| \left|\left|S_N^{-1} A_N^\top d_y\right|\right| \\
&\leq ||S_N|| \left|\left|S_N^{-1} A_N^\top v\right|\right|, \quad \text{if } (y,s) \text{ is on the central path} \\
&\leq ||S_N|| \left|\left|S_N^{-1}\right|\right| \left|\left|A_N^\top\right|\right| ||v|| \\
&= O(||v||) \\
&= O(\mu).
\end{aligned}$$

Let $\bar{v}$ be defined as a minimizer of the problem defined in Lemma 5.5.2 and $\bar{s}$ be the corresponding vector in s-space. If the current (y,s) is not on the central path, then by the statement of the algorithm, we know that

$$\left|\left|\nabla f(y) - \frac{1}{\mu} b\right|\right|_y = \left|\left|AS^{-1}e - \frac{1}{\mu} b\right|\right|_y = \left|\left|AS^{-1}e - A\bar{S}^{-1}e\right|\right|_y \leq \frac{1}{25}.$$

Note that

$$\begin{aligned}
\left|\left|S_N^{-1} A_N^\top d_y\right|\right|^2 &= \left|\left|S_N^{-1}(d_s)_N\right|\right|^2 \\
&= \left|\left|S_N^{-1} S_N (S_N^{-1} A_N^\top)[(A_N S_N^{-1})(A_N S_N^{-1})^\top]^{-1}(A_N S_N^{-1})e\right|\right|^2 \\
&= \left|\left|\Pi_{R\left((AS^{-1})_N^\top\right)} e\right|\right|^2 \\
&= e^\top \left(\Pi_{R\left((AS^{-1})_N^\top\right)}\right)^\top \Pi_{R\left((AS^{-1})_N^\top\right)} e \\
&= e^\top \Pi_{R\left((AS^{-1})_N^\top\right)} e, \quad \text{since } \Pi_{R\left((AS^{-1})_N^\top\right)} \text{ is an orthogonal projection.}
\end{aligned}$$

Therefore,

$$\begin{aligned}
& \left|\left|S_N^{-1}A_N^\top d_y\right|\right|^2 - \left|\left|\bar{S}_N^{-1}A_N^\top \bar{v}\right|\right|^2 \\
= & \left|\left|\Pi_{R\left((AS^{-1})_N^\top\right)}e\right|\right|^2 - \left|\left|\Pi_{R\left(\left(A\bar{S}^{-1}\right)_N^\top\right)}e\right|\right|^2 \\
= & \left(\left|\left|\Pi_{R\left((AS^{-1})_N^\top\right)}e\right|\right| + \left|\left|\Pi_{R\left(\left(A\bar{S}^{-1}\right)_N^\top\right)}e\right|\right|\right)\left(\left|\left|\Pi_{R\left((AS^{-1})_N^\top\right)}e\right|\right| - \left|\left|\Pi_{R\left(\left(A\bar{S}^{-1}\right)_N^\top\right)}e\right|\right|\right) \\
\leq & \left(||e|| + ||e||\right)\left(\left|\left|\Pi_{R\left((AS^{-1})_N^\top\right)}e\right|\right| - \left|\left|\Pi_{R\left(\left(A\bar{S}^{-1}\right)_N^\top\right)}e\right|\right|\right) \\
\leq & 2\sqrt{n}\left(\left|\left|\Pi_{R\left((AS^{-1})_N^\top\right)}e\right|\right| - \left|\left|\Pi_{R\left(\left(A\bar{S}^{-1}\right)_N^\top\right)}e\right|\right|\right).
\end{aligned}$$

Hence, when the current iterate (y, s) is very close to the central path, $\left|\left|S_N^{-1}A_N^\top d_y\right|\right|^2 - \left|\left|\bar{S}_N^{-1}A_N^\top \bar{v}\right|\right|^2$ is small. Then, we can still obtain

$$\begin{aligned}
||S_N||\left|\left|S_N^{-1}A_N^\top d_y\right|\right| &\leq ||S_N||\left|\left|S_N^{-1}A_N^\top v\right|\right| \\
&\leq ||S_N||\left|\left|S_N^{-1}\right|\right|\left|\left|A_N^\top\right|\right||v|| \\
&= O(||v||) \\
&= O(\mu).
\end{aligned}$$

□

From above, we showed that $||d_s|| = O(\mu)$. In addition, we need the convergence property of the corrector steps. From Section 8 in [30], we can see that after applying predictor step, we need to apply a corrector step such that the next iterate $y^{(k+1)} \in \mathcal{N}(\mu_{k+1}, \frac{1}{25})$. Therefore, we minimize the following function using Newton Method:

$$g(y) := f(y) - \frac{1}{\mu_{k+1}}b^\top y.$$

The Newton direction which minimizes the above function g is actually the unique solution d to the following system: (The first corrector direction in the Appendix A)

$$\nabla^2 f(y^{(k)})d = -\left[\nabla f(y^{(k)}) - \frac{1}{\mu_{k+1}}b\right].$$

Nesterov and Tunçel [30] showed that this corrector direction achieves quadratic convergence due to the properties of self-concordant barrier function and Newton decrement being small enough. So, this result requires a proximity (to the central path) hypothesis.

Chapter 6

Computational Experiments

In this chapter, we computationally observe and test the superlinear convergence properties of a variant of the dual algorithm of [30]. We restate the algorithm with more details. All of the following experiments are performed using this version of the algorithm.

Algorithm 3

1. Input: $\bar{\mu} \in (0,1)$, a pair of positive integers m, n such that $n > m \geq 1$.

2. Pick a randomly generated matrix A using a uniform distribution $rand(m,n)$ in MATLAB.

3. Let $b := Ae$, $c := e$, $y^{(0)} := 0$, $s^{(0)} := c - A^\top y^{(0)} = e$, $\mu_0 := 1$ and $k := 0$.

4. If $\mu_k \leq \bar{\mu}$, then output $y^{(k)}$, μ_k and **stop**. Otherwise, execute the following loop:

   (a) Compute the predictor direction $d_{y^{(k)}} = \left[AS^{-2}A^\top\right]^{-1} AS^{-1}e$ and compute $\bar{\alpha}_k = \bar{\alpha}(y^{(k)}) = \max\left\{\alpha \geq 0 : A^\top\left(y^{(k)} + \alpha d_{y^{(k)}}\right) \leq c\right\}$.

   (b) Use bi-section method to find the largest α such that $y(\alpha) = y^{(k)} + \alpha d_{y^{(k)}} \in \mathcal{N}(\frac{\mu_k}{\xi_{\bar{\alpha_k}}(\alpha_k)}, \frac{1}{6})$, where $\xi_{\bar{\alpha}}(\alpha) := 1 + \frac{\alpha\bar{\alpha}}{\bar{\alpha}-\alpha}$, $\alpha \in [0, \bar{\alpha})$.

   (c) Set $p^{(k)} = y^{(k)} + \alpha_k d_{y^{(k)}}$, and update $\mu_{k+1} = \frac{\mu_k}{\xi_{\bar{\alpha_k}}(\alpha_k)}$.

   (d) Apply one corrector step to $p^{(k)}$ to obtain $y^{(k+1)} \in \mathcal{N}(\mu_{k+1}, \frac{1}{25})$, $y^{(k+1)} := p^{(k)} + \Delta_y^{(k)}$, where $\Delta_y^{(k)} := \frac{1}{\mu}\left[f''(y^{(k)})\right]^{-1} b - \left[f''(y^{(k)})\right]^{-1} f'(y^{(k)})$.

   (e) Update $k = k + 1$, and go back to Step 4.

Note that $f'(y^{(0)}) = AIe = b$. We set $\mu_0 := 1$ and then $y^{(0)} \in \mathcal{N}(\mu_0, \beta)$, where $\beta = \frac{1}{25}$.

The experiments of this chapter are performed by using the software MATLAB R2013a, on a Intel(R) Core(TM) i7-4770 CPU @3.40GHz with 12GB of memory. In the test examples, our data: A, b are randomly generated by the command $rand(m,n)$ in MATLAB.

In order to present the information better, when $\mu_k \leq \bar{\mu}$ and the algorithm is about to stop, we can compute a "final pair" of primal-dual solutions $(\hat{x}.\hat{s})$ as follows:

- Let $\hat{s} := s^{(k)} = c - A^\top y^{(k)}$;
- Find the smallest $\lambda \geq 0$ such that $A\left(-\lambda F_*'(\hat{s})\right) \cong b$. That is, we want to find $\bar{\lambda} \geq 0$ which minimizes $||b + \lambda A\left(F_*'(\hat{s})\right)||_2^2$. Since $||b + \lambda A\left(F_*'(\hat{s})\right)||_2^2$ is a quadratic function in λ, the minimizer λ satisfies the following equation.

$$2\bar{\lambda}(h^\top h) = -2b^\top h$$

  where $h := A\left(F_*'(\hat{s})\right)$. Hence, the minimizer $\bar{\lambda} \geq 0$ satisfies $\bar{\lambda} = \max\left\{0, -\frac{b^\top h}{h^\top h}\right\}$.

- Define $\hat{x} := -\bar{\lambda} F'_*(\hat{s})$;
- Compute μ of $(\hat{x}.\hat{s})$: $\mu := \frac{\langle \hat{x}.\hat{s} \rangle}{n}$ and report $||A\hat{x} - b||^2$.

We performed the experiment 100 times where $m = 100$ and $n = 400$ using the Algorithm 3 with termination criterion: $\mu \leq \bar{\mu}$ and 300 as the maximum number of corrector steps allowed per iteration. Even though the theory guarantees, if all computations are carried out in exact arithmetic, one corrector step suffices in each iteration. In computational experiments as μ goes below 10^{-9}, due to numerical inaccuracies, the algorithm may require very many corrector steps. We collected the results on the different values of μ when the algorithm terminates under various stopping criterion $\bar{\mu}$. All results are presented in the following table.

Final value of μ	$\bar{\mu} - \frac{\bar{\mu}}{10}$	$\frac{\bar{\mu}}{10} - \frac{\bar{\mu}}{10^3}$	$\frac{\bar{\mu}}{10^3} - \frac{\bar{\mu}}{10^5}$	$\frac{\bar{\mu}}{10^5} - \frac{\bar{\mu}}{10^7}$	$< \frac{\bar{\mu}}{10^7}$	max num of corrector steps per iteration per instance on avg over 100 instances
$\bar{\mu} = 10^{-8}$	68	32	0	0	0	1
$\bar{\mu} = 10^{-9}$	46	52	2	0	0	1
$\bar{\mu} = 10^{-10}$	40	48	12	0	0	59.81
$\bar{\mu} = 10^{-12}$	21	79	0	0	0	259.72

Table 6.1: Distribution of final values of μ when $m = 100$ and $n = 400$

The following table shows the result of the experiment when $m = 100$ and $n = 800$ using the above algorithm with different termination criterion choices of $\bar{\mu}$ and 300 as the maximum number of corrector steps allowed.

Final value of μ	$\bar{\mu} - \frac{\bar{\mu}}{10}$	$\frac{\bar{\mu}}{10} - \frac{\bar{\mu}}{10^3}$	$\frac{\bar{\mu}}{10^3} - \frac{\bar{\mu}}{10^5}$	$\frac{\bar{\mu}}{10^5} - \frac{\bar{\mu}}{10^7}$	$< \frac{\bar{\mu}}{10^7}$	max num of corrector steps per iteration per instance on avg over 100 instances
$\bar{\mu} = 10^{-8}$	63	37	0	0	0	1
$\bar{\mu} = 10^{-9}$	52	47	1	0	0	1
$\bar{\mu} = 10^{-10}$	28	72	0	0	0	64.85
$\bar{\mu} = 10^{-12}$	91	9	0	0	0	248.34

Table 6.2: Distribution of final values of μ when $m = 100$ and $n = 800$

The following table shows the result of the experiment when $m = 200$ and $n = 400$ using the above algorithm with different termination criterion choices of $\bar{\mu}$ and 300 as the maximum number of corrector steps allowed.

Final value of μ	$\bar{\mu} - \frac{\bar{\mu}}{10}$	$\frac{\bar{\mu}}{10} - \frac{\bar{\mu}}{10^3}$	$\frac{\bar{\mu}}{10^3} - \frac{\bar{\mu}}{10^5}$	$\frac{\bar{\mu}}{10^5} - \frac{\bar{\mu}}{10^7}$	$< \frac{\bar{\mu}}{10^7}$	max num of corrector steps per iteration per instance on avg over 100 instances
$\bar{\mu} = 10^{-8}$	88	12	0	0	0	1
$\bar{\mu} = 10^{-9}$	69	31	0	0	0	1
$\bar{\mu} = 10^{-10}$	48	46	6	0	0	27.96
$\bar{\mu} = 10^{-12}$	34	66	0	0	0	236.65

Table 6.3: Distribution of final values of μ when $m = 200$ and $n = 400$

The following table shows the result of the experiment when $m = 200$ and $n = 800$ using the above algorithm with different termination criterion choices of $\bar{\mu}$ and 300 as the maximum number of corrector steps allowed.

Final value of μ	$\bar{\mu} - \frac{\bar{\mu}}{10}$	$\frac{\bar{\mu}}{10} - \frac{\bar{\mu}}{10^3}$	$\frac{\bar{\mu}}{10^3} - \frac{\bar{\mu}}{10^5}$	$\frac{\bar{\mu}}{10^5} - \frac{\bar{\mu}}{10^7}$	$< \frac{\bar{\mu}}{10^7}$	max num of corrector steps per iteration per instance on avg over 100 instances
$\bar{\mu} = 10^{-8}$	89	11	0	0	0	1
$\bar{\mu} = 10^{-9}$	62	38	0	0	0	1
$\bar{\mu} = 10^{-10}$	48	52	0	0	0	31.25
$\bar{\mu} = 10^{-12}$	93	7	0	0	0	257.28

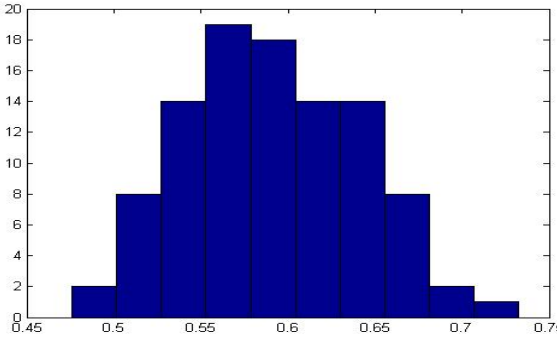
Table 6.4: Distribution of final values of μ when $m = 200$ and $n = 800$

In addition, the information on μ_k and α_k for the last five iterations for four arbitrary instances is displayed in the following table. Note that in these instances, we use 100 as the maximum number of corrector steps per iteration allowed.

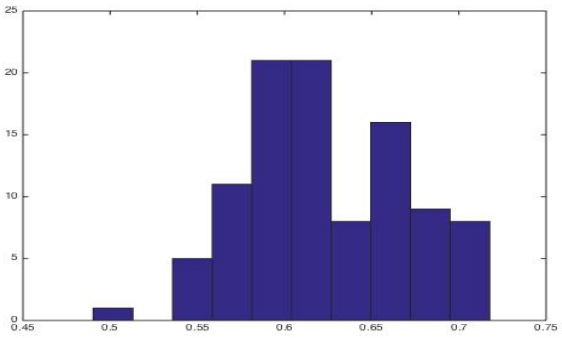
m	n	μ stopping criterion	max num of corrector steps per iteration used	μ_k	α_k
100	400	10^{-12}	1	4.80474514595144e-07	0.551063635397001
				1.4592004581876e-07	0.691190889629855
				2.05359061879439e-08	0.856899609035453
				5.23766329266451e-10	0.974066312468676
				3.67638556580905e-13	0.999286584241041
100	400	10^{-12}	100	5.60491798541484e-07	0.58558450345391
				1.42213909306164e-07	0.742373160901109
				1.32200050759574e-08	0.905608052820261
				1.35122781623248e-10	0.989621151297221
				5.47919615567409e-14	0.999592858640363
100	400	10^{-12}	1	5.99381347417163e-08	0.522369353382105
				2.04737980006565e-08	0.652600815435237
				3.70938336671908e-09	0.815737575328919
				1.66111013852665e-10	0.954456109146829
				3.75776646145919e-13	0.997700516973209
100	400	10^{-12}	100	3.13782051139735e-08	0.58389829908813
				8.27899583254401e-09	0.731669774620126
				8.61055450452595e-10	0.894227408227521
				1.13412415898217e-11	0.986604966620684
				6.55774716921468e-14	0.994214798778116

Table 6.5: Information of μ_k and α_k for the last five iterations

From the algorithm, we know that the value of α_k stands for the largest step size to keep the iterate in the larger neighbourhood. We noticed that the values of α_k seem quite random for several iterations until the algorithm is at the last few iterations, and there is no certain trend in the distribution of even the fifth last α_k over 100 instances. Therefore, we present the following figures which show us the distributions of the fifth last α_k over 100 instances for each different size of the problem.

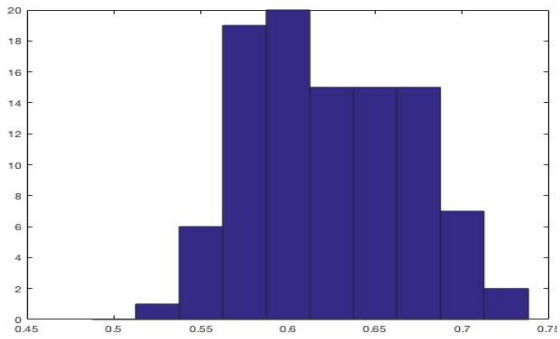


(a) The fifth last α_k where $m = 100$ and $n = 400$.

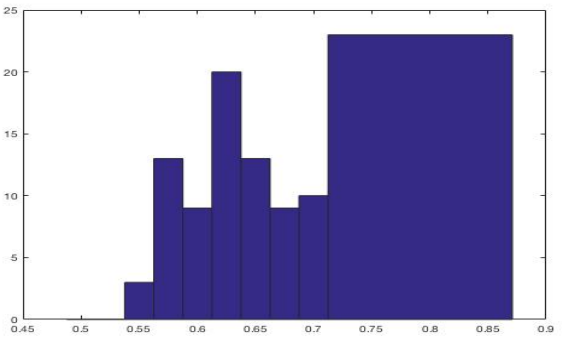


(b) The fifth last α_k where $m = 100$ and $n = 5000$

Figure 6.1: Histogram on the distributions of the fifth last α_k



(a) The fifth last α_k where $m = 100$ and $n = 10000$



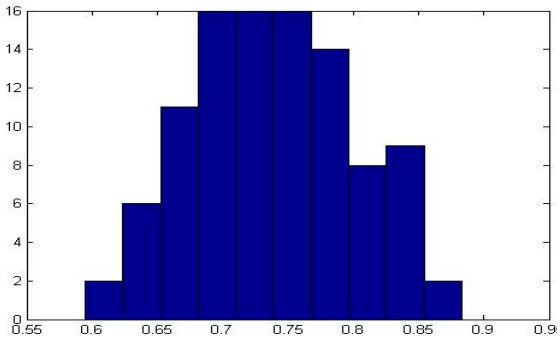
(b) The fifth last α_k where $m = 100$ and $n = 20000$

Figure 6.2: Histogram on the distributions of the fifth last α_k

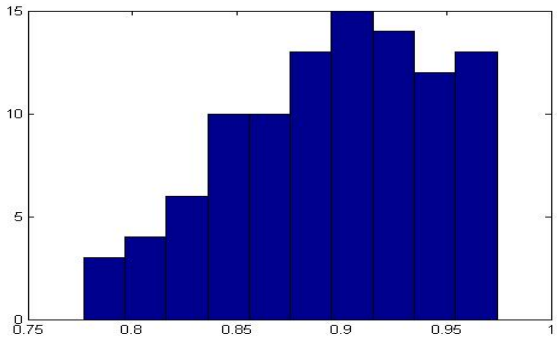
These four figures show us the distribution of the fifth last α over 100 randomly generated instances in four different cases. We can see that the distributions behave almost like a normal distribution with a little right-skewness despite the changes in the size of the problem. Next, we will present the distributions of the fourth last α_k, the third last α_k, the second last α_k and the last α_k. In each case of different size of the problem, instead of observing normal distributions, we are able to see an obvious trend in the values of α_k as

the algorithm is going to stop from those distributions. Based on these observations, we can conclude that α_k is approaching 1 as the algorithm is approaching to terminate.

The following four figures show the results of α_k for the last four iterations over 100 instances where m=100, n=400 and $\bar{\mu} = 10^{-12}$.

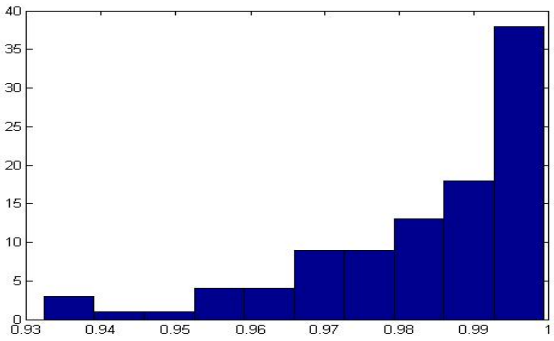


(a) The distribution of the fourth last α_k

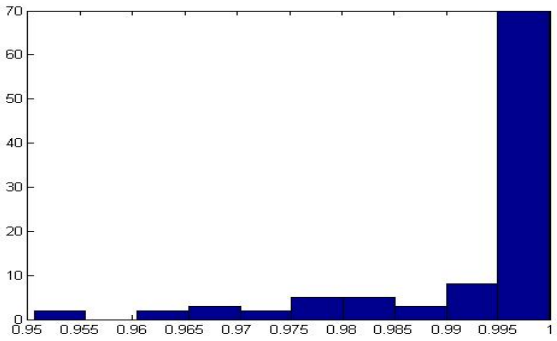


(b) The distribution of the third last α_k

Figure 6.3: Histogram on the distribution of the fourth and the third last α_k



(a) The distribution of the second last α_k



(b) The distribution of the last α_k

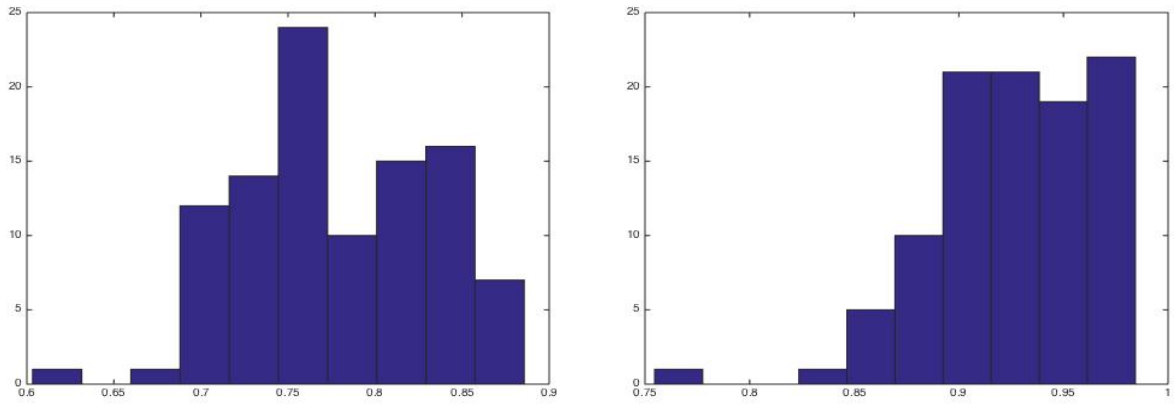
Figure 6.4: Histogram on the distribution of the second last and the last α_k

From the above figures, we observe that the distribution of the fourth last one behaves like a normal distribution and it is not skewed, i.e., the mean is approximately equal to the median. As the algorithm goes to the next iteration, we can see that the distribution

of the third last one is lefty-skewed. For the distributions of the second last and the last ones, they are skewed more to the left. Therefore, these figures show us that the value of α_k is approaching 1 when the algorithm terminates.

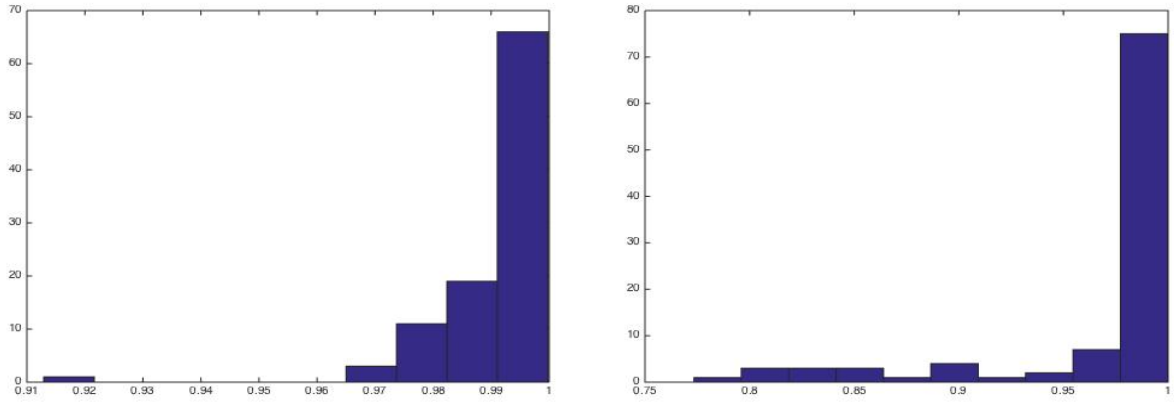
In addition, we can generate A such that it is a sparse matrix with m rows and n columns, and we require that every column of A to have several nonzero entries.

The following instance is under the scenario where, there are 5 nonzero entries for every column, $m = 100$, $n = 5000$, and $\bar{\mu} = 10^{-12}$, and then we plot the following histograms based on the distributions of α_k for the last 4 iterations.



(a) The distribution of the fourth last α_k (b) The distribution of the third last α_k

Figure 6.5: Histogram on the distribution of the fourth and the third last α_k for a sparse instance where $m = 100$ and $n = 5000$

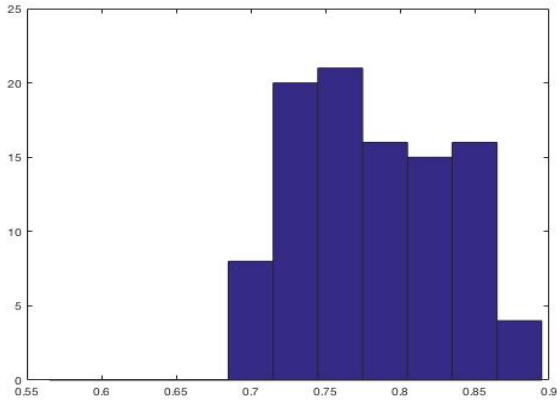


(a) The distribution of the second last α_k (b) The distribution of the last α_k

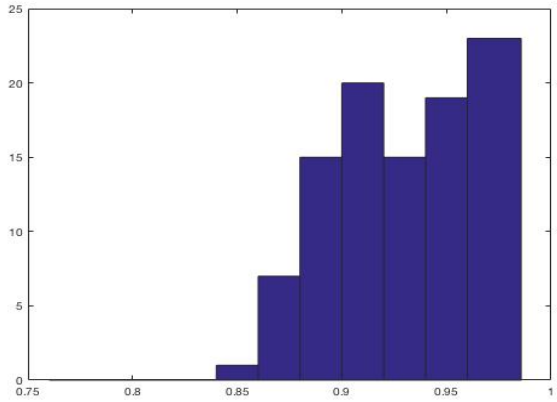
Figure 6.6: Histogram on the distribution of the second last and the last α_k for a sparse instance where $m = 100$ and $n = 5000$

The above four figures are similar to the ones of the case when $m = 100$ and $n = 400$. The distribution of the fourth one is not skewed. Then, the distributions of the third, the second last and the last ones are skewed to the left, and the distribution of the last α_k is skewed to the largest degree.

The following instance is under the scenario where $m = 100$, $n = 10000$, there are 4 nonzero entries for every column, and $\bar{\mu} = 10^{-12}$, and then we plot the following histograms based on the distributions of α_k for the last 4 iterations.

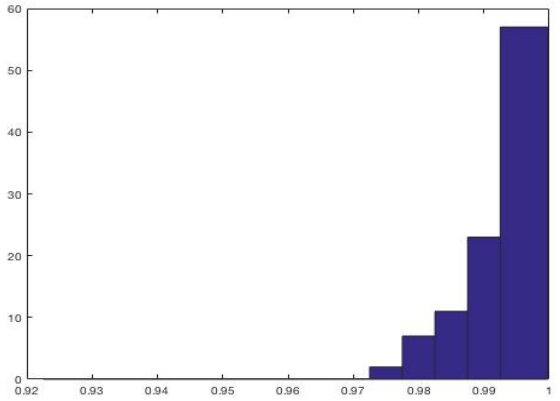


(a) The distribution of the fourth last α_k

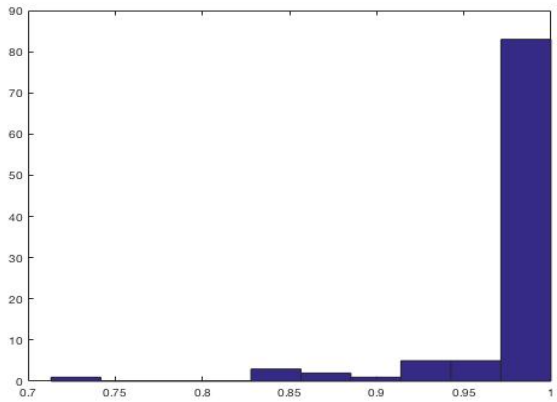


(b) The distribution of the third last α_k

Figure 6.7: Histogram on the distribution of the fourth and the third last α_k for a sparse instance where $m = 100$ and $n = 10000$



(a) The distribution of the second last α_k



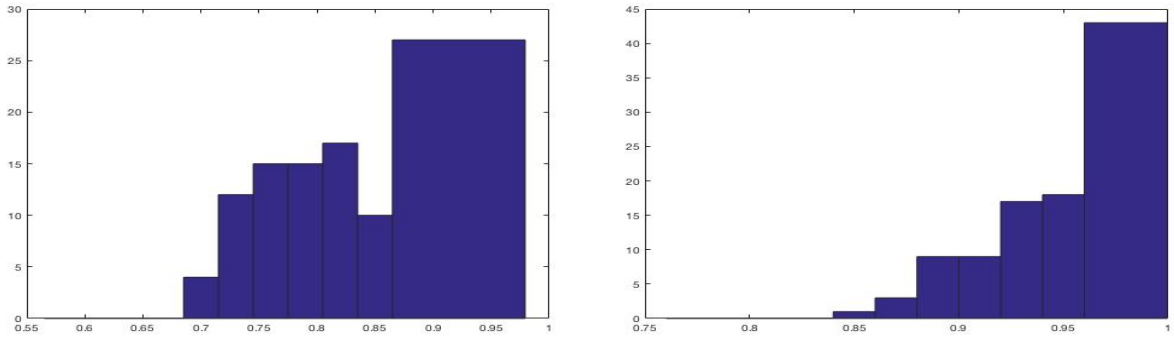
(b) The distribution of the last α_k before rescaling

Figure 6.8: Histogram on the distribution of the second last and the last α_k for a sparse instance where $m = 100$ and $n = 10000$

The above four figures are similar to the ones of the case when $m = 100$ and $n = 400$. The distribution of the fourth one is not too skewed. Then, the distributions of the third,

the second last and the last ones are skewed to the left, and the distribution of the last α_k is skewed to the largest degree.

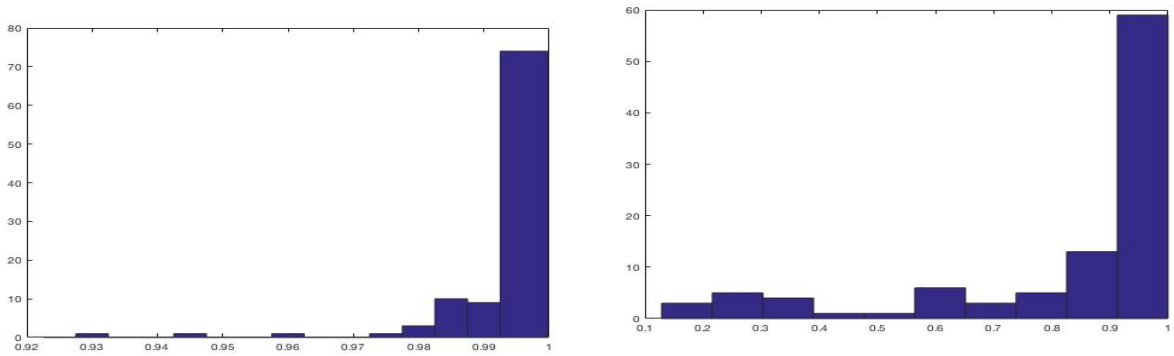
The following instance is under the scenario where $m = 100$, $n = 20000$, there are 5 nonzero entries for every column, and $\bar{\mu} = 10^{-12}$, and then we plot the following histograms based on the distributions of α_k for the last 4 iterations.



(a) The distribution of the fourth last α_k

(b) The distribution of the third last α_k

Figure 6.9: Histogram on the distribution of the fourth and the third last α_k for a sparse instance where $m = 100$ and $n = 20000$



(a) The distribution of the second last α_k

(b) The distribution of the last α_k before rescaling

Figure 6.10: Histogram on the distribution of the second and the last α_k for a sparse instance where $m = 100$ and $n = 20000$

The above four figures are similar to the ones of the case when $m = 100$ and $n = 400$. However, the distribution of the fourth one is a little left-skewed. Moreover, the distributions of the third, the second last and the last ones are skewed more to the left.

From the observation on the distributions of α_k for the last five iterations, we can see that there is a relationship between the convergence of α_k to 1 and the termination of the algorithm, i.e., the convergence of the duality gap to 0. Now, we will make some statements on the relation between the convergence of α_k to 1 and the convergence of μ_k to 0 and prove the correspondence. We will use the following figure to demonstrate the relationship. Note that to have such discrete separation between intervals (as in the figure), one has to fix constants and rates of convergence.

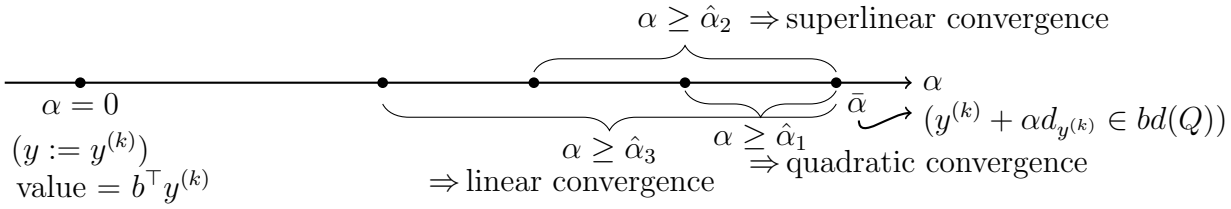


Figure 6.11: The schematic relationship between the value of α and the duality gap $f^* - \langle b, y \rangle$

- If
$$\alpha_k \geq \hat{\alpha}_1 := \frac{1}{1 + \kappa \left[f^* - b^\top y^{(k)} \right]},$$
for $\kappa := \frac{\kappa_2}{r}, r \in (0, 1)$,then it implies that $\mu_k \to 0$ Q-quadratically.

- If
$$\alpha_k \geq \hat{\alpha}_2 := \frac{1}{1 + \kappa \left[f^* - b^\top y^{(k)} \right]^r},$$
for some $r \in (0, 1)$,then it implies that $\mu_k \to 0$ Q-superlinearly.

- If
$$\alpha_k \geq \hat{\alpha}_3 := \frac{1}{r + \kappa \left[f^* - b^\top y^{(k)} \right]},$$
for some $r > 1$, then it implies that $\mu_k \to 0$ Q-linearly.

We will use the following proposition to illustrate the reasoning behind the above statements.

Proposition 6.0.1. *Let* $\hat{\alpha}_1 := \frac{1}{1+\kappa(f^*-\langle b,y\rangle)}$ *for some* $\kappa > 0$, $\hat{\alpha}_4 := \frac{1}{1+\kappa(f^*-\langle b,y\rangle)^{\frac{1}{2}}}$ *for some* $\kappa > 0$, *and* $\hat{\alpha}_2 := \frac{1}{1+\kappa(f^*-\langle b,y\rangle)^r}$ *for some* $\kappa > 0$ *and* $r \in (0,1)$. *Assume that* $y(\hat{\alpha}_1)$ *is feasible for the dual problem* (D) *and* $f^* - \langle b, y(\hat{\alpha}_1)\rangle \leq \tilde{\kappa}(f^* - \langle b, y\rangle)^2$, *for some* $\tilde{\kappa} > 0$. *Moreover,*

1. *if* $\alpha \geq \hat{\alpha}_1$, *then* $(f^* - \langle b, y\rangle)$ *converges to 0 quadratically;*
2. *if* $\alpha \geq \hat{\alpha}_4$, *then* $(f^* - \langle b, y\rangle)$ *converges to 0 with order at least* $\frac{3}{2}$;
3. *if* $\alpha \geq \hat{\alpha}_2$, *then* $(f^* - \langle b, y\rangle)$ *converges to 0 superlinearly;*

Proof. We will first prove for the first case. By applying the similar argument, we can prove the other two cases.

Recall that $y(\alpha) = y + \alpha d$, where d represents the predictor step for y. Then, we know that

$$\begin{aligned} f^* - \langle b, y(\alpha)\rangle &= f^* - \langle b, y\rangle - \alpha\langle b, d\rangle \\ &\leq f^* - \langle b, y\rangle + \frac{\alpha}{\hat{\alpha}_1}\left[\tilde{\kappa}(f^* - \langle b, y\rangle)^2 - f^* + \langle b, y\rangle\right] \\ &= f^* - \langle b, y\rangle - \alpha\left[\frac{f^* - \langle b, y\rangle - \tilde{\kappa}(f^* - \langle b, y\rangle)^2}{\hat{\alpha}_1}\right] \\ &= (1-\alpha)(f^* - \langle b, y\rangle) + \alpha(\tilde{\kappa} - \kappa)\left[f^* - \langle b, y\rangle\right]^2 + \alpha\tilde{\kappa}\kappa\left[f^* - \langle b, y\rangle\right]^3 \end{aligned}$$

If we set $\alpha = \hat{\alpha}_1$, then

$$(1-\alpha)(f^* - \langle b, y\rangle) = \frac{\kappa}{1 + \kappa(f^* - \langle b, y\rangle)}(f^* - \langle b, y\rangle)^2.$$

Also note that if $\alpha \approx 1$, then $(1-\alpha)(f^* - \langle b, y\rangle) \approx 0$. Therefore, if $\alpha \geq \hat{\alpha}_1$, then $(f^* - \langle b, y\rangle)$ converges to 0 Q-quadratically. □

Therefore, we can conclude that if α_k is larger enough, then $(f^* - \langle b, y\rangle)$ converges to 0 quadratically, i.e., μ_k converges to 0 Q-quadratically. As shown in the figures of the distribution of the last five values of α, we see that α_k is converging very fast to 1. Then, it means that we can achieve Q-quadratic convergence of μ_k to 0.

LU decomposition and QR decomposition, instead of Cholesky decomposition are also been adopted. However, I did not observe a big difference in the results between these three decomposition methods.

m	n	$Methods$	$\bar{\mu}$	max num of corrector steps allowed	$\bar{\mu} - \frac{\bar{\mu}}{10}$	$\frac{\bar{\mu}}{10} - \frac{\bar{\mu}}{10^3}$	$\frac{\bar{\mu}}{10^3} - \frac{\bar{\mu}}{10^5}$	$\frac{\bar{\mu}}{10^5} - \frac{\bar{\mu}}{10^7}$	$< \frac{\bar{\mu}}{10^7}$
100	400	Cholesky	10^{-12}	300	25	75	0	0	0
		LU			26	74	0	0	0
		QR			25	75	0	0	0
100	400	Cholesky	10^{-12}	300	27	73	0	0	0
		LU			28	72	0	0	0
		QR			29	71	0	0	0
100	400	Cholesky	10^{-12}	100	21	79	0	0	0
		LU			21	79	0	0	0
		QR			23	77	0	0	0
100	400	Cholesky	10^{-12}	100	26	74	0	0	0
		LU			27	73	0	0	0
		QR			27	73	0	0	0

Table 6.6: Distribution of final values of μ using three different decomposition methods

Chapter 7

Superlinear Convergence in Semidefinite Programming, Conclusion and Future Research

After analyzing the behaviour of the algorithm in the special case of Linear Programming, let us investigate the behaviour of the algorithm in the case of Semidefinite Programming. Similar to Linear Programming, Semidefinite Programming has many nice and special properties compared to general convex optimization. Therefore, based on the algorithm proposed in [30], we should be able to approach superlinear convergence analysis that maintains polynomial iteration complexity bound and use weaker assumptions in the case of Semidefinite Programming than the general convex optimization case.

Superlinear and quadratic convergence properties of polynomial-time interior-point methods in the case of Linear Programming is interesting and useful. However, some may argue that that in case of Linear Programming, there are very effective finite termination algorithms (purification algorithms, projection based routines etc.) which generate exact optimal solutions if the current iterate is close enough to the optimal solution set (in addition to work based on Tapia indicators, see for instance Ye [49], Mehrotra and Ye [27], Vavasis and Ye [47]). So, in this sense, the utility of superlinear convergence results for LP, in practice seems a bit less critical. On the other hand, the case of Semidefinite Programming is more complicated than the case of Linear Programming in the sense that the cone of positive semidefinite matrices is not polyhedral while the cone of nonnegative vectors is a polyhedral cone. Regardless of how close an approximately optimal solution is to the optimal solution set, we do not yet have an efficient algorithm (in case of SDP) to

"jump" to an exact optimal solution. These issues make superlinear convergence properties more critical in SDP compared to LP. Moreover, Halická, de Klerk and Roos (in [16]) show that the central path does not converge to the analytic centre in general for the SDP case by providing counterexamples in SDP case and Second Order Cone case. It will make the analysis of the limiting behaviour of the central paths more complicated since it is shown that the central path always converges to the analytic centre of the optimal solution set in the case of Linear Programming.

We use the following arguments in [30] to illustrate the possibility of using weaker assumptions in the case of Semidefinite Programming.

For the cone of positive semidefinite matrices $K = K^* = \mathbb{S}^n_+$, we choose

$$F(X) = -\ln\det X, \quad F_*(S) = -n - \ln\det S.$$

Then,

$$\langle I, \nabla^2 F_*(S_\mu) S^* \rangle = \langle I, S_\mu^{-1} S^* S_\mu^{-1} \rangle.$$

It seems difficult to get an upper bound for this value in terms of $||S^*||^2_{S_\mu} = \langle S_\mu^{-1} S^* S_\mu^{-1}, S^* \rangle$. However, we can apply the following approach:

$$\langle I, S_\mu^{-1} S^* S_\mu^{-1} \rangle = \mu^{-2} \langle X_\mu^2, S^* \rangle = \mu^{-2} \langle (X_\mu - X^*)^2, S^* \rangle.$$

Therefore, we get an upper bound for $||\nabla^2 F_*(S_\mu) S^*||_T$ assuming $||X_\mu - X^*||_T \in O(\mu)$. This last condition has been used in superlinear convergence analyses in the literature for Semidefinite Programming case, as we noted in Chapter 4.

Based on the above discussion, decomposing the iterates S according the partition B and N would be a useful start.

Furthermore, notice that the Proposition 5.2.1 does not directly apply to Semidefinite Programming case. Since we need the fact that the feasible region of the dual problem (D) is a polyhedron in the proof of Proposition 5.2.1 and the cone of the positive semidefinite matrices is not polyhedral, we are not able to show that $\inf_{y \text{ is feasible}} \cos(\angle(y_* - y, b)) > 0$ in the Semidefinite Programming case. Hence, Proposition 5.2.1 does not directly apply to Semidefinite Programming case.

References

[1] Coin-or branch and cut. https://projects.coin-or.org/Cbc.

[2] Cplex optimizer. https://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/.

[3] Csdp. https://projects.coin-or.org/Csdp/.

[4] Dsdp. http://www.mcs.anl.gov/hs/software/DSDP/.

[5] Gnu linear programming kit. http://www.gnu.org/software/glpk/.

[6] Gurobi optimizer. http://www.gurobi.com/.

[7] Mosek. https://www.mosek.com/.

[8] Sdpa. http://sdpa.sourceforge.net/.

[9] Sdpt3. http://www.math.nus.edu.sg/~mattohkc/sdpt3.html.

[10] Sedumi. http://sedumi.ie.lehigh.edu/.

[11] J. X. da Cruz Neto, O. P. Ferreira, and R. D. C. Monteiro. Asymptotic behavior of the central path for a special class of degenerate SDP problems. *Math. Program.*, 103(3, Ser. A):487–514, 2005.

[12] G. B. Dantzig. *Linear Programming and Extensions*. Princeton landmarks in mathematics and physics. Princeton University Press, 1963.

[13] A. S. El-Bakry, R. A. Tapia, and Y. Zhang. A study of indicators for identifying zero variables in interior-point methods. *SIAM Rev.*, 36(1):45–72, 1994.

[14] M. A. Feldstein and R. M. Firestone. *Hermite Interpolatory Theory and Parallel Numerical Analysis*. Division in Applied Mathematics. Brown University, Providence, Rhode Island, 1967.

[15] D. Goldfarb and K. Scheinberg. Interior point trajectories in semidefinite programming. *SIAM Journal on Optimization*, 8(4):871–886, 1998.

[16] M. Halická, E. de Klerk, and C. Roos. On the convergence of the central path in semidefinite optimization. *SIAM Journal on Optimization*, 12(4):1090–1099, 2002.

[17] M. Halická, E. De Klerk, and C. Roos. Limiting behavior of the central path in semidefinite optimization. *Optim. Methods Softw.*, 20(1):99–113, 2005.

[18] A. J. Hoffman. On approximate solutions of systems of linear inequalities. In *Selected Papers Of Alan J Hoffman: With Commentary*, pages 174–176. 2003.

[19] M. Iri and H. Imai. A multiplicative barrier function method for linear programming. *Algorithmica*, 1(1-4):455–482, 1986.

[20] L. V. Kantorovich. Functional analysis and applied mathematics. *Uspekhi Matematicheskikh Nauk*, 3(6):89–185, 1948.

[21] L. V. Kantorovich and G. P. Akilov. Functional analysis in normed spaces, volume 46 of international series of monographs in pure and applied mathematics, 1964.

[22] N. Karmarkar. A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of Computing*, pages 302–311. ACM, 1984.

[23] L. G. Khachiyan. Polynomial algorithms in linear programming. *USSR Computational Mathematics and Mathematical Physics*, 20(1):53–72, 1980.

[24] M. Kojima and S. Shida, M.and Shindoh. Local convergence of predictor-corrector infeasible-interior-point algorithms for sdps and sdlcps. *Mathematical Programming*, 80(2):129–160, 1998.

[25] Z. 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(1):59–81, 1998.

[26] S. Mehrotra. Quadratic convergence in a primal-dual method. *Mathematics of Operations Research*, 18(3):741–751, 1993.

[27] S. Mehrotra and Y. Ye. Finding an interior point in the optimal face of linear programs. *Mathematical Programming*, 62(1-3):497–515, 1993.

[28] S. Mizuno, M. J. Todd, and Y. Ye. On adaptive-step primal-dual interior-point algorithms for linear programming. *Mathematics of Operations research*, 18(4):964–981, 1993.

[29] Y. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming: Theory and Applications*. SIAM, 1994.

[30] Y. Nesterov and L. Tunçel. Local superlinear convergence of polynomial-time interior-point methods for hyperbolicity cone optimization problems. *SIAM Journal on Optimization*, 26(1):139–170, 2016.

[31] J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonlinear equations in several variables*, volume 30 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000. Reprint of the 1970 original.

[32] F. A. Potra. On Q-order and R-order of convergence. *J. Optim. Theory Appl.*, 63(3):415–431, 1989.

[33] F. A. Potra and J. Stoer. On a class of superlinearly convergent polynomial time interior point methods for sufficient LCP. *SIAM Journal on Optimization*, 20(3):1333–1363, 2009.

[34] F.A. Potra. Q-superlinear convergence of the iterates in primal-dual interior-point methods. *Math. Program.*, 91(1, Ser. A):99–115, 2001.

[35] F.A. Potra. Primal-dual affine scaling interior point methods for linear complementarity problems. *SIAM Journal on Optimization*, 19(1):114–143, 2008.

[36] A. Schrijver. *Theory of Linear and Integer Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1986.

[37] S. Smale. Newton's method estimates from data at one point. In *The merging of disciplines: new directions in pure, applied, and computational mathematics (Laramie, Wyo., 1985)*, pages 185–196. Springer, New York, 1986.

[38] J. F. Sturm. Implementation of interior point methods for mixed semidefinite and second order cone optimization problems. *Optim. Methods Softw.*, 17(6):1105–1154, 2002.

[39] J. F. Sturm and S. Zhang. Symmetric primal-dual path-following algorithms for semidefinite programming. *Applied Numerical Mathematics*, 29(3):301–315, 1999.

[40] K. Tanabe. Centered newton method for mathematical programming. In *System modelling and optimization*, pages 197–206. Springer, 1988.

[41] R.A. Tapia. *Role of slack variables in quasi-Newton methods for constrained optimization*. Jan 1979.

[42] M. J. Todd. A study of search directions in primal-dual interior-point methods for semidefinite programming. *Optim. Methods Softw.*, 11/12(1-4):1–46, 1999. Interior point methods.

[43] M. J. Todd and Y. Ye. A centered projective algorithm for linear programming. *Mathematics of Operations Research*, 15(3):508–529, 1990.

[44] T. Tsuchiya. Quadratic convergence of the Iri-Imai algorithm for degenerate linear programming problems. *J. Optim. Theory Appl.*, 87(3):703–726, 1995.

[45] L. Tuncel. *Polyhedral and Semidefinite Programming Methods in Combinatorial Optimization*. Fields Institute monographs. American Mathematical Soc., 2010.

[46] R. H. Tütüncü. Quadratic convergence of potential-reduction methods for degenerate problems. *Math. Program.*, 90(1, Ser. A):169–203, 2001.

[47] S. A. Vavasis and Y. Ye. A primal-dual interior point method whose running time depends only on the constraint matrix. *Mathematical Programming*, 74(1):79–120, 1996.

[48] S. J. Wright. A path-following interior-point algorithm for linear and quadratic problems. *Annals of Operations Research*, 62(1):103–130, 1996.

[49] Y. Ye. On the finite convergence of interior-point algorithms for linear programming. *Mathematical Programming*, 57(1-3):325–335, 1992.

[50] Y. Ye, O. Güler, R. A. Tapia, and Y. Zhang. A quadratically convergent o($\sqrt{nl}$)-iteration algorithm for linear programming. *Mathematical programming*, 59(1-3):151–162, 1993.

APPENDICES

Appendix A

Different ways of computing corrector steps for dual path-following algorithms

In the Algorithm 1, we need to take one predictor step and one (or more) corrector step(s) in each iteration. For the predictor step, we use $d_y = [\nabla^2 f(y)]^{-1} \nabla f(y)$. However, for corrector step(s), we think of different ways of deriving it.

The purpose of taking corrector step(s) is to make the next iterate $y^{(k+1)}$ get back to the smaller neighbourhood. We can achieve this goal by minimizing the distance from the current iterate $y^{(k)}$ to the central path or considering the constrained problem $\left\{\min f(y) : b^\top y = b^\top y^{(k)}\right\}$ and finding the Newton's direction of this problem. Therefore, we could find different types of corrector steps through minimizing the neighbourhood parameter $\mathcal{N}(\mu, \beta)$ for the central path, applying Newton's method to the KKT system of the problem (P_μ), using the dual problem (D) and projecting the Newton's direction, eliminating one of the variables $y_1, y_2, \ldots, y_m$, or using optimality conditions of the constrained problem $\left\{\min f(y) : b^\top y = b^\top y^{(k)}\right\}$. Hence, we derive the following five corrector directions, which are pure dual step(s), according to five different approaches.

A.1 Minimizing the neighbourhood parameter

Consider the following optimization problem:

$$(\bar{P}_\mu) \quad \min \quad f(y) - \tfrac{1}{\mu} b^\top y,$$

for some μ fixed by the algorithm.

We can see that the hessian of the objective function $\nabla^2 f(y)$ satisfies $\nabla^2 f(y) \succ 0$. Hence, $(\bar{P}_\mu)$ is an unconstrained minimization problem whose objective function is strictly convex. Then, we know that $(\bar{P}_\mu)$ has a unique minimizer characterized by

$$f'(y) - \tfrac{1}{\mu} b = 0.$$

Therefore, we are able to obtain the Newton's direction for minimizing $(f(y) - \frac{1}{\mu} b^\top y)$ as follows:

$$\begin{aligned} & - \left[f''(y^{(k)})\right]^{-1} \left[f'(y^{(k)}) - \frac{1}{\mu} b\right] \\ = \quad & \frac{1}{\mu} \left[f''(y^{(k)})\right]^{-1} b - \left[f''(y^{(k)})\right]^{-1} f'(y^{(k)}). \end{aligned}$$

The resulting algorithm is quadratically convergent, provided that the initial y lies in $\mathcal{N}(\mu, \frac{1}{6})$. See Lemma 8.3 in [30].

A.2 Using the primal-dual symmetric system

Now, let us derive the corrector direction based on the primal-dual system and Newton's Method applied to the KKT system of the problem (P_μ) defined in 3.3. Applying Newton's Method, we obtain the following system:

$$\begin{aligned} Adx &= 0 && \text{(A.1)} \\ A^\top dy + ds &= 0 && \text{(A.2)} \\ Sdx + Xds &= -Xs + \mu e && \text{(A.3)} \end{aligned}$$

where X is the n-by-n diagonal matrix with diagonal entries x_i, for $i \in \{1, 2, \ldots, n\}$.

Equation (A.3) gives us:

$$dx \;=\; -x + \mu S^{-1}e - XS^{-1}ds.$$

Using (A.2) to eliminate ds and substitute into (1), we can get:

$$0 \;=\; Adx \;=\; -b + \mu AS^{-1}e + \left(AXS^{-1}A^{\top}\right) dy.$$

Algorithm 1 does not generate primal iterates $x^{(k)}$. However, the iterates stay close to the central path. Therefore, $x_\mu \approx \mu S^{-1}e$ and $X_\mu \approx \mu S^{-1}$. Thus, we replace X by μS^{-1} in $\left(AXS^{-1}A^{\top}\right)$.

Since $AXS^{-1}A^{\top}$ is positive definite, dy is the unique Newton's direction at (x, y, s) to compute the point on the central path corresponding to μ.

Then, we need to solve $\mu\left(AS^{-2}A^{\top}\right) dy = b - \mu AS^{-1}e$ for dy and let $y^{(k+1)} = y^{(k)} + dy$.

Therefore,

$$dy \;=\; \tfrac{1}{\mu}\left[f''(y^{(k)})\right]^{-1} b - \left[f''(y^{(k)})\right]^{-1} f'(y^{(k)}),$$

which leads to the same algorithm as in A.1, with the same properties.

In the next three sections, we consider some other approaches that a continuous optimizer might utilize to derive a computationally effective algorithm for the corrector steps.

A.3 Projected Newton's method

The corrector direction we obtained above is derived from the primal-dual system using Newton's Method. Moreover, we can only use the dual problem and Newton's Method to derive another corrector direction.

For $k \geq 0$, consider the following problem:

$$\begin{array}{lll} (P_c) \quad \min & f(y) & \\ & b^{\top}y & = \; b^{\top}y^{(k)}, \end{array}$$

where $y^{(k)}$ is the dual variable for the current iterate k and $b^{\top}y^{(k)}$ is the current objective value.

If the above problem is unconstrained, by using Newton's Method, we know that the Newton's direction is $-\left[f''(y^{(k)})\right]^{-1} f'(y^{(k)})$. Then for this constrained problem, in order

to find the Newton's direction, we may choose to project $-\left[f''(y^{(k)})\right]^{-1} f'(y^{(k)})$ onto the hyperplane $\left\{y \in \mathbb{R}^m : b^\top y = 0\right\}$. Then, we can calculate the projection matrix as follows:

$$\left[I - b\left(b^\top b\right)^{-1} b^\top\right] \;=\; \left[I - \frac{1}{||b||_2^2} bb^\top\right].$$

Therefore, the corresponding corrector direction is:

$$\begin{aligned} & -\left[I - \frac{1}{||b||_2^2} bb^\top\right] \left[f''(y^{(k)})\right]^{-1} f'(y^{(k)}) \\ = \;& \frac{1}{||b||_2^2} bb^\top \left[f''(y^{(k)})\right]^{-1} f'(y^{(k)}) - \left[f''(y^{(k)})\right]^{-1} f'(y^{(k)}) \\ = \;& \frac{b^\top \left[f''(y^{(k)})\right]^{-1} f'(y^{(k)})}{||b||_2^2} b - \left[f''(y^{(k)})\right]^{-1} f'(y^{(k)}). \end{aligned}$$

A.4 Eliminating one variable

Another approach to find the corrector direction is to eliminate a variable from (P_c) using the equation:

$$b^\top y \;=\; b^\top y^{(k)} =: \kappa.$$

For instance, we may assume $b_m \neq 0$, then

$$b^\top y \;=\; b^\top y^{(k)} \;\Leftrightarrow\; y_m \;=\; \frac{\left(b^\top y^{(k)} - \sum_{i=1}^{m-1} b_i y_i\right)}{b_m}.$$

So, (P_c) is equivalent to the unconstrained problem:

$$\min \quad h(y_1, y_2, \ldots, y_{m-1}) := f\left(y_1, y_2, \ldots, y_{m-1}, \frac{\left(b^\top y^{(k)} - \sum_{i=1}^{m-1} b_i y_i\right)}{b_m}\right),$$

where $h : \mathbb{R}^{m-1} \to \mathbb{R}$ is a function with $m-1$ variables.

By Newton's Method for unconstrained problem, the corrector direction is:

$$-h''\left(y_1, y_2, \ldots, y_{m-1}\right)^{-1} h'\left(y_1, y_2, \ldots, y_{m-1}\right).$$

The function h can be thought as a composition of two functions: $h = f \circ g$, where

$$g(y_1, y_2, \ldots, y_{m-1}) := \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{m-1} \\ \frac{\left(b^\top y^{(k)} - \sum_{i=1}^{m-1} b_i y_i\right)}{b_m} \end{bmatrix} \tag{A.4}$$

$$= \frac{\kappa}{b_m} e_m + Z \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{m-1} \end{bmatrix} \tag{A.5}$$

is a function from $\mathbb{R}^{m-1}$ to $\mathbb{R}^m$ and $Z = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -\frac{b_1}{b_m} & -\frac{b_2}{b_m} & -\frac{b_3}{b_m} & \cdots & -\frac{b_{m-1}}{b_m} \end{pmatrix}$.

Then,

$$\begin{aligned} h'(y) &= Z^\top f' \left(y_1, y_2, \ldots, y_{m-1}, \frac{\left(b^\top y^{(k)} - \sum_{i=1}^{m-1} b_i y_i\right)}{b_m} \right), \\ h''(y) &= Z^\top f'' \left(y_1, y_2, \ldots, y_{m-1}, \frac{\left(b^\top y^{(k)} - \sum_{i=1}^{m-1} b_i y_i\right)}{b_m} \right) Z. \end{aligned}$$

For the first step, starting with $y := y^{(k)}$, we have:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{m-1} \end{bmatrix}^{new} := \begin{bmatrix} y_1^{(k)} \\ y_2^{(k)} \\ \vdots \\ y_{m-1}^{(k)} \end{bmatrix} - \left(Z^\top f'' \left(y^{(k)} \right) Z \right)^{-1} Z^\top f'(y^{(k)}). \tag{A.6}$$

We denote the m-dimensional vector obtained from the above$(m-1)$-dimensional new y

vector, $y^{new} \in \mathbb{R}^m$. Then,

$$y^{new} := \frac{\kappa}{b_m} e_m + Z \begin{bmatrix} y_1^{(k)} \\ y_2^{(k)} \\ \vdots \\ y_{m-1}^{(k)} \end{bmatrix} - Z \left(Z^\top f'' \left(y^{(k)} \right) Z \right)^{-1} Z^\top f'(y^{(k)}) \tag{A.7}$$

$$= y^{(k)} - Z \left(Z^\top f'' \left(y^{(k)} \right) Z \right)^{-1} Z^\top f'(y^{(k)}). \tag{A.8}$$

Let $\bar{b}^\top$ denote the last row of Z and H_1 denote the submatrix of $f''(y)$ consisting of the first $m-1$ rows and $m-1$ columns. Let H_2 be the first $m-1$ components of the last column of $f''(y)$ and H_3 be the mth entry of the last column. Then, we know that

$$\begin{aligned} Z^\top f''(y) Z &= \begin{bmatrix} I_{m-1} & \bar{b} \end{bmatrix} \begin{bmatrix} H_1 & H_2 \\ H_2^\top & H_3 \end{bmatrix} \begin{bmatrix} I_{m-1} \\ \bar{b}^\top \end{bmatrix} \\ &= \begin{bmatrix} H_1 + \bar{b} H_2^\top & H_2 + H_3 \bar{b} \end{bmatrix} \begin{bmatrix} I_{m-1} \\ \bar{b}^\top \end{bmatrix} \\ &= H_1 + (\bar{b} H_2^\top + H_2 \bar{b}^\top) + H_3 \bar{b} \bar{b}^\top \\ &= H_1 + \begin{bmatrix} \bar{b} & H_2 + H_3 \bar{b} \end{bmatrix} \begin{bmatrix} H_2^\top \\ \bar{b}^\top \end{bmatrix}. \end{aligned}$$

Note that the 2-by-2 matrix $\left\{ I + \begin{bmatrix} H_2^\top \\ \bar{b}^\top \end{bmatrix} H_1^{-1} \begin{bmatrix} \bar{b} & H_2 + H_3 \bar{b} \end{bmatrix} \right\}$ is nonsingular.

$$\begin{aligned} I + \begin{bmatrix} H_2^\top \\ \bar{b}^\top \end{bmatrix} H_1^{-1} \begin{bmatrix} \bar{b} & H_2 + H_3 \bar{b} \end{bmatrix} &= I + \begin{bmatrix} H_2^\top H_1^{-1} \bar{b} & H_2^\top H_1^{-1} (H_2 + H_3 \bar{b}) \\ \bar{b}^\top H_1^{-1} \bar{b} & \bar{b}^\top H_1^{-1} (H_2 + H_3 \bar{b}) \end{bmatrix} \\ &= \begin{bmatrix} 1 + H_2^\top H_1^{-1} \bar{b} & H_2^\top H_1^{-1} (H_2 + H_3 \bar{b}) \\ \bar{b}^\top H_1^{-1} \bar{b} & 1 + \bar{b}^\top H_1^{-1} (H_2 + H_3 \bar{b}) \end{bmatrix}. \end{aligned}$$

The two rows of the above matrix are linearly independent, so it is nonsingular.

Then, by the Sherman-Morrison-Woodbury formula, we have

$$\begin{aligned} \left[Z^\top f''(y) Z \right]^{-1} &= H_1^{-1} + H_1^{-1} \begin{bmatrix} \bar{b} & H_2 + H_3 \bar{b} \end{bmatrix} \left\{ I + \begin{bmatrix} H_2^\top \\ \bar{b}^\top \end{bmatrix} H_1^{-1} \begin{bmatrix} \bar{b} & H_2 + H_3 \bar{b} \end{bmatrix} \right\}^{-1} \begin{bmatrix} H_2^\top \\ \bar{b}^\top \end{bmatrix} H_1^{-1} \\ &= H_1^{-1} + \begin{bmatrix} H_1^{-1} \bar{b} & H_1^{-1} (H_2 + H_3 \bar{b}) \end{bmatrix} \left\{ I + \begin{bmatrix} H_2^\top \\ \bar{b}^\top \end{bmatrix} H_1^{-1} \begin{bmatrix} \bar{b} & H_2 + H_3 \bar{b} \end{bmatrix} \right\}^{-1} \begin{bmatrix} H_2^\top H_1^{-1} \\ \bar{b}^\top H_1^{-1} \end{bmatrix} \\ &= H_1^{-1} + \begin{bmatrix} H_1^{-1} \bar{b} & H_1^{-1} (H_2 + H_3 \bar{b}) \end{bmatrix} C^{-1} \begin{bmatrix} H_2^\top H_1^{-1} \\ \bar{b}^\top H_1^{-1} \end{bmatrix}, \end{aligned}$$

where $C = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$, $c_{11} = 1 + H_2^\top H_1^{-1}\bar{b}$, $c_{22} = 1 + \bar{b}^\top H_1^{-1}(H_2 + H_3\bar{b})$, $c_{12} = H_2^\top H_1^{-1}(H_2 + H_3\bar{b})$, and $c_{21} = \bar{b}^\top H_1^{-1}\bar{b}$.

Hence, we have that:

$$
\begin{aligned}
& Z\left(Z^\top f''\left(y^{(k)}\right) Z\right)^{-1} Z^\top \\
= & \begin{bmatrix} I_{m-1} \\ \bar{b}^\top \end{bmatrix} \left\{ H_1^{-1} + \begin{bmatrix} H_1^{-1}\bar{b} & H_1^{-1}(H_2 + H_3\bar{b}) \end{bmatrix} C^{-1} \begin{bmatrix} H_2^\top H_1^{-1} \\ \bar{b}^\top H_1^{-1} \end{bmatrix} \right\} \begin{bmatrix} I_{m-1} & \bar{b} \end{bmatrix} \\
= & \begin{bmatrix} I_{m-1} \\ \bar{b}^\top \end{bmatrix} H_1^{-1} \begin{bmatrix} I_{m-1} & \bar{b} \end{bmatrix} + \begin{bmatrix} I_{m-1} \\ \bar{b}^\top \end{bmatrix} \left\{ \begin{bmatrix} H_1^{-1}\bar{b} & H_1^{-1}(H_2 + H_3\bar{b}) \end{bmatrix} C^{-1} \begin{bmatrix} H_2^\top H_1^{-1} \\ \bar{b}^\top H_1^{-1} \end{bmatrix} \right\} \begin{bmatrix} I_{m-1} & \bar{b} \end{bmatrix} \\
= & \begin{bmatrix} H_1^{-1} & H_1^{-1}\bar{b} \\ \bar{b}^\top H_1^{-1} & \bar{b}^\top H_1^{-1}\bar{b} \end{bmatrix} + \left\{ \begin{bmatrix} I_{m-1} \\ \bar{b}^\top \end{bmatrix} \begin{bmatrix} H_1^{-1}\bar{b} & H_1^{-1}(H_2 + H_3\bar{b}) \end{bmatrix} \right\} C^{-1} \left\{ \begin{bmatrix} H_2^\top H_1^{-1} \\ \bar{b}^\top H_1^{-1} \end{bmatrix} \begin{bmatrix} I_{m-1} & \bar{b} \end{bmatrix} \right\},
\end{aligned}
$$

where $C^{-1} = \frac{1}{(c_{11} + H_3 c_{21})c_{11} - c_{21}c_{12}} \cdot \begin{bmatrix} c_{22} & -c_{12} \\ -c_{21} & c_{11} \end{bmatrix}$ Let $\bar{c} := (c_{11} + H_3 c_{21})c_{11} - c_{21}c_{12}$

Then,

$$
\begin{aligned}
& \left\{ \begin{bmatrix} I_{m-1} \\ \bar{b}^\top \end{bmatrix} \begin{bmatrix} H_1^{-1}\bar{b} & H_1^{-1}(H_2 + H_3\bar{b}) \end{bmatrix} \right\} C^{-1} \left\{ \begin{bmatrix} H_2^\top H_1^{-1} \\ \bar{b}^\top H_1^{-1} \end{bmatrix} \begin{bmatrix} I_{m-1} & \bar{b} \end{bmatrix} \right\} \\
= & \begin{bmatrix} H_1^{-1}\bar{b} & H_1^{-1}(H_2 + H_3\bar{b}) \\ \bar{b}^\top H_1^{-1}\bar{b} & \bar{b}^\top H_1^{-1}(H_2 + H_3\bar{b}) \end{bmatrix} C^{-1} \begin{bmatrix} H_2^\top H_1^{-1} & H_2^\top H_1^{-1}\bar{b} \\ \bar{b}^\top H_1^{-1} & \bar{b}^\top H_1^{-1}\bar{b} \end{bmatrix} \\
= & \frac{1}{\bar{c}} \begin{bmatrix} H_1^{-1}\bar{b} & H_1^{-1}(H_2 + H_3\bar{b}) \\ \bar{b}^\top H_1^{-1}\bar{b} & \bar{b}^\top H_1^{-1}(H_2 + H_3\bar{b}) \end{bmatrix} \begin{bmatrix} 1 + \bar{b}^\top H_1^{-1}(H_2 + H_3\bar{b}) & -H_2^\top H_1^{-1}(H_2 + H_3\bar{b}) \\ -\bar{b}^\top H_1^{-1}\bar{b} & 1 + H_2^\top H_1^{-1}\bar{b} \end{bmatrix} \\
& \begin{bmatrix} H_2^\top H_1^{-1} & H_2^\top H_1^{-1}\bar{b} \\ \bar{b}^\top H_1^{-1} & \bar{b}^\top H_1^{-1}\bar{b} \end{bmatrix} \\
= & \frac{1}{\bar{c}} \begin{bmatrix} (1 + \bar{b}^\top H_1^{-1}H_2)H_1^{-1}\bar{b} & (1 + H_2^\top H_1^{-1}\bar{b})H_1^{-1}H_2 \\ -(\bar{b}^\top H_1^{-1}\bar{b})H_1^{-1}H_2 & +(H_3 - H_2^\top H_1^{-1}H_2)H_1^{-1}\bar{b} \\ \bar{b}^\top H_1^{-1}\bar{b} & (1 + \bar{b}^\top H_1^{-1}H_2)\bar{b}^\top H_1^{-1}H_2 \\ & +(H_3 - H_2^\top H_1^{-1}H_2)\bar{b}^\top H_1^{-1}\bar{b} \end{bmatrix} \begin{bmatrix} H_2^\top H_1^{-1} & H_2^\top H_1^{-1}\bar{b} \\ \bar{b}^\top H_1^{-1} & \bar{b}^\top H_1^{-1}\bar{b} \end{bmatrix} \\
= & \frac{1}{\bar{c}} \begin{bmatrix} (1 + \bar{b}^\top H_1^{-1}H_2)\left[H_1^{-1}\bar{b}H_2^\top H_1^{-1} + (H_1^{-1}\bar{b}H_2^\top H_1^{-1})^\top\right] & \left[(1 + \bar{b}^\top H_1^{-1}H_2)(\bar{b}^\top H_1^{-1}H_2)\right] H_1^{-1}\bar{b} \\ +(H_3 - H_2^\top H_1^{-1}H_2)H_1^{-1}\bar{b}\bar{b}^\top H_1^{-1} & +(\bar{b}^\top H_1^{-1}\bar{b})(H_3 - H_2^\top H_1^{-1}H_2)H_1^{-1}\bar{b} \\ -(\bar{b}^\top H_1^{-1}\bar{b})[H_1^{-1}H_2H_2^\top H_1^{-1} & +(\bar{b}^\top H_1^{-1}\bar{b})H_1^{-1}H_2 \\ \left[(\bar{b}^\top H_1^{-1}\bar{b})H_2^\top + (1 + \bar{b}^\top H_1^{-1}H_2)(\bar{b}^\top H_1^{-1}H_2)\bar{b}^\top\right] H_1^{-1} & (2 + \bar{b}^\top H_1^{-1}H_2)(\bar{b}^\top H_1^{-1}H_2)(\bar{b}^\top H_1^{-1}\bar{b}) \\ +(H_3 - H_2^\top H_1^{-1}H_2)(\bar{b}^\top H_1^{-1}\bar{b})\bar{b}^\top H_1^{-1} & +(H_3 - H_2^\top H_1^{-1}H_2)(\bar{b}^\top H_1^{-1}\bar{b})^2 \end{bmatrix}.
\end{aligned}
$$

We know that the above matrix is an m-by-m matrix with rank 2. So, $\left\{Z\left(Z^\top f''\left(y^{(k)}\right)Z\right)^{-1}Z^\top\right\}$ is a rank 2 update of the matrix $\begin{bmatrix} H_1^{-1} & H_1^{-1}\bar{b} \\ \bar{b}^\top H_1^{-1} & \bar{b}^\top H_1^{-1}\bar{b} \end{bmatrix}$. Hence, it means that the matrix $\left\{Z\left(Z^\top f''\left(y^{(k)}\right)Z\right)^{-1}Z^\top\right\}$ is close to the matrix $\begin{bmatrix} H_1^{-1} & H_1^{-1}\bar{b} \\ \bar{b}^\top H_1^{-1} & \bar{b}^\top H_1^{-1}\bar{b} \end{bmatrix}$.

Let R denote the above matrix, i.e, the rank 2 update. Let $f'(y)_{1:m-1}$ denote the $(m-1)$-dimensional vector obtained from taking the first $m-1$ entries of $f'(y^{(k)})$ and let$f'(y)_m$ denote the last entry of $f'(y^{(k)})$.

Therefore, we have

$$\begin{aligned} & Z\left(Z^\top f''\left(y^{(k)}\right)Z\right)^{-1}Z^\top f'(y^{(k)}) \\ = & \begin{bmatrix} H_1^{-1} & H_1^{-1}\bar{b} \\ \bar{b}^\top H_1^{-1} & \bar{b}^\top H_1^{-1}\bar{b} \end{bmatrix} f'(y^{(k)}) + Rf'(y^{(k)}) \\ = & \begin{bmatrix} H_1^{-1} & H_1^{-1}\bar{b} \\ \bar{b}^\top H_1^{-1} & \bar{b}^\top H_1^{-1}\bar{b} \end{bmatrix} \begin{bmatrix} f'(y)_{1:m-1} \\ f'(y)_m \end{bmatrix} + Rf'(y^{(k)}) \\ = & \begin{bmatrix} H_1^{-1}f'(y)_{1:m-1} + (f'(y)_m)H_1^{-1}\bar{b} \\ \\ \bar{b}^\top H_1^{-1}f'(y)_{1:m-1} + (\bar{b}^\top H_1^{-1}\bar{b})f'(y)_m \end{bmatrix} + Rf'(y^{(k)}). \end{aligned}$$

Using the idea of finding Schur Complement, we have the following:

$$\begin{aligned} \begin{bmatrix} H_1 & H_2 \\ H_2^\top & H_3 \end{bmatrix} &= L \begin{bmatrix} H_1 & 0 \\ 0^\top & H_3 - H_2^\top H_1^{-1}H_2 \end{bmatrix} L^\top \\ \iff [f''(y)]^{-1} &= L^{-1} \begin{bmatrix} H_1^{-1} & 0 \\ 0^\top & \frac{1}{H_3 - H_2^\top H_1^{-1}H_2} \end{bmatrix} L^{-\top} \end{aligned}$$

, where $L = \begin{bmatrix} I_{m-1} & 0 \\ H_2^\top H_1^{-1} & 1 \end{bmatrix}$ and $L^{-1} = \begin{bmatrix} I_{m-1} & 0 \\ -H_2^\top H_1^{-1} & 1 \end{bmatrix}$. Therefore, $[f''(y)]^{-1}$ is a rank-1 perturbation of the matrix $\begin{bmatrix} H_1^{-1} & 0 \\ 0^\top & \frac{1}{H_3 - H_2^\top H_1^{-1}H_2} \end{bmatrix}$. So, $\left\{Z^\top\left[Z^\top f''(y)Z\right]^{-1}Z\right\}$ is at most a rank-3 perturbation of the matrix $\begin{bmatrix} H_1^{-1} & H_1^{-1}\bar{b} \\ \bar{b}^\top H_1^{-1} & \bar{b}^\top H_1^{-1}\bar{b} \end{bmatrix}$.

A.5 Using optimality conditions and Newton's method

Based on the problem (P_c), we can write down the optimality conditions of (P_c) using KKT Theorem:

$$\begin{array}{rcl} f'(y) - \lambda b & = & 0 \\ b^\top y & = & b^\top y^{(k)} \\ \lambda & \in & \mathbb{R}. \end{array}$$

Then, we can apply Newton's Method to the above system to obtain:

$$f''(y)dy - (d\lambda)b = \lambda b - f'(y) \tag{A.9}$$

$$b^\top dy = 0. \tag{A.10}$$

Multiply $b^\top f''(y)^{-1}$ on the left to equation (A.9) and substituting (A.10) into it gives us:

$$d\lambda = \frac{b^\top f''(y)^{-1} f'(y)}{b^\top f''(y)^{-1} b} - \lambda.$$

Then, we substitute into (A.9) to get the formula for the corrector direction:

$$\begin{array}{rcl} dy & = & (d\lambda + \lambda) f''(y)^{-1} b - f''(y)^{-1} f'(y) \\ & = & \left(\dfrac{b^\top f''(y)^{-1} f'(y)}{b^\top f''(y)^{-1} b} \right) f''(y)^{-1} b - f''(y)^{-1} f'(y). \end{array}$$

A.6 Comparison of five corrector directions

In order to analyze the efficiency of computing the five different corrector directions, we want to first compare all five corrector directions.

From A.1 and A.2, we note that the expressions of the first and the second corrector steps are the same.

Now, we will first compare the first corrector direction and the fifth one.

If the current point is on the central path, then we know that $f'(y) = \frac{1}{\mu} b$. Therefore, $\left(\frac{b^\top f''(y)^{-1} f'(y)}{b^\top f''(y)^{-1} b} \right) = \frac{1}{\mu}$. So, the coefficient of $\left[f''(y^{(k)}) \right]^{-1} b$ in the fifth corrector direction is $\frac{1}{\mu}$. Then, the fifth corrector direction is the same as the first one.

Otherwise, we can write the coefficient of $\left[f''(y^{(k)})\right]^{-1} b$ in the fifth corrector direction as follows:

$$\begin{aligned}\left(\frac{b^\top f''(y)^{-1} f'(y)}{b^\top f''(y)^{-1} b}\right) &= \frac{1}{b^\top f''(y)^{-1} b}\left[b^\top f''(y)^{-1}\frac{1}{\mu}b - b^\top f''(y)^{-1}\frac{1}{\mu}b + b^\top f''(y)^{-1} f'(y)\right] \\ &= \frac{1}{\mu} + \frac{b^\top f''(y)^{-1}\left(f'(y) - \frac{1}{\mu}b\right)}{b^\top f''(y)^{-1} b}.\end{aligned}$$

Therefore, the difference in the coefficient of $\left[f''(y^{(k)})\right]^{-1} b$ between the fifth corrector direction and the first corrector direction is:

$$\begin{aligned}\left|\frac{b^\top f''(y)^{-1}\left(f'(y) - \frac{1}{\mu}b\right)}{b^\top f''(y)^{-1} b}\right| &\leq \frac{\left|\left|f''(y)^{-\frac{1}{2}}b\right|\right|_2 \cdot \left|\left|f''(y)^{-\frac{1}{2}}(f'(y) - \frac{1}{\mu}b)\right|\right|_2}{\left|\left|f''(y)^{-\frac{1}{2}}b\right|\right|_2^2} \\ &= \frac{\left|\left|f''(y)^{-\frac{1}{2}}(f'(y) - \frac{1}{\mu}b)\right|\right|_2}{\left|\left|f''(y)^{-\frac{1}{2}}b\right|\right|_2} \\ &= \frac{\gamma(y,\mu)}{\left|\left|f''(y)^{-\frac{1}{2}}b\right|\right|_2}.\end{aligned}$$

Next, we will compare the first corrector direction with the third corrector direction. Since they both have the common term $\left[f''(y^{(k)})\right]^{-1} f'(y^{(k)})$, it suffices to just compare the first term in each expression.

Consider the first term of the third corrector step as follows:

$$\frac{b^\top\left[f''(y^{(k)})\right]^{-1} f'(y^{(k)})}{||b||_2^2} b - \frac{1}{\mu}\left[f''(y^{(k)})\right]^{-1} b + \frac{1}{\mu}\left[f''(y^{(k)})\right]^{-1} b.$$

If the current point is on the central path, then we will have:

$$\begin{aligned}&\frac{b^\top\left[f''(y^{(k)})\right]^{-1} f'(y^{(k)})}{||b||_2^2} b - \frac{1}{\mu}\left[f''(y^{(k)})\right]^{-1} b \\ =\ & \frac{1}{\mu}\left[\frac{||b||_y}{||b||_2} b - \left[f''(y^{(k)})\right]^{-1} b\right] \\ =\ & \frac{1}{\mu}\left[\frac{||b||_y}{||b||_2} I - \left[f''(y^{(k)})\right]^{-1}\right] b.\end{aligned}$$

If b is an eigenvector of $f''(y^{(k)})$ determining the eigenvalue $\frac{||b||_2}{||b||_y}$, then these two corrector directions coincide.

In our preliminary computational experiments we found that the corrector direction derived by the methods given in Appendices A.1 and A.2 worked the best.