NEWTON-TYPE METHODS UNDER GENERALIZED SELF-CONCORDANCE AND INEXACT ORACLES

Tianxiao Sun

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2018

Approved by:

Quoc Tran-Dinh

Shu Lu

Nilay Argon

Gabor Pataki

Kai Zhang

ABSTRACT

TIANXIAO SUN: Newton-type methods under generalized self-concordance and inexact oracles
(Under the direction of Quoc Tran-Dinh and Shu Lu)

Many modern applications in machine learning, image/signal processing, and statistics require to solve large-scale convex optimization problems. These problems share some common challenges such as high-dimensionality, nonsmoothness, and complex objectives and constraints. Due to these challenges, the theoretical assumptions for existing numerical methods are not satisfied. In numerical methods, it is also impractical to do exact computations in many cases (e.g. noisy computation, storage or time limitation). Therefore, new approaches as well as inexact computations to design new algorithms should be considered.

In this thesis, we develop fundamental theories and numerical methods, especially second-order methods, to solve some classes of convex optimization problems, where first-order methods are inefficient or do not have a theoretical guarantee. We aim at exploiting the underlying smoothness structures of the problem to design novel Newton-type methods. More specifically, we generalize a powerful concept called self-concordance introduced by Nesterov and Nemirovski to a broader class of convex functions. We develop several basic properties of this concept and prove key estimates for function values and its derivatives. Then, we apply our theory to design different Newton-type methods such as damped-step Newton methods, full-step Newton methods, and proximal Newton methods. Our new theory allows us to establish both global and local convergence guarantees of these methods without imposing unverifiable conditions as in classical Newton-type methods. Numerical experiments show that our approach has several advantages compared to existing works.

In the second part of this thesis, we introduce new global and local inexact oracle settings, and apply them to develop inexact proximal Newton-type schemes for optimizing general composite convex problems equipped with such inexact oracles. These schemes allow us to measure errors theoretically and systematically and still lead to desired convergence results. Moreover, they can be applied to solve a wider class of applications arising in statistics and machine learning.

To my parents,

Lidan Bai and Zhenwei Sun,

who have loved and supported me throughout my life.

ACKNOWLEDGEMENTS

I would like to gratefully acknowledge the guidance, support, and encouragement of my doctoral advisors, Dr. Quoc Tran-Dinh and Dr. Shu Lu. I have been extremely lucky to have two supervisors who cared so much about my research and training, and responded to my questions and queries so promptly. I would like to thank Dr. Ion Necoara. My collaboration with you has been a very enjoyable and valuable part of my graduate study. I would like to thank my committee members Dr. Kai Zhang, Dr. Nilay Argon and Dr. Gabor Pataki during my time at STOR Department of UNC, for your valuable suggestions about my work. Without your expertise, I would not have been able to continue and complete my projects presented in this dissertation. I would also like to thank for the partial support for my research of the NSF-grant No.DMS-1619884, USA.

Secondly, I would like to thank all my friends during my PhD period, especially Mr. Yuqin Mu, Mr. Haipeng Gao, Ms. Yanni Lai, Mr. Hongsheng Liu and Dr. Yifan Cui. I treat you as my examples and the source of inspiration in my heart all the time. Besides, you have brought me so much joy, and taught me a lot about mathematics as well as how to be a good person.

Thirdly, I would like to thank all the scholars, professors, and my peers in the field of operations research that I have recognized or haven't recognized. No matter who you are, what languages you speak, and how big the contributions you made, we are ultimately doing the same thing – optimizing the world. Because of our common goal, I am linked with all of you closely, honored and humbled.

Besides, I would like to thank myself. Research is simple, but not easy. Countless of midnight you are struggling in front of the computer, reading and editing papers, and debugging codes once and once again – making your own efforts to move forward, and finally you made this tiny contribution, which pushes you to do faster, stabler and more accurate in future. Never forget why you started, and your mission can be accomplished.

Last but not least, I would like to thank my parents, Lidan Bai and Zhenwei Sun, for your endless love and support to me. It was your patience, experience, and continuous encouragement that helped me conquer all the difficulties and hardships during my research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

$\mathrm{dom}(f)$	Domain of function f, all $\mathbf{x}$'s such that $\lvert f(\mathbf{x})\rvert < \infty$
$\mathbb{I}_p$	Identity matrix of size p
$\mathcal{F}_L^1$	Lipschitz gradient function class with parameter L
$\lVert\cdot\rVert$	Matrix or vector norm; ℓ_2-norm if the norm is unspecified
$\mathcal{C}^n(S)$	n times continuously differentiable function on set S
$\mathbf{x}_f^\star$	Optimal solution of the [constrained] optimization problem $\min_{\mathbf{x}} f(\mathbf{x})$
$\widetilde{\mathcal{F}}_{M,\nu}^p$	p dimensional *generalized self-concordant* class of parameters M and ν
$\mathcal{S}_{++}^p$	p dimensional positive definite cone
$\mathcal{S}_{+}^p$	p dimensional positive semi-definite cone
$\mathbb{R}^p$	Set of p dimensional real-valued vectors
$\mathbb{R}_+^p$	Set of p dimensional real-valued nonnegative vectors
$\mathbb{R}_{++}^p$	Set of p dimensional real-valued positive vectors
$\mathrm{prox}_g(\cdot)$	Standard proximal operator of g
$\mathcal{L}_F(\mathbf{x})$	Sublevel set of F at a point $\mathbf{x}$, all $\mathbf{y} \in \mathrm{dom}(F)$ such that $F(\mathbf{y}) \leq F(\mathbf{x})$
$\mathrm{Dom}(f)$	The closure of the domain of f, all $\mathbf{x}$'s such that $\lvert f(\mathbf{x})\rvert \leq \infty$
$f^*(\cdot)$	The Fenchel/convex conjugate of function $f(\cdot)$
$\mathrm{ri}(S)$	The relative interior of set S
$\lambda_{\min}(\mathbf{H})$	The smallest eigenvalue of a positive [semi-]definite matrix $\mathbf{H}$
$\lVert\cdot\rVert_{\mathbf{x}}$	Weighted norm based on Hessian matrix at point $\mathbf{x}$ ($\nabla^2 f(\mathbf{x})$)
$\lVert\!\lVert\cdot\rVert\!\rVert_{\mathbf{x}}$	Weighted norm based on some positive definite matrix at point $\mathbf{x}$ ($H(\mathbf{x})$)
w.r.t	With respect to

CHAPTER 1

Introduction

1.1 Introduction

This thesis is about the theory and foundation of Newton-type methods, under the generalized self-concordant structure and inexact global and local oracles. The Newton method is an important computational tool for solving smooth minimization problems as well as smooth equations. Its variant including proximal Newton methods, primal-dual methods, stochastic Newton-type methods, and quasi-Newton methods, can be viewed as an advanced tool for non-smooth, constrained, large-scale, or distributed settings of these problems. In the first part of this thesis, we study a smooth structure called *generalized self-concordance*. Like the self-concordance concept introduced by Nesterov and Nemirovski in the early 1990s, generalized self-concordance serves as a powerful but more general analogous structure that allows ones to develop Newton-type methods with rigorous theoretical convergence guarantees, while treating a broader class of convex and smooth functions than the former one. In the second part of this thesis, we introduce new global and local inexact oracle concepts for a wide class of convex functions in composite convex minimizations, and use them to develop inexact Newton-type methods. This topic is motivated by the fact that many numerical methods, especially second-order methods, naturally use inexact computations as well as oracles due to limited memory, noisy data or limited computational time. Unfortunately, this inexact computational issue has not been theoretically characterized in a full setting. We cover this topic in the second part of this thesis.

Our work in this thesis provides an alternative view of using smoothness structures of convex functions, as well as a control of bounds of function values, derivatives and subproblem inexactness derived from our oracle settings, to develop Newton-type methods. Both theories in this thesis lead to the best convergence rates compared with existing methods. We will also see

that the theory and algorithms developed using generalized self-concordance and our inexact oracles have many important and interesting applications. Moreover, our numerical experiments show that the new theories provide competitive or better results compared to state-of-the-art algorithms.

In the rest of this chapter, we first present our motivation and describe our research goals. Next, we clarify our contribution and review current research state. Finally, we give a short outline of each chapter coved in this thesis.

1.1.1 Motivation

In recent years, there has been a huge interest in Newton-type methods for solving convex optimization problems and monotone equations due to the development of new techniques and mathematical tools in optimization, machine learning, and randomized algorithms [6, 14, 27, 29, 34, 60, 72, 76, 87, 89, 93, 94]. Several combinations of Newton-type methods and other techniques such as proximal operators [10], cubic regularization [76], gradient regularization [89], randomized algorithms such as sketching [87], subsampling [34], and fast eigen-decomposition [46] create a new research direction and have attracted a great attention in solving nonsmooth and large-scale optimization problems.

A wide range of problems especially in signal and image processing can be expressed in a particular composite [primal] form, where the dual may be much easier to solve than the primal one, due to the splitting structure of the dual settings. Correspondingly, one often study methods which have ability to split the problem by activating each of the functions through elementary processing steps which can be computed in parallel [56]. The primal-dual methods make this possible, by exploiting the structure of the problem in a flexible manner, especially when combined with the Lipschitz property of objective gradient or/and Hessian mappings.

While standard assumptions often required in both primal and dual methods, such as non-singularity, Lipschitz gradient and Hessian conditions do not hold for many examples, Nesterov and Nemirovskii [75] introduced a powerful concept called self-concordance to overcome this drawback and developed new Newton scheme to achieve global and local convergence without requiring any additional assumption, or a globalization strategy. The self-concordance notion

was initially invented to study interior-point methods, but it is less well-known in other communities. Recent works [1, 68, 102, 104, 109] have greatly popularized this concept to solve other problems arising from machine learning, statistics, image processing, and variational inequalities.

Unfortunately, although the above self-concordant extension is improtant, it still cannot cover some other basic or important convex functions, such as power function, entropy, arcsine distribution, just to name a few. Motivated by this fact, we extend it to a more general smooth structure, develop the corresponding Newton-type schemes and customize it to the dual settings, which covers the standard self-concordant optimization as a special case.

All the schemes above are mainly dealing with exact Newton methods. However, from computational viewpoint, error measurement during evaluation, storage and transfer of data happens frequently in sequential methods or distributed and parallel computation. Besides, due to technical or complexity limitation, we often need inexact evaluations of function values and derivatives. While existing inexact methods mostly focus on the inexactness of subproblem [63] or first-order oracles [28], there is no intensive work on inexact second-order oracles to the best of our knowledge. Motivative by this, we introduce new global and local inexact second-order oracle concepts, which allow one to develop novel inexact Newton-type variants that have the desired convergence guarantees by direct control of function value and derivative tolerances, and include the subproblem inexactness routines as special cases.

1.1.2 The goals of this research

Motivated by [1, 103, 109], our first goal is to generalize the self-concordance concept in [75] to a broader class of smooth and convex functions. To develop the corresponding methods, we require the generalization from univariate to multivariate case to preserve some key properties. Unfortunately, the preliminary attempt shows that the natural generalization has several drawbacks when developing the theory. Besides, similar extensions in [1, 103] for a class of logistic-type functions are still limited and creates certain difficulty for developing further theories. Therefore, we first introduce a new definition of generalized self-concordance that fixed all these drawbacks.

Our second goal is to develop a unified mechanism to analyze the convergence (including global and local convergence) of the following Newton-type scheme:

$$\mathbf{x}^{k+1} := \mathbf{x}^k - s_k F'(\mathbf{x}^k)^{-1} F(\mathbf{x}^k), \tag{1.1}$$

where F' is the Jacobian of F, $s_k \in (0, 1]$ is a given step-size, and F can be presented as the right-hand-side of a monotone equation $F(x) = 0$ or the optimality condition of a convex optimization or a convex-concave saddle-point problem. Despite the Newton scheme (1.1) is invariant to a change of variables [27], its convergence property relies on the growth of the Hessian along the Newton iterative process. In classical settings, the Lipschitz continuity of the Hessian and the nondegeneracy of the Hessian in a neighborhood of the solution set are key assumptions to achieve local quadratic convergence rate [27]. These assumptions have been considered to be standard, but they are often very difficult to check in practice, especially the second one. A natural idea is to classify the functionals of the underlying problem into a known class of functions to choose a suitable method for solving it. While first-order methods for convex optimization essentially rely on the Lipschitz gradient function assumption, Newton methods usually use the Lipschitz continuity of the Hessian and its nondegeneracy property to obtain a well-defined Newton direction as we have mentioned. For self-concordant functions, the second condition automatically holds, while the first does not. However, both full-step and damped-step Newton methods still work in this case by appropriately choosing a suitable metric. This situation has been observed and standard assumptions have been modified in different directions to still guarantee the convergence of Newton methods, see [27] for an intensive study of generic Newton methods, and [74, 75] for the self-concordant function class.

Thirdly, we want to combine our theory and methods with other broader settings in order to make our scheme stabler and more efficient. For example, in the field of machine learning , one has to deal with truly massive datasets and to train very large models, which naturally leads to high-dimensional optimization problems. Hence, computational accuracy and efficiency constitute two major issues that need to be thoroughly addressed. We wish to develop inexact oracle theory that can lower the computational accuracy of function value, derivative(s), solution of subproblem, or/and (proximal) Newton decrement, while still guarantee desired solution

accuracy and convergence rates both globally and locally. As a special case, we customize our theory to handle the primal-dual setting. As its name implied, this approach acts by solving the primal problem as well as dual formulation simultaneously, or even merely in the dual space. By doing so, we are able to exploit the structure of the underlying model more efficiently.

1.1.3 Literature review and the state of current research

The Newton-type method in convex optimization is often referred to as a second-order method. It is widely used to solve both unconstrained and constrained optimization problems [45, 65]. It is popular among past several decades, mainly because of its fast local convergence rate given the method is convergent. However, due to the unclearness of the global convergence and the high per-iteration computational cost, it has been dominated by first-order methods for solving modern large-scale optimization problems. For example, (a) the alternating direction method of multipliers (ADMM) which is closely related to [33], is a simple but powerful algorithm that is well adapted to parallel and distributed optimization algorithms [18, 25], and in particular to problems arising in image processing, applied statistics and machine learning, where the objectives can be even nondifferentiable. (b) The Frank-Wolfe method, also known as the conditional gradient method, was originally developed for smooth convex optimization on a polytope, dated back from Frank and Wolfe [37]. It is still popular among many application such as sparse convex optimization [54], particle filtering [58], and support vector machine [82], due to its low per-iteraion cost and good practical performance. (c) The methods that use Nesterov's acceleration and smoothness techniques [73] such as fast iterative shrinkage thresholding algorithm (FISTA) [4] and Nesterov's algorithm (NESTA) [5] are even used as a criterion to test the performance of new methods. However, those kind of methods often require strong smoothness structure assumptions such as the most commonly used Lipschitz gradient and strongly convexity, which do not hold in many important applications. This brings the Newton-type method back to its life. To overcome the difficulty of strong impractical assumptions in first-order methods, the self-concordance concept was introduced in 1990s by Nestorov and Nemirovski [75], as an innovative way of exploiting smoothness structures of convex optimization problems. Since the self-concordance theory was introduced lately, its first extension

was proposed by [1] for a class of logistic regression. In [103], the authors extended [1] to study proximal Newton method for logistic, multinomial logistic, and exponential loss functions. By augmenting a strongly convex regularizer, Zhang and Lin [109] showed that the regularized logistic regression is indeed standard self-concordant. In [2] Bach continued exploiting his results in [1] to show that the averaging stochastic gradient method can achieve the same best known convergence rate as in strongly convex case without adding a regularizer. In our recent work [104], we developed a new generalized Newton-type framework to solve a large class of self-concordant inclusion problem, and can achieve the same worst-case complexity as in standard path-following method for smooth convex programming [75].

To overcome the difficulties of high per-iteration complexity of traditional Newton methods and further accelerate the algorithm, both decentralized storage of big data as well as accompanying distributed computation are necessary or at least highly desirable. In [102], the authors exploited standard self-concordance theory in [75] to develop several classes of optimization algorithms including proximal Newton, proximal-quasi Newton and proximal gradient methods to solve composite convex minimization problems. In a recent paper [40], Gao and Goldfarb studied quasi-Newton methods for self-concordant minimization problems. In our recent work [97], we made a broader generalization of the self-concordant concept, and developed the corresponding Newton and quasi-Newton-type methods, which covers [1, 29, 109] as special cases. In addition to deterministic approaches, randomized algorithms and stochastic methods have been also well developed. Along with stochastic gradient descents and coordinate descent schemes, subsampled and sketching Newton-type methods have recently gained a great attention. Lu [66] extended [102] to study randomized block coordinate descent methods. In addition, we refer to [55, 87, 93] for further related works.

To accommodate with data-related errors and reduce the computational complexity, inexact methods have been widely studied recently. Among the first-order frameworks, [28] provides a general inexact first-order oracle that covers a wide class of objective functions, including nonsmooth functions, and covering many other existing inexact first-order oracles as special cases. However, [28] only studied a global first-order inexact oracle to analyze the behavior for first-order methods of smooth convex optimization. Such an oracle cannot be used to study the local behavior of second-order methods, in particular, for self-concordant functions. In quasi-

Newton algorithms, secant equations are usually used to approximate the Hessian mapping [77]. We show in Chapter 5 that this setup can also be cast into our Newton-type methods with inexact oracles. Alternative to deterministic inexact oracles, stochastic gradient type schemes can be viewed as optimization methods with inexact oracles [96]. Function values and gradients are approximated by a stochastic sampling scheme to obtain inexact oracles. Finally, derivative-free optimization can be considered as optimization methods with inexact oracles as well [23].

With the rapid development of computational power, the big advance in acceleration techniques, and the considerable progress in algorithms, we believe that Newton-type methods will eventually play a major role in the future.

1.2 Contribution

Our contribution of this thesis is twofolds: **theory** and **numerical algorithms**, which can be summarized as follows.

Theoretical contribution:

(a) We generalize the self-concordant notion in [74] to a more broader class of smooth convex functions, which we call generalized self-concordance. We identify several link functions that can be cast into our generalized self-concordant class. We also prove several fundamental properties and show that the *generalized self-concordant* class is closed with respect to the basic affine transformation, for a given range of parameters or under suitable assumptions. In addition, we develop lower and upper bounds on the Hessian, gradient, and function values for *generalized self-concordant* functions. These estimates are key to develop and analyze several numerical optimization methods including Newton-type methods.

(b) We introduce new global and local inexact second-order oracles for a large class of convex functions. Such a global inexact oracle covers a wide range of convex functions including smooth convex functions with Lipschitz gradient continuity, nonsmooth Lipschitz continuous convex functions with bounded domain, and self-concordant convex functions. For

the local inexact oracle, we limit our consideration to the class of self-concordant functions. Relying on these global and local inexact oracles, we develop several key properties that are useful for algorithm development.

Algorithmic contribution:

(a) We propose a class of (proximal) Newton methods including damped-step and full-step schemes to minimize a (composite) *generalized self-concordant* function. We show explicitly how to choose a suitable step-size to guarantee a descent direction in the damped-step scheme, and prove a local quadratic convergence for both damped-step and full-step schemes using a suitable metric.

(b) We develop a proximal-Newton algorithm based on inexact oracles and approximate computations of the proximal-Newton directions to solve composite minimization (5.1). Our global inexact oracle allows us to prove a general convergence result for the proposed proximal-Newton method. When limited to self-concordant class for f, by using the new local inexact oracle, we show how to adapt the inner accuracy parameters of the oracles so that our algorithm still enjoys a global convergence guarantee, while having either R-linear, R-superlinear, or R-quadratic local convergence rate.

(c) Finally, we customize our inexact method to handle a class of convex programs in the primal-dual setting, where our method is applied to solve the dual problem. This particular application provides a new primal-dual method for handling some classes of convex optimization problems including constrained formulations.

Let us emphasize the following of our contribution. First, we observe that the self-concordance notion is a powerful concept and has been widely used in interior-point methods as well as in other optimization schemes [49, 66, 102, 109]. Therefore, generalizing it to a broader class of smooth convex functions can substantially cover a number of new applications, and is helpful to develop new methods for solving classical problems including logistic and multimonomial logistic regression, optimization involving exponential objectives, and distance-weighted discrimination problems in classification (see Table 3.1 below). Second, verifying theoretical assumptions for convergence guarantees of a Newton method is not trivial, our theory allows

one to classify the underlying functions into different subclasses by using different parameters ν and M_φ and to choose suitable algorithms to solve the corresponding optimization problem. Third, the theory developed in this chapter can potentially apply to other optimization methods such as gradient-type, sketching and sub-sampling Newton, and Frank-Wolfe algorithms as done in the literature [80, 87, 93, 102]. Fourth, our generalization also shows that it is possible to impose additional structure such as self-concordant barrier to develop path-following scheme for solving a subclass of the composite convex minimization problems of the form (2.7). Fifth, our global inexact second-order oracle is defined via a weighted local norm and via a non-quadratic term and thus very different from the inexact first-order oracle from [28]. The global convergence result is independent of the self-concordance of f, and holds for a large class of functions, including Lipschitz gradient convex functions analyzed in [28]. Our inexact algorithm covers the inexact methods and quasi-Newton methods developed in [31, 40, 66, 104, 109] as special cases. Finally, we believe that our generalized self-concordant theory is not limited to convex optimization, but can be extended to solve convex-concave saddle-point problems, and monotone equations/inclusions involving *generalized self-concordant* functions, and our inexact oracle theory can be used to further develop other methods such as sub-sampled Newton-type methods rather than just the inexact proximal-Newton method in this thesis.

1.3 Outline of the thesis

The rest chapters are organized as follows.

- In Chapter 2, we provide some preliminary results and mathematical tools used in the entire thesis, including a brief overview of Newton method and its variant, the concept of standard self-concordance, [scaled] proximal operator, Fenchel conjugate, and Nesterov's smoothing technique.

- In Chapter 3, we introduce the class of *generalized self-concordant* functions, which covers standard self-concordant functions as special cases. Then, we establish several properties and key estimates of this function class, which can be used to design new numerical methods.

- In Chapter 4, we apply the theory introduced in Chapter 3 to develop several Newton-type methods for solving a class of smooth convex optimization problems involving the generalized self-concordant functions. We provide an explicit step-size for damped-step Newton-type scheme which can guarantee a global convergence without performing any globalization strategy. We also prove a local quadratic convergence of this method and its full-step variant without requiring the Lipschitz continuity of the objective Hessian. Then, we extend our result to develop proximal Newton-type methods for a class of composite convex minimization problems involving generalized self-concordant functions. We also achieve both local and global convergence without additional assumptions. Finally, we verify our theoretical results via several numerical examples, and compare them with existing methods.

- In Chapter 5, we introduce new global and local inexact second-order oracle concepts for a wide class of convex functions in composite convex optimization. We also provide examples to show that the class of convex functions equipped with the newly introduced inexact second-order oracles is larger than the standard self-concordant function class. Furthermore, we investigate several properties of convex and/or self-concordant functions under the inexact second-order oracles which are useful for algorithm development. Next, we apply our theory to develop inexact proximal Newton-type schemes for minimizing general composite convex problems equipped with such inexact oracles, with global convergence guarantees. When the first objective term is self-concordant, we establish different local convergence results for our method. We also apply our framework to derive a new primal-dual method for composite convex minimization problems. Finally, we provide some numerical examples to illustrate the benefit of our new algorithms based on this concept of inexact second-order oracles.

- In Chapter 6, we summarize the main conclusions of this thesis, and list several related research directions which remain on-going.

CHAPTER 2

Mathematical Tools and Preliminary Results

In this chapter, we briefly present necessary mathematical concepts and preliminary results which will be used in this thesis.

2.1 The Newton method

The *Newton method* is a fundamental scheme in optimization, which can be found in many numerical analysis textbooks such as [75]. For the following unconstrained minimization problem

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}), \tag{2.1}$$

where $f \in \mathcal{C}^2(\mathbb{R}^p)$, the Newton scheme refers to the following iteration step:

$$\mathbf{x}^{k+1} := \mathbf{x}^k - \nabla^2 f(\mathbf{x}^k)^{-1} \nabla f(\mathbf{x}^k). \tag{2.2}$$

Since the Hessian matrix may not always be positive definite, the global convergence behavior for the Newton iteration step (2.2) is not clear. To guarantee the global convergence, there are two common strategies: linesearch and trust-region. One can see [77, Chapter 3 and 4] for further references. In addition, the traditional Newton method has some drawbacks. For example, Assumption (a) is hard to verify in practice, since we have limited information about the optimal solution. Besides, Assumption (b) is strong in many applications. As mentioned in Section 1.1.3, the Newton method has many variant versions and is extremely important for modern applications in scientific computing. We briefly recall two variants in Section 2.4.

Let $\mathbf{x}^\star$ be the local minimum of the above optimization problem. Given the following assumptions

(a) $\nabla^2 f(\mathbf{x}^\star) \succeq l\mathbb{I}_p$ with some constant $l > 0$;

(b) $\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\|_2 \leq M\|\mathbf{x} - \mathbf{y}\|_2$ for all $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^p$;

(c) The initial point $\mathbf{x}^0$ is close enough to the optimal solution $\mathbf{x}^\star$, i.e.:

$$\|\mathbf{x}^0 - \mathbf{x}^\star\|_2 \leq \bar{r} := \frac{2l}{3M}.$$

Then we conclude that

(a) The iterative scheme (2.2) is well-defined;

(b) $\|\mathbf{x}^k - \mathbf{x}^\star\|_2 \leq \bar{r}$ for all $k \geq 0$;

(c) The sequence $\{\mathbf{x}^k\}$ converges quadratically to $\mathbf{x}^\star$, and the following relation holds:

$$\|\mathbf{x}^{k+1} - \mathbf{x}^\star\|_2 \leq \frac{M\|\mathbf{x}^k - \mathbf{x}^\star\|_2^2}{2(l - M\|\mathbf{x}^k - \mathbf{x}^\star\|_2)}.$$

2.2 Local norm and self-concordance

Let f be a $\mathcal{C}^3(\mathbb{R}^p)$ function. By the definition of third directional derivative, we have

$$\nabla^3 f(\mathbf{x})[\mathbf{u}] := \lim_{t \to 0} \frac{1}{t}[\nabla^2 f(\mathbf{x} + t\mathbf{u}) - \nabla^2 f(\mathbf{x})].$$

In this case, Assumption (b) of the Newton method in Subsection 2.1 becomes

$$|\left\langle \nabla^3 f(\mathbf{x})[\mathbf{u}]\mathbf{v}, \mathbf{v} \right\rangle| \leq M\|\mathbf{u}\|_2\|\mathbf{v}\|_2^2. \tag{2.3}$$

We note that, the Newton scheme is affine invariant w.r.t affine transformation of variables, while (2.3) is not. To maintain the affine invariant property, the following *local norm* and *self-concordant* concept are naturally introduced.

Local norm: Given a matrix $\mathbf{H} \in \mathcal{S}_{++}$, we define the weighted norm of $\mathbf{u} \in \mathbb{R}^p$ w.r.t $\mathbf{H}$ as $\|\mathbf{u}\|_{\mathbf{H}} := \langle \mathbf{H}\mathbf{u}, \mathbf{u} \rangle^{1/2}$. Its dual norm is $\|\mathbf{v}\|_{\mathbf{H}}^* = \left\langle \mathbf{H}^{-1}\mathbf{v}, \mathbf{v} \right\rangle^{1/2}$, which can be easily computed through definition. Especially, when $\mathbf{H} = \mathbb{I}$, the identity matrix, we have $\|\mathbf{u}\|_{\mathbf{H}} = \|\mathbf{u}\|_{\mathbf{H}}^* = \|\mathbf{u}\|_2$, the standard Euclidean norm. Let $f : \mathbb{R}^p \to \mathbb{R}$ be a three times continuously differentiable

function, i.e., $f(\mathbf{x}) \in \mathcal{C}^3(\mathrm{dom}(f))$. If $\nabla^2 f(\mathbf{x}) \succ 0$ at a given $\mathbf{x} \in \mathrm{dom}(f)$, then we define the *local norm* of $\mathbf{u}$ as $\|\mathbf{u}\|_{\nabla^2 f(\mathbf{x})}$, the weighted norm of $\mathbf{u}$ w.r.t $\nabla^2 f(\mathbf{x})$, shortly written as $\|\mathbf{u}\|_{\mathbf{x}}$ if the context is clear. The corresponding dual norm of $\mathbf{v} \in \mathbb{R}^p$, denoted by $\|\mathbf{v}\|_{\mathbf{x}}^*$, is defined as $\|\mathbf{v}\|_{\mathbf{x}}^* := \max\{\langle \mathbf{v}, \mathbf{u}\rangle \mid \|\mathbf{u}\|_{\mathbf{x}} \leq 1\} = \left\langle \nabla^2 f(\mathbf{x})^{-1}\mathbf{v}, \mathbf{v}\right\rangle^{1/2}$.

Self-concordant function: Following the standard definition of self-concordance [74], we call a function f *self-concordant* if the inequality

$$|\left\langle \nabla^3 f(\mathbf{x})[\mathbf{u}]\mathbf{u}, \mathbf{u}\right\rangle| \leq M_f \|\mathbf{u}\|_{\mathbf{x}}^3$$

holds for any $\mathbf{x} \in \mathrm{dom}(f)$ and $\mathbf{u} \in \mathbb{R}^p$ with some constant $M_f \geq 0$. If $M_f = 2$, the function f is called *standard self-concordant*. One simple example of standard self-concordant function is $f(x) = -\ln(x)$, where $x \in \mathbb{R}_+$.

To apply the Newton-type method, sometimes we need to transfer the constrained problem into unconstrained problem. Then we need to add a barrier function on the objective to prevent the variable running close to the boundary in the iterative scheme. If the barrier shares some properties related to self-concordance, then we call it *self-concordant barrier*. The standard definition is given as follows.

Self-concordant barrier: Let $B(\mathbf{x})$ be a standard self-concordant function. We call it a *ν-self-concordant barrier* for the set $\mathrm{Dom}(B) := \mathrm{cl}\ (\mathrm{dom}(B))$, if

$$\max_{\mathbf{u}\in\mathbb{R}^p}\{2\left\langle \nabla B(\mathbf{x}), \mathbf{u}\right\rangle - \left\langle \nabla^2 B(\mathbf{x})\mathbf{u}, \mathbf{u}\right\rangle\} \leq \nu$$

for all $\mathbf{x} \in \mathrm{dom}(B)$. The value ν is called the parameter of the barrier.

We will use the properties of self-concordant function when developing the inexact oracle theory in Chapter 5, and also generalize the self-concordance concepts together with corresponding properties, convergence theory, and algorithms in Chapters 3 and 4.

2.3 Proximal operator

Proximal operator: The *proximal operator* was first introduced in the early 1960s work by Moreau [69]. It is frequently used in optimization algorithms associated with nonsmooth optimization problems. It shows its popularity along with many well-known optimization methods such as proximal Newton methods (which will be discussed in Subsection 2.4) and proximal gradient methods. The proximal operator $\text{prox}_f : \mathbb{R}^p \to \mathbb{R}^p$ of f is defined by

$$\text{prox}_f(\mathbf{v}) := \arg\min_{\mathbf{x}} \left\{ f(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{v}\|_2^2 \right\}. \tag{2.4}$$

Two commonly used examples are as follows:

(a) *Indicator function:* If f is the indicator function of a closed convex set C, then prox_f is the projection onto C:

$$\text{prox}_f(\mathbf{v}) = \arg\min_{\mathbf{x} \in C} \|\mathbf{x} - \mathbf{v}\|_2^2 = P_C(\mathbf{v}).$$

(b) *The ℓ_1-norm:* If $f(\mathbf{x}) := \|\mathbf{x}\|_1$, then prox_f is the *soft-thresholding* operator:

$$\text{prox}_f(\mathbf{v})_i = \text{sign}(v_i) \max\{|v_i| - 1, 0\}.$$

Because of (a), the proximal operator can be viewed as an extension of the projection onto convex sets.

The proximity operator enjoys many properties of the projection, in particular it is firmly nonexpansive:

$$\|\text{prox}_f(\mathbf{x}) - \text{prox}_f(\mathbf{y})\|_2^2 \leq \left\langle \mathbf{x} - \mathbf{y}, \text{prox}_f(\mathbf{x}) - \text{prox}_f(\mathbf{y}) \right\rangle, \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^p, \tag{2.5}$$

which will be used to prove the convergence results in Chapters 4 and 5.

Scaled proximal operator: Since we use local norms in the definition of self-concordance and generalized self-concordance in the next chapters, we also introduce the *scaled proximal operator* together as follows. Given a matrix $\mathbf{H} \in \mathcal{S}_{++}^p$, we define a scaled proximal operator of

g in (2.7) as

$$\text{prox}_{\mathbf{H}^{-1}g}(\mathbf{x}) := \arg\min_{\mathbf{z}} \big\{ g(\mathbf{z}) + \tfrac{1}{2}\|\mathbf{z}-\mathbf{x}\|_{\mathbf{H}}^2 \big\}. \tag{2.6}$$

If $\mathbf{H} := H(\mathbf{y}) \in \mathcal{S}_{++}^p$, then we denote the above operator $\text{prox}_{\mathbf{H}^{-1}g}(\mathbf{x})$ as $\mathcal{P}_{\mathbf{y}}(\mathbf{x})$. We will use this notation in Chapter 5. Using the optimality condition of the minimization problem in (2.6), we can show that

$$\mathbf{y} = \text{prox}_{\mathbf{H}^{-1}g}(\mathbf{x}) \iff 0 \in \mathbf{H}(\mathbf{y}-\mathbf{x}) + \partial g(\mathbf{y}) \iff \mathbf{x} \in \mathbf{y} + \mathbf{H}^{-1}\partial g(\mathbf{y}) \equiv (\mathbb{I} + \mathbf{H}^{-1}\partial g)(\mathbf{y}).$$

Since g is proper, closed, and convex, $\text{prox}_{\mathbf{H}^{-1}g}$ is well-defined and single-valued. In particular, if we take $\mathbf{H} = \mathbb{I}$, the identity matrix, then $\text{prox}_{\mathbf{H}^{-1}g}(\cdot) = \text{prox}_g(\cdot)$, the standard proximal operator of g. If we can efficiently compute $\text{prox}_{\mathbf{H}^{-1}g}(\cdot)$ by a closed form or by polynomial time algorithms, then we say that g is *proximally tractable*. There exist several convex functions whose proximal operator is tractable. Examples such as ℓ_1-norm, coordinate-wise separable convex functions, and the indicator of simple convex sets can be found in the literature including [3, 38, 84].

2.4 Two variants of the Newton method

Proximal Newton method: The proximal Newton method was developed in early 1990s, see, e.g, [10, 91], which is known as the generalized Newton method. But this method is recently popularized in [52, 60, 102]. Proximal algorithms can be viewed as an analogous tool for nonsmooth, constrained, large-scale, or distributed versions of the unconstrained optimization described in Section 2.1, and have plenty of interesting interpretations and are connected to many different topics in optimization and applied mathematics. Many surveys written on various aspects of this topic over the years can be found easily, such as [22, 62, 84], and even for nonconvex optimization [78]. However, in this section we just recall the basic algorithm scheme and convergence results as follows.

We consider the composite minimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^p} \{F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}, \tag{2.7}$$

where f is a convex, continuously differentiable loss function, and g is a convex and proximal tractable, but not necessarily differentiable penalty function or regularizer. Such problems include the LASSO [98], the graphical LASSO [39], and trace-norm matrix completion [15]. The proximal Newton scheme refers to the following iterative scheme:

$$\begin{cases} \mathbf{z}^k & := \arg\min_{\mathbf{x}\in\mathrm{dom}(g)} \left\{ \left\langle \nabla f(\mathbf{x}^k), \mathbf{x}-\mathbf{x}^k \right\rangle + \frac{1}{2}\left\langle \nabla^2 f(\mathbf{x}^k)(\mathbf{x}-\mathbf{x}^k), \mathbf{x}-\mathbf{x}^k \right\rangle + g(\mathbf{x}) \right\} \\ & = \mathrm{prox}_{\nabla^2 f(\mathbf{x}^k)^{-1} g}\Big(\mathbf{x}^k - \nabla^2 f(\mathbf{x}^k)^{-1}\nabla f(\mathbf{x}^k)\Big). \\ \mathbf{x}^{k+1} & := \mathbf{x}^k + \tau_k(\mathbf{z}^k - \mathbf{x}^k), \end{cases} \tag{2.8}$$

where τ_k is the step-size determined by the backtracking linesearch in this section.

Given the following assumptions

(a) $m\mathbb{I} \preceq \nabla^2 f \preceq L\mathbb{I}$;

(b) $\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\|_2 \leq M\|\mathbf{x}-\mathbf{y}\|_2$ for all $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^p$.

Then we have the conclusion that

(a) The proximal Newton method (2.8) converges globally.

(b) Further more, it achieves local quadratic convergence rate:

$$\|\mathbf{x}^{k+1} - \mathbf{x}^\star\|_2 \leq \frac{M}{2m}\|\mathbf{x}^k - \mathbf{x}^\star\|_2^2.$$

Similar to the Newton method, the assumptions required here are too strong in practice. Therefore, we will continue studying this formulation in detail in Chapters 4 and 5, by reducing the above assumptions and giving an explicit step-size using [generalized] self-concordant theory.

Quasi-Newton method: The first quasi-Newton algorithm (DFP formula) was proposed by William C. Davidon, but it was soon superseded by its dual - the BFGS formula. Currently the most common quasi-Newton algorithms are the SR1 formula (for symmetric rank-one), the widespread BFGS method (suggested independently by Broyden, Fletcher, Goldfarb, and Shanno, in 1970), and its low-memory extension L-BFGS. The Broyden's class is a linear combination of the DFP and BFGS methods. We refer the reader to [77, Chapter 6] for a systematic review.

In quasi-Newton methods, like steepest descent, only the gradient of the objective function is required at each iteration. By measuring the Hessian mapping via secant equations in a proper way, it can achieve a local superlinear convergence, while avoiding the computation of the inverse Hessian matrix. In particular, we recall two important results as below.

Consider again the unconstrained convex minimization problem (2.1), the quasi-Newton scheme refers to the following iterative scheme:

$$\mathbf{x}^{k+1} := \mathbf{x}^k - \alpha_k \mathbf{B}_k^{-1} \nabla f(\mathbf{x}^k), \tag{2.9}$$

where $\mathbf{B}_k$ is a sequence of nonsingular matrices constructed in a certain way.

Theorem 2.4.1. [26, Dennis-Moré] Let $\mathbf{x}^\star$ be an optimal solution for (2.1), and $\nabla^2 f(\mathbf{x}^\star) \succ 0$. Let $\mathbf{E}_k := \mathbf{B}_k - \nabla^2 f(\mathbf{x}^\star)$. Assume that the sequence $\{\mathbf{x}^k\}$ generated by (2.9) converges to $\mathbf{x}^\star$. Then, $\{\mathbf{x}^k\}$ converges to $\mathbf{x}^\star$ superlinearly if and only if

$$\lim_{k\to\infty} \frac{\|\mathbf{E}_k(\mathbf{x}^{k+1} - \mathbf{x}^k)\|}{\|\mathbf{x}^{k+1} - \mathbf{x}^k\|} = 0,$$

where $\|\cdot\|$ represents an arbitrary vector norm in $\mathbb{R}^p$.

The above theorem characterizes the conditions for quasi-Newton method to achieve a superlinear convergence rate. While the next theorem provides a more general rule for its inexact updating version.

Given a starting point $\mathbf{x}^0$ and a sequence of positive scalars $\{\eta_k\}$, we update $\mathbf{x}^{k+1}$ following the condition

$$\|\nabla f(\mathbf{x}^k) + \nabla^2 f(\mathbf{x}^k)(\mathbf{x}^{k+1} - \mathbf{x}^k)\| \le \eta_k \|\nabla f(\mathbf{x}^k)\|, \tag{2.10}$$

where $\|\cdot\|$ represents an arbitrary vector norm in $\mathbb{R}^p$.

Theorem 2.4.2. [24, Theorem 3.4, Corollary 3.5] Let $\mathbf{x}^\star$ be an optimal solution of (2.1), and $\nabla^2 f(\mathbf{x}^\star) \succ 0$. Let $\eta_k \to 0$. If the sequence $\{\mathbf{x}^k\}$ generated from (2.10) converges to $\mathbf{x}^\star$, Then it converges to $\mathbf{x}^\star$ superlinearly.

Remark 2.4.1. The proximal Newton-type methods are developed recently to solve the convex composite problem (2.7), which combined both methods above together. One can see [20,

40, 60] for further references. In this thesis, we will consider the inexact scheme (2.10) as a subproblem by (1) specifying a certain [local] norm and different sequences of η_k; (2) replacing the derivatives with approximations in the inexact oracle; and (3) in the composite setting (2.7). Both convergence analysis and algorithm schemes can be found in Chapter 5.

2.5 Fenchel conjugates

Sometimes solving an optimization problem in its dual space is more convenient than its primal setting. When forming the dual problem, the *Fenchel conjugate* is frequently used as an expression of a maximization problem related to the original objective or constraint (when forming the Lagrangian function, see Chapter 5). The Fenchel conjugate, also known as convex conjugate was first introduced by Fenchel [35]. The *conjugate function* f^* of f in X, is defined in its dual space X^*, as follows

$$f^*(\mathbf{y}) := \sup\{\langle \mathbf{y}, \mathbf{x}\rangle - f(\mathbf{x}) \mid \mathbf{x} \in X\}. \tag{2.11}$$

By definition, the conjugate function f^* is always convex, and shares an important inequality that links the conjugate and original functions together:

$$\langle \mathbf{x}, \mathbf{y}\rangle \leq f(\mathbf{x}) + f^*(\mathbf{y}), \text{ for all } \mathbf{x} \in X, \text{ and } \mathbf{y} \in X^*.$$

Here are some commonly used examples:

(a) *Affine functions:* The conjugate of an affine function $f(\mathbf{x}) := \langle \mathbf{a}, \mathbf{x}\rangle - b$ is

$$f^*(\mathbf{y}) := \begin{cases} b & \text{if } \mathbf{y} = \mathbf{a} \\ +\infty & \text{otherwise.} \end{cases}$$

(b) *Absolute value:* The conjugate of the absolute value $f(x) := |x|$ is the indicator of the closed interval $[-1, 1]$:

$$f^*(y) := \begin{cases} 0 & \text{if } |y| \leq 1 \\ +\infty & \text{otherwise.} \end{cases}$$

(c) *Exponential function:* The conjugate of the exponential function $f(x) := e^x$ is

$$f^*(y) := \begin{cases} y\ln(y) - y, & x > 0 \\ 0, & x = 0 \\ +\infty & x < 0. \end{cases}$$

2.6 Nesterov's smoothing techniques

Nonsmooth convex functions or models appear frequently in practice. However, the optimization methods for smooth functions are more efficient and well-developed than the methods of nonsmooth ones. Therefore there is a need of good smoothing techniques that help us deal with nonsmooth functions.

Nesterov's smoothing techniques refer to an efficient approach for constructing efficient schemes for nonsmooth convex optimization, introduced by Nesterov [71]. Historically, the first numerical schemes for nonsmooth convex minimization were subgradient methods [88], with time complexity $\mathcal{O}(\varepsilon^{-2})$, where ε is the desired absolute accuracy of the approximate solution measured by the function value. For the black-box model of the objective function, it was shown that this efficiency of the simplest subgradient method cannot be improved uniformly in dimension of variables [53]. However, we never meet a pure black box model in practice. Motivated by this, Nesterov introduced this smoothing technique which makes a proper use of the structure of the problem, with time complexity improved to $\mathcal{O}(\varepsilon^{-1})$.

In detail, given a proper, closed, possibly nonsmooth, and convex function $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$, one can smooth f using the following Nesterov's smoothing technique

$$f_\gamma(\mathbf{x}) := \max_{\mathbf{u} \in \mathrm{dom}(f^*)} \{\langle \mathbf{A}\mathbf{x}, \mathbf{u}\rangle - f^*(\mathbf{u}) - \gamma\omega(\mathbf{u})\}, \tag{2.12}$$

where f^* is the Fenchel conjugate of f, $\omega : \mathrm{dom}(\omega) \subseteq \mathbb{R}^p \to \mathbb{R}$ is a continuous and σ-strongly convex function (prox-function) such that $\mathrm{dom}(f^*) \subseteq \mathrm{dom}(\omega)$, $\mathbf{A}$ is a linear operator, and $\gamma > 0$ is called the smoothness parameter. Without loss of generality, we assume that $\omega(\mathbf{u}_0) = 0$, where $\mathbf{u}_0 := \arg\min\{\omega(\mathbf{u}) : \mathbf{u} \in \mathrm{dom}(f^*)\}$ (prox-center). Then we have the following theorem:

Theorem 2.6.1. The function f_γ is well-defined and continuously differentiable at any $\mathbf{x} \in \mathrm{dom}(f)$. Moreover, this function is convex and its gradient $\nabla f_\gamma(\mathbf{x}) = \mathbf{A}^\top \mathbf{u}_\gamma^\star$ is Lipschitz continuous with constant $L_\gamma = \frac{1}{\gamma\sigma}\|\mathbf{A}\|_{1,2}^2$, where $\mathbf{u}_\gamma$ is the optimal solution of (2.12), and

$$\|\mathbf{A}\|_{1,2} := \max_{\mathbf{x},\mathbf{u}} \{\langle \mathbf{A}\mathbf{x}, \mathbf{u}\rangle : \|\mathbf{x}\|_1 = 1, \|\mathbf{u}\|_2 = 1\}.$$

For examples and the optimal scheme for smooth optimization, we refer the reader to [71, Section 3,4]. In this thesis, we combine this smoothing technique by choosing a proper smoothing function ω to build a new approximation of nonsmooth convex functions, which enjoys the *generalized self-concordant* properties. Details and examples are provided in Section 3.6.

CHAPTER 3

Theory of *generalized self-concordant* functions

3.1 Introduction

In this chapter we develop our generalized self-concordance theory. On the one hand, it is a generalization of the well-known self-concordance notion developed in [75]. On the other hand, it also covers the work in [1, 29, 109] as specific examples. Several specific applications and extensions of self-concordance notion can also be found in the literature including [49, 57, 80, 86].

The rest of this chapter is organized as follows. Section 3.2 develops fundamental concepts and examples of *generalized self-concordant* functions. Section 3.3 gives the foundation theory including some basic properties. Section 3.4 shows the relationship between *generalized self-concordant* and special function structures. Section 3.5 highlights the property of generalized self-concordance in conjugate form. Section 3.6 introduces *generalized self-concordant* approximation of nonsmooth functions. Section 3.7 provides the key bounds for Hessian, gradient and function values of *generalized self-concordant* functions, which will be used to develop our main theory in next chapter.

3.2 Fundamental concepts and examples

We introduce the fundamental concepts and motivating examples of *generalized self-concordant* functions in this section.

3.2.1 Univariate *generalized self-concordant* functions

Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a three times continuously differentiable function on the open domain $\mathrm{dom}(\varphi)$. Then, we write $\varphi \in \mathcal{C}^3\left(\mathrm{dom}(\varphi)\right)$. In this case, φ is convex if and only if $\varphi''(t) \geq 0$ for all $t \in \mathrm{dom}(\varphi)$. We introduce the following definition.

Definition 3.2.1. *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a $\mathcal{C}^3\left(dom(\varphi)\right)$ and univariate function with open domain $dom(\varphi)$, and $\nu > 0$ and $M_\varphi \geq 0$ be two constants. We say that φ is (M_φ, ν)-generalized self-concordant if*

$$|\varphi'''(t)| \leq M_\varphi \varphi''(t)^{\frac{\nu}{2}}, \quad \forall t \in dom(\varphi). \tag{3.1}$$

We denote this class of functions by $\widetilde{\mathcal{F}}_{M_\varphi,\nu}(\mathrm{dom}(\varphi))$ (shortly, $\widetilde{\mathcal{F}}_{M_\varphi,\nu}$ when $\mathrm{dom}(\varphi)$ is explicitly defined).

The inequality (3.1) also indicates that $\varphi''(t) \geq 0$ for all $t \in \mathrm{dom}(f)$. Hence, φ is convex. Clearly, if $\varphi(t) = \frac{a}{2}t^2 + bt$ for any constants $a \geq 0$ and $b \in \mathbb{R}$, we have $\varphi''(t) = a$ and $\varphi'''(t) = 0$. The inequality (3.1) is automatically satisfied for any $\nu > 0$ and $M_\varphi \geq 0$. The smallest value of M_φ is zero. Hence, any convex quadratic function belongs to $\widetilde{\mathcal{F}}_{0,\nu}$ for any $\nu > 0$. While (3.1) holds for any other constant $\hat{M}_\varphi \geq M_\varphi$, we often require that M_φ is the smallest constant satisfying (3.1).

Example 3.1. Let us now provide some common examples satisfying Definition 3.2.1.

(a) *Standard self-concordant functions:* If we choose $\nu = 3$, then (3.1) becomes $|\varphi'''(t)| \leq M_\varphi \varphi''(t)^{3/2}$, which is the standard self-concordant functions in $\mathbb{R}$ introduced in [75].

(b) *Logistic functions:* In [1], Bach modified the standard self-concordant inequality in [75] to obtain $|\varphi'''(t)| \leq M_\varphi \varphi''(t)$, and showed that the well-known logistic loss $\varphi(t) := \ln(1+e^{-t})$ satisfies this definition. In [103] the authors also exploited this definition, and developed a class of first-order and second-order methods to solve composite convex minimization problems. Hence, $\varphi(t) := \ln(1 + e^{-t}) \in \widetilde{\mathcal{F}}_{1,2}$.

(c) *Exponential functions:* The exponential function $\varphi(t) := e^{-t} \in \widetilde{\mathcal{F}}_{1,2}$. This function is often used, e.g., in Ada-boost [59], or in matrix scaling [21].

(d) *Distance-weighted discrimination* (DWD): We consider a more general function $\varphi(t) := \frac{1}{t^q}$ on $\mathbb{R}_{++}$ and $q \geq 1$ studied in [67] for DWD using in support vector machine. As shown in Table 3.1, $\varphi \in \widetilde{\mathcal{F}}_{M_\varphi,\nu}$ with $M_\varphi = \frac{q+2}{\sqrt[(q+2)]{q(q+1)}}$ and $\nu = \frac{2(q+3)}{q+2} \in (2, 3)$.

(e) *Entropy function:* We consider the well-known entropy function $\varphi(t) := t\ln(t)$ for $t > 0$. We can easily show that $|\varphi'''(t)| = \frac{1}{t^2} = \varphi''(t)^2$. Hence $\varphi \in \widetilde{\mathcal{F}}_{1,4}$.

(f) *Arcsine distribution:* Consider the function $\varphi(t) := \frac{1}{\sqrt{1-t^2}}$ for $t \in (-1, 1)$. This function is convex and smooth. Moreover, we verify that $\varphi \in \widetilde{\mathcal{F}}_{M_\varphi,\nu}$ with $\nu = \frac{14}{5} \in (2, 3)$ and $M_\varphi = \frac{3\sqrt{495-105\sqrt{21}}}{(7-\sqrt{21})^{7/5}} < 3.25$. We can generalize this function to $\varphi(t) := [(t-a)(b-t)]^{-q}$ for $t \in (a, b)$, where $a < b$ and $q > 0$. Then, we can show that $\nu = \frac{2(q+3)}{q+2} \in (2, 3)$.

(g) *Robust Regression:* Consider a monomial function $\varphi(t) := t^q$ for $q \in (1, 2)$ studied in [107] for robust regression using in statistics. Then $\varphi \in \widetilde{\mathcal{F}}_{M_\varphi,\nu}$ with $M_\varphi = \frac{2-q}{\sqrt[(2-q)]{q(q-1)}}$ and $\nu = \frac{2(3-q)}{2-q} \in (4, +\infty)$.

$\diamond$

As concrete examples, the following table, Table 3.1, provides a non-exhaustive list of *generalized self-concordant* (gsc) functions used in the literature.

Table 3.1: Examples of univariate gsc functions ($\mathcal{F}_L^1$ means that $\nabla\varphi$ is Lipschitz continuous).

Function	Form of $\varphi(t)$	ν	M	dom(φ)	Application	$\mathcal{F}_L^1$	References
Log-barrier	$-\ln(t)$	3	2	$\mathbb{R}_{++}$	Poisson	no	[13, 74, 75]
Entropy-barrier	$t\ln(t) - \ln(t)$	3	2	$\mathbb{R}_{++}$	Interior-point	no	[74]
Logistic	$\ln(1 + e^t)$	2	1	$\mathbb{R}$	Classification	yes	[50]
Exponential	e^{-t}	2	1	$\mathbb{R}$	AdaBoost	no	[21, 59]
Negative power	t^{-q}, $(q > 0)$	$\frac{2(q+3)}{q+2}$	$\frac{q+2}{\sqrt[(q+2)]{q(q+1)}}$	$\mathbb{R}_{++}$	DWD	no	[67]
Arcsine distribution	$\frac{1}{\sqrt{1-t^2}}$	$\frac{14}{5}$	< 3.25	$(-1, 1)$	Random walks	no	[42]
Positive power	t^q, $(q \in (1, 2))$	$\frac{2(3-q)}{2-q}$	$\frac{2-q}{\sqrt[(2-q)]{q(q-1)}}$	$\mathbb{R}_+$	Regression	no	[107]
Entropy	$t\ln(t)$	4	1	$\mathbb{R}_+$	KL divergence	no	[1]

Remark 3.2.1. All examples given in Table 3.1 fall into the case $\nu \geq 2$. However, we note that Definition 3.2.1 also covers [109, Lemma 1] as a special case when $\nu \in (0, 2)$. Unfortunately, as we will see in what follows, it is unclear how to generalize several properties of generalized self-concordance from univariate to multivariable functions for $\nu \in (0, 2)$, except for strongly convex functions.

Table 3.1 only provides common *generalized self-concordant* functions using in practice. However, it is possible to combine these functions to obtain mixture functions that preserve the *generalized self-concordant* inequality given in Definition 3.2.1. For instance, the barrier entropy $t\ln(t) - \ln(t)$ is a standard self-concordant function, and it is the sum of the entropy

$t\ln(t)$ and the negative logarithmic function $-\ln(t)$, which are *generalized self-concordant* with $\nu = 4$ and $\nu = 3$, respectively.

3.2.2 Multivariate *generalized self-concordant* functions

Let $f : \mathbb{R}^p \to \mathbb{R}$ be a $\mathcal{C}^3(\mathrm{dom}(f))$ smooth and convex function with open domain $\mathrm{dom}(f)$. Given $\nabla^2 f$ the Hessian of f, $\mathbf{x} \in \mathrm{dom}(f)$, and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$, we consider the function $\psi(t) := \left\langle \nabla^2 f(\mathbf{x} + t\mathbf{v})\mathbf{u}, \mathbf{u} \right\rangle$. Then, it is obvious to show that

$$\psi'(t) := \left\langle \nabla^3 f(\mathbf{x} + t\mathbf{v})[\mathbf{v}]\mathbf{u}, \mathbf{u} \right\rangle.$$

for $t \in \mathbb{R}$ such that $\mathbf{x} + t\mathbf{v} \in \mathrm{dom}(f)$, where $\nabla^3 f$ is the third-order derivative of f. It is clear that $\psi(0) = \left\langle \nabla^2 f(\mathbf{x})\mathbf{u}, \mathbf{u} \right\rangle = \|\mathbf{u}\|_{\mathbf{x}}^2$. By using the local norm, we generalize Definition 3.2.1 to multivariate functions $f : \mathbb{R}^p \to \mathbb{R}$ as follows.

Definition 3.1. A $\mathcal{C}^3$-convex function $f : \mathbb{R}^p \to \mathbb{R}$ is said to be an (M_f, ν)-generalized self-concordant function of the order $\nu > 0$ and the constant $M_f \geq 0$ if, for any $\mathbf{x} \in \mathrm{dom}(f)$ and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$, it holds

$$| \left\langle \nabla^3 f(\mathbf{x})[\mathbf{v}]\mathbf{u}, \mathbf{u} \right\rangle | \leq M_f \|\mathbf{u}\|_{\mathbf{x}}^2 \|\mathbf{v}\|_{\mathbf{x}}^{\nu-2} \|\mathbf{v}\|_2^{3-\nu}. \tag{3.2}$$

Here, we use a convention that $\frac{0}{0} = 0$ for the case $\nu < 2$ or $\nu > 3$. We also adopt the previous univariate *generalized self-concordant* notation $\widetilde{\mathcal{F}}_{M_f,\nu}(\mathrm{dom}(f))$ (shortly, $\widetilde{\mathcal{F}}^p_{M_f,\nu}$ or $\widetilde{\mathcal{F}}_{M_f,\nu}$ when $\mathrm{dom}(f)$ is explicitly defined) to denote this class of functions.

Let us consider the following two extreme cases:

1. If $\nu = 2$, (3.2) leads to $| \left\langle \nabla^3 f(\mathbf{x})[\mathbf{v}]\mathbf{u}, \mathbf{u} \right\rangle | \leq M_f \|\mathbf{u}\|_{\mathbf{x}}^2 \|\mathbf{v}\|_2$, which collapses to the definition introduced in [1] by letting $\mathbf{u} = \mathbf{v}$.

2. If $\nu = 3$ and $\mathbf{u} = \mathbf{v}$, (3.2) reduces to $| \left\langle \nabla^3 f(\mathbf{x})[\mathbf{u}]\mathbf{u}, \mathbf{u} \right\rangle | \leq M_f \|\mathbf{u}\|_{\mathbf{x}}^3$, Definition 3.1 becomes the standard self-concordant definition introduced in [74, 75].

We emphasize that Definition 3.1 is not symmetric, but can avoid the use of multilinear mappings as required in [1, 75]. However, by [75, Proposition 9.1.1] or [74, Lemma 4.1.2], Definition 3.1 with $\nu = 3$ is equivalent to [74, Definition 4.1.1] for standard self-concordant functions.

3.3 Basic properties of *generalized self-concordant* functions

We first show that if f_1 and f_2 are two *generalized self-concordant* functions, then $\beta_1 f_1 + \beta_2 f_2$ is also *generalized self-concordant* for any $\beta_1, \beta_2 > 0$ according to Definition 3.1.

Proposition 3.3.1 (Sum of *generalized self-concordant* functions)**.** Let $f_i \in \widetilde{\mathcal{F}}_{M_{f_i},\nu}$ satisfying (3.2), where $M_{f_i} \geq 0$ and $\nu \geq 2$ for $i = 1, \ldots, m$. Then, for $\beta_i > 0$, $i = 1, 2, \ldots, m$, the function $f(\mathbf{x}) := \sum_{i=1}^m \beta_i f_i(\mathbf{x})$ is well-defined on $\mathrm{dom}(f) = \bigcap_{i=1}^m \mathrm{dom}(f_i)$, and $f \in \widetilde{\mathcal{F}}_{M_f,\nu}$ with the same order $\nu \geq 2$ and the constant

$$M_f := \max\{\beta_i^{1-\frac{\nu}{2}} M_{f_i} \mid 1 \leq i \leq m\} \geq 0.$$

Proof. It is sufficient to prove for $m = 2$. For $m > 2$, it follows from $m = 2$ by induction. By [74, Theorem 3.1.5], f is a closed and convex function. In addition, $\mathrm{dom}(f) = \mathrm{dom}(f_1) \cap \mathrm{dom}(f_2)$. Let us fix some $\mathbf{x} \in \mathrm{dom}(f)$ and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$. Then, by Definition 3.1, we have

$$|\left\langle \nabla^3 f_i(\mathbf{x})[\mathbf{v}]\mathbf{u}, \mathbf{u}\right\rangle| \leq M_{f_i} \left\langle \nabla^2 f_i(\mathbf{x})\mathbf{u}, \mathbf{u}\right\rangle \left\langle \nabla^2 f_i(\mathbf{x})\mathbf{v}, \mathbf{v}\right\rangle^{\frac{\nu-2}{2}} \|\mathbf{v}\|_2^{3-\nu}, \quad i = 1, 2.$$

Denote $w_i := \left\langle \nabla^2 f_i(\mathbf{x})\mathbf{u}, \mathbf{u}\right\rangle \geq 0$ and $s_i := \left\langle \nabla^2 f_i(\mathbf{x})\mathbf{v}, \mathbf{v}\right\rangle \geq 0$ for $i = 1, 2$. We can derive

$$\begin{aligned}
\frac{|\left\langle \nabla^3 f(\mathbf{x})[\mathbf{v}]\mathbf{u}, \mathbf{u}\right\rangle|}{\left\langle \nabla^2 f(\mathbf{x})\mathbf{u}, \mathbf{u}\right\rangle \left\langle \nabla^2 f(\mathbf{x})\mathbf{v}, \mathbf{v}\right\rangle^{\frac{\nu-2}{2}}} &\leq \frac{\beta_1|\left\langle \nabla^3 f_1(\mathbf{x})[\mathbf{v}]\mathbf{u}, \mathbf{u}\right\rangle| + \beta_2|\left\langle \nabla^3 f_2(\mathbf{x})[\mathbf{v}]\mathbf{u}, \mathbf{u}\right\rangle|}{\left\langle \nabla^2 f(\mathbf{x})\mathbf{u}, \mathbf{u}\right\rangle \left\langle \nabla^2 f(\mathbf{x})\mathbf{v}, \mathbf{v}\right\rangle^{\frac{\nu-2}{2}}} \\
&\leq \left[\frac{M_{f_1}\beta_1 w_1 s_1^{\frac{\nu-2}{2}} + M_{f_2}\beta_2 w_2 s_2^{\frac{\nu-2}{2}}}{(\beta_1 w_1 + \beta_2 w_2)(\beta_1 s_1 + \beta_2 s_2)^{\frac{\nu-2}{2}}}\right]_{[T]} \|\mathbf{v}\|_2^{3-\nu}. \quad (3.3)
\end{aligned}$$

Let $\xi := \frac{\beta_1 w_1}{\beta_1 w_1 + \beta_2 w_2} \in [0, 1]$ and $\eta := \frac{\beta_1 s_1}{\beta_1 s_1 + \beta_2 s_2} \in [0, 1]$. Then, $\frac{\beta_2 w_2}{\beta_1 w_1 + \beta_2 w_2} = 1 - \xi \geq 0$ and $\frac{\beta_2 s_2}{\beta_1 s_1 + \beta_2 s_2} = 1 - \eta \geq 0$. Hence, the term $[T]$ in the square brackets of (3.3) becomes

$$h(\xi, \eta) := \beta_1^{1-\frac{\nu}{2}} M_{f_1} \xi \eta^{\frac{\nu-2}{2}} + \beta_2^{1-\frac{\nu}{2}} M_{f_2}(1-\xi)(1-\eta)^{\frac{\nu-2}{2}}, \quad \xi, \eta \in [0, 1].$$

Since $\nu \geq 2$ and $\xi, \eta \in [0, 1]$, we can upper bound $h(\xi, \eta)$ as

$$h(\xi, \eta) \leq \beta_1^{1-\frac{\nu}{2}} M_{f_1} \xi + \beta_2^{1-\frac{\nu}{2}} M_{f_2}(1-\xi), \quad \forall \xi \in [0, 1].$$

The right-hand side function is linear in ξ on $[0,1]$. It achieves the maximum at its boundary. Hence, we have

$$\max_{\xi\in[0,1],\eta\in[0,1]} h(\xi,\eta) \le \max\{\beta_1^{1-\frac{\nu}{2}} M_{f_1}, \beta_2^{1-\frac{\nu}{2}} M_{f_2}\}.$$

Using this estimate into (3.3), we can show that $f(\cdot) := \beta_1 f_1(\cdot) + \beta_2 f_2(\cdot)$ is (M_f,ν)-*generalized self-concordant* with $M_f := \max\{\beta_1^{1-\frac{\nu}{2}} M_{f_1}, \beta_2^{1-\frac{\nu}{2}} M_{f_2}\}$. □

Using Proposition 3.3.1, we can also see that if $f \in \widetilde{\mathcal{F}}_{M_f,\nu}$ and $\beta > 0$, then $g(\mathbf{x}) := \beta f(\mathbf{x}) \in \widetilde{\mathcal{F}}_{M_g,\nu}$ with the constant $M_g := \beta^{1-\frac{\nu}{2}} M_f$. The convex quadratic function $q(\mathbf{x}) := \frac{1}{2}\langle \mathbf{Q}\mathbf{x},\mathbf{x}\rangle + \mathbf{c}^\top\mathbf{x}$ with $\mathbf{Q} \in \mathcal{S}^p_+$ belongs to $\widetilde{\mathcal{F}}_{0,\nu}$ for any $\nu > 0$. Hence, by Proposition 3.3.1, if $f \in \widetilde{\mathcal{F}}_{M_f,\nu}$, then $f(\mathbf{x}) + \frac{1}{2}\langle \mathbf{Q}\mathbf{x},\mathbf{x}\rangle + \mathbf{c}^\top\mathbf{x} \in \widetilde{\mathcal{F}}_{M_f,\nu}$.

Next, we consider an affine transformation of a *generalized self-concordant* function.

Proposition 3.3.2 (Affine transformation)**.** Let $\mathcal{A}(\mathbf{x}) := \mathbf{A}\mathbf{x} + \mathbf{b}$ be an affine transformation from $\mathbb{R}^p$ to $\mathbb{R}^q$, and $f \in \widetilde{\mathcal{F}}_{M_f,\nu}$ with $\nu > 0$. Then, the following statements hold:

(a) If $\nu \in (0,3]$, then $g(\mathbf{x}) := f(\mathcal{A}(\mathbf{x})) \in \widetilde{\mathcal{F}}_{M_g,\nu}$ with $M_g := M_f\|\mathbf{A}\|^{3-\nu}$.

(b) If $\nu > 3$ and $\lambda_{\min}(\mathbf{A}^\top\mathbf{A}) > 0$, then $g(\mathbf{x}) := f(\mathcal{A}(\mathbf{x})) \in \widetilde{\mathcal{F}}_{M_g,\nu}$ with $M_g := M_f\lambda_{\min}(\mathbf{A}^\top\mathbf{A})^{\frac{3-\nu}{2}}$, where $\lambda_{\min}(\mathbf{A}^\top\mathbf{A})$ is the smallest eigenvalue of $\mathbf{A}^\top\mathbf{A}$.

Proof. Since $g(\mathbf{x}) = f(\mathcal{A}(\mathbf{x})) = f(\mathbf{A}\mathbf{x}+\mathbf{b})$, it is easy to show that $\nabla^2 g(\mathbf{x}) = \mathbf{A}^\top\nabla^2 f(\mathcal{A}(\mathbf{x}))\mathbf{A}$ and $\nabla^3 g(\mathbf{x})[\mathbf{v}] = \mathbf{A}^\top(\nabla^3 f(\mathcal{A}(\mathbf{x})[\mathbf{A}\mathbf{v}])\mathbf{A}$. Let us denote by $\tilde{\mathbf{x}} := \mathbf{A}\mathbf{x}+\mathbf{b}$, $\tilde{\mathbf{u}} := \mathbf{A}\mathbf{u}$, and $\tilde{\mathbf{v}} := \mathbf{A}\mathbf{v}$. Then, using Definition 3.1, we have

$$\begin{aligned}
|\langle\nabla^3 g(\mathbf{x})[\mathbf{v}]\mathbf{u},\mathbf{u}\rangle| &= |\langle\mathbf{A}^\top(\nabla^3 f(\tilde{\mathbf{x}})[\tilde{\mathbf{v}}])\mathbf{A}\mathbf{u},\mathbf{u}\rangle| = |\langle\nabla^3 f(\tilde{\mathbf{x}})[\tilde{\mathbf{v}}]\tilde{\mathbf{u}},\tilde{\mathbf{u}}\rangle| \\
&\overset{(3.2)}{\le} M_f\langle\nabla^2 f(\tilde{\mathbf{x}})\tilde{\mathbf{u}},\tilde{\mathbf{u}}\rangle\langle\nabla^2 f(\tilde{\mathbf{x}})\tilde{\mathbf{v}},\tilde{\mathbf{v}}\rangle^{\frac{\nu}{2}-1}\|\tilde{\mathbf{v}}\|_2^{3-\nu} \\
&= M_f\langle\mathbf{A}^\top\nabla^2 f(\mathcal{A}(\mathbf{x}))\mathbf{A}\mathbf{u},\mathbf{u}\rangle\langle\mathbf{A}^\top\nabla^2 f(\mathcal{A}(\mathbf{x}))\mathbf{A}\mathbf{v},\mathbf{v}\rangle^{\frac{\nu}{2}-1}\|\mathbf{A}\mathbf{v}\|_2^{3-\nu} \\
&= M_f\langle\nabla^2 g(\mathbf{x})\mathbf{u},\mathbf{u}\rangle\langle\nabla^2 g(\mathbf{x})\mathbf{v},\mathbf{v}\rangle^{\frac{\nu}{2}-1}\|\mathbf{A}\mathbf{v}\|_2^{3-\nu}.
\end{aligned} \tag{3.4}$$

(a) If $\nu \in (0,3]$, then we have $\|\mathbf{A}\mathbf{v}\|_2^{3-\nu} \le \|\mathbf{A}\|^{3-\nu}\|\mathbf{v}\|_2^{3-\nu}$. Hence (3.4) implies

$$|\langle\nabla^3 g(\mathbf{x})[\mathbf{v}]\mathbf{u},\mathbf{u}\rangle| \le M_f\|\mathbf{A}\|^{3-\nu}\langle\nabla^2 g(\mathbf{x})\mathbf{u},\mathbf{u}\rangle\langle\nabla^2 g(\mathbf{x})\mathbf{v},\mathbf{v}\rangle^{\frac{\nu}{2}-1}\|\mathbf{v}\|_2^{3-\nu},$$

which shows that $g \in \widetilde{\mathcal{F}}_{M_g,\nu}$ with $M_g := M_f\|\mathbf{A}\|^{3-\nu}$.

(b) Note that $\|\mathbf{A}\mathbf{v}\|_2^2 = \mathbf{v}^\top\mathbf{A}^\top\mathbf{A}\mathbf{v} \geq \lambda_{\min}(\mathbf{A}^\top\mathbf{A})\|\mathbf{v}\|_2^2 \geq 0$, where $\lambda_{\min}(\mathbf{A}^\top\mathbf{A})$ is the smallest eigenvalue of $\mathbf{A}^\top\mathbf{A}$. If $\lambda_{\min}(\mathbf{A}^\top\mathbf{A}) > 0$ and $\nu > 3$, then we have $\|\mathbf{A}\mathbf{v}\|_2^{3-\nu} \leq \lambda_{\min}(\mathbf{A}^\top\mathbf{A})^{\frac{3-\nu}{2}}\|\mathbf{v}\|_2^{3-\nu}$. Combining this estimate and (3.4), we can show that $g \in \widetilde{\mathcal{F}}_{M_g,\nu}$ with $M_g := M_f\lambda_{\min}(\mathbf{A}^\top\mathbf{A})^{\frac{3-\nu}{2}}$. □

Remark 3.3.1. Proposition 3.3.2 shows that generalized self-concordance is preserved via an affine transformations if $\nu \in (0, 3]$. If $\nu > 3$, then it requires $\mathbf{A}$ to be over-completed, i.e., $\lambda_{\min}(\mathbf{A}^\top\mathbf{A}) > 0$. Hence, the theory developed in the sequel remains applicable for $\nu > 3$ if $\mathbf{A}$ is over-completed.

Combining Proposition 3.3.1 and 3.3.2(a), we have a corollary of $\widetilde{\mathcal{F}}_{M_f,\nu}$ class for the sum of functions from different dimensional spaces.

Corollary 3.3.3. Let $f_i \in \widetilde{\mathcal{F}}^{d_i}_{M_i,\nu}$ for $i = 1, \ldots, m$. If $\nu \in [2, 3]$, then $f(\mathbf{x}) := \sum_{i=1}^m f_i(\mathbf{x}_i)$ also belongs to $\widetilde{\mathcal{F}}^{p}_{M,\nu}$ with $p = \sum_{i=1}^m d_i$, and the same parameters as in Proposition 3.3.1.

Proof. Similar as the proof of Proposition 3.3.1, it is sufficient to prove for the case $m = 2$. Define $\tilde{f}_1(\mathbf{x}) := f_1([\mathbb{I}_{d_1}, 0]\mathbf{x}) = f_1(\mathbf{x}_1)$ and $\tilde{f}_2(\mathbf{x}) := f_2([0, \mathbb{I}_{d_2}]\mathbf{x}) = f_2(\mathbf{x}_2)$. Since $\|[\mathbb{I}_{d_1}, 0]\| = \|[0, \mathbb{I}_{d_2}]\| = 1$, by Proposition 3.3.2(a), $\tilde{f}_i \in \widetilde{\mathcal{F}}^{d_1+d_2}_{M_i,\nu}, i = 1, 2$. By Proposition 3.3.1, f also belongs to $\mathcal{G}^{d_1+d_2}(M, \nu)$ with the same parameters as in Proposition 3.3.1. □

The following result is an extension of standard self-concordant functions ($\nu = 3$), whose proof is very similar to [74, Theorems 4.1.3, 4.1.4] by replacing the parameters $M_f = 2$ and $\nu = 3$ with the general parameters $M_f \geq 0$ and $\nu > 0$ (or $\nu \geq 2$), respectively. We omit the detailed proof.

Proposition 3.3.4. Let $f \in \widetilde{\mathcal{F}}_{M_f,\nu}$ with $\nu > 0$. Then:

(a) If $\nu \geq 2$ and $\mathrm{dom}(f)$ contains no straight line, then $\nabla^2 f(\mathbf{x}) \succ 0$ for any $\mathbf{x} \in \mathrm{dom}(f)$.

(b) If there exists $\bar{\mathbf{x}} \in \mathrm{bd}(\mathrm{dom}(f))$, the boundary of $\mathrm{dom}(f)$, then, for any $\bar{\mathbf{x}} \in \mathrm{bd}(\mathrm{dom}(f))$, and any sequence $\{\mathbf{x}_k\} \subset \mathrm{dom}(f)$ such that $\lim_{k\to\infty} \mathbf{x}_k = \bar{\mathbf{x}}$, we have $\lim_{k\to\infty} f(\mathbf{x}_k) = +\infty$.

Note that Proposition 3.3.4(a) only holds for $\nu \geq 2$. If we consider $g(\mathbf{x}) := f(\mathcal{A}(\mathbf{x}))$ for a given affine operator $\mathcal{A}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$, then the non-degenerateness of $\nabla^2 g$ is only guaranteed if $\mathbf{A}$ is full-rank. Otherwise, it is non-degenerated in a given subspace of $\mathbf{A}$.

3.4 Generalized self-concordant functions with special structures

We first show that if a *generalized self-concordant* function is strongly convex or Lipschitz gradient, then we can increase or decrease the parameter ν if necessary. Particularly, the original $\widetilde{\mathcal{F}}_{M_f,\nu}$ class can be cast into the special case $\nu = 2$ or $\nu = 3$.

Proposition 3.4.1. Let $f \in \widetilde{\mathcal{F}}_{M,\nu}$ with $\nu \geq 2$. Then:

(a) If f is μ-strongly convex w.r.t ℓ_2-norm for some $\mu > 0$, then f also belongs to $\widetilde{\mathcal{F}}_{\tilde{M},\tilde{\nu}}$ class with $\tilde{M} := M/(\sqrt{\mu})^{\tilde{\nu}-\nu}$, given that $\nu \leq \tilde{\nu}$.

(b) If f has L-Lipschitz gradient w.r.t ℓ_2-norm, then f also belongs to $\widetilde{\mathcal{F}}_{\tilde{M},\tilde{\nu}}$ class with $\tilde{M} := M(\sqrt{L})^{\nu-\tilde{\nu}}$, given that $\tilde{\nu} \leq \nu$.

Proof. If f is μ-strongly convex, then $\mu\|\mathbf{v}\|_2^2 \leq \langle \nabla^2 f(\mathbf{x})\mathbf{v}, \mathbf{v}\rangle$, hence $\|\mathbf{v}\|_2 \leq \|\mathbf{v}\|_{\mathbf{x}}/\sqrt{\mu}$.

$$\begin{aligned} |\nabla^3 f(\mathbf{x})[\mathbf{v}]\mathbf{u}, \mathbf{u}| &\leq M\|\mathbf{u}\|_{\mathbf{x}}^2 \|\mathbf{v}\|_{\mathbf{x}}^{\nu-2} \|\mathbf{v}\|_2^{3-\nu} \\ &= M\left(\frac{\|\mathbf{v}\|_2}{\|\mathbf{v}\|_{\mathbf{x}}}\right)^{\tilde{\nu}-\nu} \|\mathbf{u}\|_{\mathbf{x}}^2 \|\mathbf{v}\|_{\mathbf{x}}^{\tilde{\nu}-2} \|\mathbf{v}\|_2^{3-\tilde{\nu}} \\ &\leq M/(\sqrt{\mu})^{\tilde{\nu}-\nu} \|\mathbf{u}\|_{\mathbf{x}}^2 \|\mathbf{v}\|_{\mathbf{x}}^{\tilde{\nu}-2} \|\mathbf{v}\|_2^{3-\tilde{\nu}}, \end{aligned}$$

where the first inequality is by definition, and the second is from the strongly convexity.
If f has L-Lipschitz gradient, then $\langle \nabla^2 f(\mathbf{x})\mathbf{v}, \mathbf{v}\rangle \leq L\|\mathbf{v}\|_2^2$, hence $\|\mathbf{v}\|_{\mathbf{x}} \leq \sqrt{L}\|\mathbf{v}\|_2$. Then

$$\begin{aligned} |\nabla^3 f(\mathbf{x})[\mathbf{v}]\mathbf{u}, \mathbf{u}| &\leq M\|\mathbf{u}\|_{\mathbf{x}}^2 \|\mathbf{v}\|_{\mathbf{x}}^{\nu-2} \|\mathbf{v}\|_2^{3-\nu} \\ &= M\left(\frac{\|\mathbf{v}\|_{\mathbf{x}}}{\|\mathbf{v}\|_2}\right)^{\nu-\tilde{\nu}} \|\mathbf{u}\|_{\mathbf{x}}^2 \|\mathbf{v}\|_{\mathbf{x}}^{\tilde{\nu}-2} \|\mathbf{v}\|_2^{3-\tilde{\nu}} \\ &\leq M(\sqrt{L})^{\nu-\tilde{\nu}} \|\mathbf{u}\|_{\mathbf{x}}^2 \|\mathbf{v}\|_{\mathbf{x}}^{\tilde{\nu}-2} \|\mathbf{v}\|_2^{3-\tilde{\nu}}, \end{aligned}$$

where the first inequality is by definition, and the second is from the Lipschitz property. □

Remark 3.4.1. If we take $\tilde{\nu} = 3$ and $\nu \leq 3$ in case (a), or $\tilde{\nu} = 2$ in case (b), then we get back to the special case shown in [97, Proposition 4].

Proposition 3.4.1 provides two important properties jointly linked with special function structures. If a *generalized self-concordant* function f is Lipschitz gradient, we can always classify it into the special case $\nu = 2$. Therefore, we can exploit both structures: generalized self-concordance and Lipschitz gradient to develop better algorithms. Combining Proposition 3.3.1 and 3.4.1, we have the following corollary:

Corollary 3.4.2. Let $g \in \widetilde{\mathcal{F}}^p_{M_g,\nu_g}$ and $h \in \widetilde{\mathcal{F}}^p_{M_h,\nu_h}$, and $f := \alpha g + \beta h$ be their sum, for $\alpha, \beta > 0$. Assume that $2 \leq \nu_g \leq \nu_h$, then

(a) If g is μ-strongly convex, then f also belongs to $\widetilde{\mathcal{F}}^p_{M_f,\nu_h}$ with

$$M_f := \max\{\alpha^{1-\frac{\nu_h}{2}} M_g/(\sqrt{\mu})^{\nu_h-\nu_g}, \beta^{1-\frac{\nu_h}{2}} M_h\}.$$

(b) If h has L-Lipschitz gradient, then f also belongs to $\widetilde{\mathcal{F}}^p_{M_f,\nu_g}$ with

$$M_f := \max\{\alpha^{1-\frac{\nu_g}{2}} M_g, \beta^{1-\frac{\nu_g}{2}} M_h(\sqrt{L})^{\nu_h-\nu_g}\}.$$

Combining Corollary 3.3.3 and Proposition 3.4.1, we have the following corollary:

Corollary 3.4.3. Let $g \in \widetilde{\mathcal{F}}^{d_1}_{M_g,\nu_g}$ and $h \in \widetilde{\mathcal{F}}^{d_2}_{M_h,\nu_h}$, and $f(\mathbf{x}) := \alpha g(\mathbf{x}_1) + \beta h(\mathbf{x}_2)$ be their sum, for $\alpha, \beta > 0$ and $\mathbf{x} := (\mathbf{x}_1^T, \mathbf{x}_2^T)^T \in \mathbb{R}^{d_1+d_2}$. Assume that $2 \leq \nu_g \leq \nu_h \leq 3$, then we have the same conclusion as Corollary 3.4.2.

Given n smooth convex univariate functions $\varphi_i : \mathbb{R} \to \mathbb{R}$ satisfying (3.1) for $i = 1, \ldots, n$ with the same order $\nu > 0$, we consider the function $f : \mathbb{R}^p \to \mathbb{R}$ defined by the following:

$$f(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^{n} \varphi_i(\mathbf{a}_i^\top \mathbf{x} + b_i), \tag{3.5}$$

where $\mathbf{a}_i \in \mathbb{R}^p$ and $b_i \in \mathbb{R}$ are given vectors and scalars, respectively for $i = 1, \cdots, n$. This convex function is called a finite sum and widely used in machine learning and statistics. The decomposable structure in (3.5) often appears in generalized linear models [9, 14], and empirical

risk minimization [109], where φ_i is referred to as a loss function as can be found, e.g., in Table 3.1.

Next, we show that if φ_i is *generalized self-concordant* with $\nu \in [2, 3]$, then f is also *generalized self-concordant*. This result is a direct consequence of Proposition 3.3.1 and Proposition 3.3.2.

Corollary 3.4.4. If φ_i in (3.5) satisfies (3.1) for $i = 1, \ldots, n$ with the same order $\nu \in [2, 3]$ and $M_{\varphi_i} \geq 0$, then f defined by (3.5) also belongs to $\widetilde{\mathcal{F}}_{M_f,\nu}$ in the sense of Definition 3.1 with the same order ν and the constant $M_f := n^{\frac{\nu}{2}-1} \max\{M_{\varphi_i} \|\mathbf{a}_i\|_2^{3-\nu} \mid 1 \leq i \leq n\}$.

Finally, we show that if we regularize f in (3.5) by a strongly convex quadratic term, then the resulting function becomes self-concordant. The proof can follow the same path as [109, Lemma 2].

Proposition 3.4.5. Let $f(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^n \varphi_i(\mathbf{a}_i^\top \mathbf{x} + b_i) + \psi(\mathbf{x})$, where $\psi(\mathbf{x}) := \frac{1}{2} \langle \mathbf{Q}\mathbf{x}, \mathbf{x} \rangle + \mathbf{c}^\top \mathbf{x}$ is strongly convex quadratic function with $\mathbf{Q} \in \mathcal{S}_{++}^p$. If φ_i satisfies (3.1) for $i = 1, \cdots, n$ with the same order $\nu \in (0, 3]$ and a constant $M_{\varphi_i} > 0$, then $f \in \widetilde{\mathcal{F}}_{\hat{M}_f,3}$ in the sense of Definition 3.1 with $\hat{M}_f := \lambda_{\min}(\mathbf{Q})^{\frac{\nu-3}{2}} \max\{M_{\varphi_i} \|\mathbf{a}_i\|_2^{3-\nu} \mid 1 \leq i \leq n\}$.

3.5 Fenchel's conjugate of *generalized self-concordant* functions

Primal-dual theory is fundamental in convex optimization. Hence, it is important to study the Fenchel conjugate of *generalized self-concordant* functions.

Let $f : \mathbb{R}^p \to \mathbb{R}$ be an (M_f, ν)-*generalized self-concordant* function. We consider Fenchel's conjugate f^* of f as

$$f^*(\mathbf{x}) = \sup_{\mathbf{u}} \{\langle \mathbf{x}, \mathbf{u} \rangle - f(\mathbf{u}) \mid \mathbf{u} \in \mathrm{dom}(f)\}. \tag{3.6}$$

Since f is proper, closed, and convex, f^* is well-defined and also proper, closed, and convex. Moreover, since f is smooth and convex, by Fermat's rule, if $u^*(\mathbf{x})$ satisfies $\nabla f(u^*(\mathbf{x})) = \mathbf{x}$, then f^* is well-defined at $\mathbf{x}$. This shows that $\mathrm{dom}(f^*) = \{\mathbf{x} \in \mathbb{R}^p \mid \nabla f(u^*(\mathbf{x})) = \mathbf{x} \text{ is solvable}\}$.

Example 3.2. Let us look at some univariate functions. By using (3.6), we can directly show that:

1. If $\varphi(s) = \ln(1 + e^s)$, then $\varphi^*(t) = t\ln(t) + (1-t)\ln(1-t)$.

2. If $\varphi(s) = s\ln(s)$, then $\varphi^*(t) = e^{t-1}$.

3. If $\varphi(s) = e^s$, then $\varphi^*(t) = t\ln(t) - t$.

$\diamond$

Intuitively, these examples show that if φ is *generalized self-concordant* , then its conjugate φ^* is also *generalized self-concordant* . For more examples, see [3, Chapter 13]. Let us generalize this result in the following proposition, whose proof is given in Appendix A.1.1.

Proposition 3.5.1. If f is $\widetilde{\mathcal{F}}_{M_f,\nu}$ in $\mathrm{dom}(f) \subseteq \mathbb{R}^p$ such that $\nabla^2 f(\mathbf{x}) \succ 0$ for $\mathbf{x} \in \mathrm{dom}(f)$, then the conjugate function f^* of f given by (3.6) is well-defined, and belongs to $\widetilde{\mathcal{F}}_{M_{f^*},\nu_*}$ on

$$\mathrm{dom}(f^*) := \{\mathbf{x} \in \mathbb{R}^p \mid f(\mathbf{u}) - \langle \mathbf{x}, \mathbf{u}\rangle \text{ is bounded from below on } \mathrm{dom}(f)\},$$

where $M_{f^*} = M_f$ and $\nu_* = 6 - \nu$ provided that $\nu \in [3, 6)$ if $p > 1$ and $\nu \in (0, 6)$ if $p = 1$.

Moreover, we have $\nabla f^*(\mathbf{x}) = u^*(\mathbf{x})$ and $\nabla^2 f^*(\mathbf{x}) = \nabla^2 f(u^*(\mathbf{x}))^{-1}$, where $u^*(\mathbf{x})$ is a unique solution of the maximization problem $\max_{\mathbf{u}}\{\langle \mathbf{x}, \mathbf{u}\rangle - f(\mathbf{u}) \mid \mathbf{u} \in \mathrm{dom}(f)\}$ in (3.6) for any $\mathbf{x} \in \mathrm{dom}(f^*)$.

Proposition 3.5.1 allows us to apply our generalized self-concordance theory in this paper to the dual problem of a convex problem involving *generalized self-concordant* functions, especially, when the objective function of the primal problem is *generalized self-concordant* with $\nu \in (3, 4]$. The Fenchel conjugates are certainly useful when we develop optimization algorithms to solve constrained convex optimization involving *generalized self-concordant* functions, see, e.g., [30, 31].

3.6 Generalized self-concordant approximation of nonsmooth convex functions

Several well-known convex functions are nonsmooth. However, they can be approximated (up to an arbitrary accuracy) by a *generalized self-concordant* function via smoothing. Smoothing techniques clearly allow us to enrich the applicability of our theory to nonsmooth convex problems.

Inspired by Nesterov's smoothing techniques (introduced in Section 2.6), our goal is to choose an appropriate smoothing function ω such that the smoothed function f_γ is well-defined and *generalized self-concordant* for any fixed smoothness parameter $\gamma > 0$.

Example 3.3. Let us provide a few examples with well-known nonsmooth convex functions:

(a) Consider the ℓ_1-norm function $f(\mathbf{x}) := \|\mathbf{x}\|_1$ in $\mathbb{R}^p$. Then, it can be rewritten as

$$\|\mathbf{x}\|_1 = \max_{\mathbf{u}}\{\langle \mathbf{x}, \mathbf{u}\rangle \mid \|\mathbf{u}\|_\infty \leq 1\} = \max_{\mathbf{u},\mathbf{v}}\big\{ \langle \mathbf{x}, \mathbf{u} - \mathbf{v}\rangle \mid \sum_{i=1}^{p}(u_i + v_i) = 1,\ \mathbf{u}, \mathbf{v} \in \mathbb{R}^p_+\big\}.$$

We can smooth this function by f_γ by choosing $\omega(\mathbf{u}, \mathbf{v}) := \ln(2p) + \sum_{i=1}^p (u_i \ln(u_i) + v_i \ln(v_i))$. In this case, we obtain $f_\gamma(\mathbf{x}) = \gamma \ln\left(\sum_{i=1}^p \left(e^{x_i/\gamma} + e^{-x_i/\gamma}\right)\right) - \gamma \ln(2p)$. This function is clearly *generalized self-concordant* with $\nu = 2$, see [103, Lemma 4].

However, if we choose $\omega(\mathbf{u}) := p - \sum_{i=1}^p \sqrt{1 - u_i^2}$, then we get $f_\gamma(\mathbf{x}) = \sum_{i=1}^p \sqrt{x_i^2 + \gamma^2} - \gamma p$. In this case, $f_\gamma \in \widetilde{\mathcal{F}}_{M_{f_\gamma},\nu}$ with $\nu = \frac{8}{3}$ and $M_{f_\gamma} = 3\gamma^{-\frac{2}{3}}$.

(b) The hinge loss function $\varphi(t) := \max\{0, 1 - t\}$ can be written as $\varphi(t) = \frac{1}{2}|1 - t| + \frac{1}{2}(1 - t)$. Hence, we can smooth this function by $\varphi_\gamma(t) := \gamma \ln\left(\frac{e^{\frac{(1-t)}{\gamma}} + e^{-\frac{(1-t)}{\gamma}}}{2}\right) + \frac{1}{2}(1 - t)$ with a smoothness parameter $\gamma > 0$. Clearly, φ_γ is *generalized self-concordant* with $\nu = 2$.

$\diamond$

In many practical problems, the conjugate f^* of f can be written as the sum $f^* = \varphi + \delta_{\mathcal{U}}$, where φ is a *generalized self-concordant* function, and $\delta_{\mathcal{U}}$ is the indicator function of a given nonempty, closed, and convex set $\mathcal{U}$. In this case, f_γ in (2.12) becomes

$$f_\gamma(\mathbf{x}) := \sup_{\mathbf{u}}\{\langle \mathbf{x}, \mathbf{u}\rangle - \varphi(\mathbf{u}) - \gamma\omega(\mathbf{u}) \mid \mathbf{u} \in \mathcal{U}\}. \tag{3.7}$$

If ω is *generalized self-concordant* such that $\nu_\varphi = \nu_\omega$, and $\mathcal{U} = \overline{\operatorname{dom}(\omega) \cap \operatorname{dom}(\varphi)}$, then f_γ is also *generalized self-concordant* with $\nu_{f_\gamma} = 6 - \nu_\varphi$ as shown in Proposition 3.5.1.

3.7 Key bounds on Hessian, gradient and function values

Now, we develop some key bounds on the local norms, Hessian, gradient and function values of *generalized self-concordant* functions.

For this purpose, given $\nu \geq 2$, we define the following quantity for any $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$:

$$d_\nu(\mathbf{x}, \mathbf{y}) := M\|\mathbf{y} - \mathbf{x}\|_2^{3-\nu}\|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}}^{\nu-2}. \tag{3.8}$$

Here, if $\nu > 3$, then we require $\mathbf{x} \neq \mathbf{y}$. Otherwise, we set $d_\nu(\mathbf{x}, \mathbf{y}) := 0$ if $\mathbf{x} = \mathbf{y}$. In addition, we also define the function $\bar{\bar{\omega}}_\nu : \mathbb{R} \to \mathbb{R}_+$ as

$$\bar{\bar{\omega}}_\nu(\tau) := \begin{cases} \left(\frac{1}{1-\frac{\nu-2}{2}\tau}\right)^{\frac{2}{\nu-2}} & \text{if } \nu > 2 \\ e^\tau & \text{if } \nu = 2. \end{cases} \tag{3.9}$$

with $\mathrm{dom}(\bar{\bar{\omega}}_\nu) = \left(-\infty, \frac{2}{\nu-2}\right)$ if $\nu > 2$, and $\mathrm{dom}(\bar{\bar{\omega}}_\nu) = \mathbb{R}$ if $\nu = 2$. We also adopt the Dikin ellipsoidal notion from [75] as $W^0(\mathbf{x}; r) := \{\mathbf{y} \in \mathbb{R}^p \mid \frac{\nu-2}{2} d_\nu(\mathbf{x}, \mathbf{y}) < r\}$.

The next proposition provides some bounds on the local norm defined by *generalized self-concordant* function f. These bounds are given for the local distance $\|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}}$ and $\|\mathbf{y} - \mathbf{x}\|_{\mathbf{y}}$ between two points $\mathbf{x}$ and $\mathbf{y}$ in $\mathrm{dom}(f)$.

Proposition 3.7.1 (Bounds of local norms)**.** If $\nu > 2$, then, for any $\mathbf{x} \in \mathrm{dom}(f)$, we have $W^0(\mathbf{x}; 1) \subseteq \mathrm{dom}(f)$. For any $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$, let $d_\nu(\mathbf{x}, \mathbf{y})$ be defined by (3.8), and $\bar{\bar{\omega}}_\nu(\cdot)$ be defined by (3.9). Then, we have

$$\bar{\bar{\omega}}_\nu\left(-d_\nu(\mathbf{x}, \mathbf{y})\right)^{\frac{1}{2}} \|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}} \leq \|\mathbf{y} - \mathbf{x}\|_{\mathbf{y}} \leq \bar{\bar{\omega}}_\nu\left(d_\nu(\mathbf{x}, \mathbf{y})\right)^{\frac{1}{2}} \|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}}. \tag{3.10}$$

If $\nu > 2$, then the right-hand side inequality of (3.10) holds if $d_\nu(\mathbf{x}, \mathbf{y}) < \frac{2}{\nu-2}$.

Proof. We first consider the case $\nu > 2$. Let $\mathbf{u} \in \mathbb{R}^p$ and $\mathbf{u} \neq 0$. Consider the following univariate function

$$\phi(t) := \left\langle \nabla^2 f(\mathbf{x} + t\mathbf{u})\mathbf{u}, \mathbf{u} \right\rangle^{1-\frac{\nu}{2}} = \|\mathbf{u}\|_{\mathbf{x}+t\mathbf{u}}^{2-\nu}.$$

It is easy to compute the derivative of this function, and obtain

$$\phi'(t) = \left(\frac{2-\nu}{2}\right) \frac{\left\langle \nabla^3 f(\mathbf{x} + t\mathbf{u})[\mathbf{u}]\mathbf{u}, \mathbf{u} \right\rangle}{\left\langle \nabla^2 f(\mathbf{x} + t\mathbf{u})\mathbf{u}, \mathbf{u} \right\rangle^{\frac{\nu}{2}}} = \left(\frac{2-\nu}{2}\right) \frac{\left\langle \nabla^3 f(\mathbf{x} + t\mathbf{u})[\mathbf{u}]\mathbf{u}, \mathbf{u} \right\rangle}{\|\mathbf{u}\|_{\mathbf{x}+t\mathbf{u}}^{\nu}}.$$

Using Definition 3.1 with $\mathbf{u} = \mathbf{v}$ and $\mathbf{x} + t\mathbf{u}$ instead of $\mathbf{x}$, we have $|\phi'(t)| \leq \frac{\nu-2}{2} M_f \|\mathbf{u}\|_2^{3-\nu}$. This implies that $\phi(t) \geq \phi(0) - \frac{\nu-2}{2} M_f \|\mathbf{u}\|_2^{3-\nu}|t|$. On the other hand, we can see that $\mathrm{dom}(\phi) = \{t \in \mathbb{R} \mid \phi(t) > 0\}$. Hence, we have $\mathrm{dom}(\phi)$ contains $\left(-\frac{2\phi(0)}{(\nu-2)M_f\|\mathbf{u}\|_2^{3-\nu}}, \frac{2\phi(0)}{(\nu-2)M_f\|\mathbf{u}\|_2^{3-\nu}}\right)$. Using this fact and the definition of ϕ, we can show that $\mathrm{dom}(f)$ contains $\{\mathbf{y} := \mathbf{x} + t\mathbf{u} \mid |t| < \frac{2\|\mathbf{u}\|_{\mathbf{x}}^{2-\nu}}{(\nu-2)M_f\|\mathbf{u}\|_2^{3-\nu}}\}$. However, since $|t| = \frac{\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^{\nu-2}}{\|\mathbf{u}\|_{\mathbf{x}}^{\nu-2}} \frac{\|\mathbf{y}-\mathbf{x}\|_2^{3-\nu}}{\|\mathbf{u}\|_2^{3-\nu}}$, the condition $|t| < \frac{2\|\mathbf{u}\|_{\mathbf{x}}^{2-\nu}}{(\nu-2)M_f\|\mathbf{u}\|_2^{3-\nu}}$ is equivalent to $d_\nu(\mathbf{x}, \mathbf{y}) < \frac{2}{\nu-2}$. This shows that $W^0(\mathbf{x}; 1) \subseteq \mathrm{dom}(f)$.

Since $\left| \int_0^1 \phi'(t)\mathrm{d}t \right| \leq \int_0^1 |\phi'(t)|\mathrm{d}t$, integrating $\phi'(t)$ over the interval $[0, 1]$ we get

$$\left| \|\mathbf{u}\|_{\mathbf{x}+\mathbf{u}}^{2-\nu} - \|\mathbf{u}\|_{\mathbf{x}}^{2-\nu} \right| \leq \frac{\nu-2}{2} M_f \|\mathbf{u}\|_2^{3-\nu}.$$

Using $\mathbf{u} = \mathbf{y} - \mathbf{x}$ in the last inequality, we get $|\|\mathbf{y}-\mathbf{x}\|_{\mathbf{y}}^{2-\nu} - \|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^{2-\nu}| \leq \frac{\nu-2}{2} M_f \|\mathbf{y}-\mathbf{x}\|_2^{3-\nu}$, which is equivalent to

$$\begin{aligned}
\|\mathbf{y}-\mathbf{x}\|_{\mathbf{y}}^{\nu-2} &\leq \|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^{\nu-2} \left(1 - \tfrac{\nu-2}{2} M_f \|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^{\nu-2} \|\mathbf{x}-\mathbf{y}\|_2^{3-\nu}\right)^{-1} = \|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^{\nu-2} \left(1 - \tfrac{\nu-2}{2} d_\nu(\mathbf{x}, \mathbf{y})\right)^{-1} \\
\|\mathbf{y}-\mathbf{x}\|_{\mathbf{y}}^{\nu-2} &\geq \|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^{\nu-2} \left(1 + \tfrac{\nu-2}{2} M_f \|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^{\nu-2} \|\mathbf{x}-\mathbf{y}\|_2^{3-\nu}\right)^{-1} = \|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^{\nu-2} \left(1 + \tfrac{\nu-2}{2} d_\nu(\mathbf{x}, \mathbf{y})\right)^{-1},
\end{aligned}$$

given that $d_\nu(\mathbf{x}, \mathbf{y}) < \frac{2}{\nu-2}$. Taking the power of $\frac{1}{\nu-2} > 0$ on both sides, we get (3.10) for the case $\nu > 2$.

Now, we consider the case $\nu = 2$. Let $0 \neq \mathbf{u} \in \mathbb{R}^p$. We consider the following function

$$\phi(t) := \ln\left(\left\langle \nabla^2 f(\mathbf{x} + t\mathbf{u})\mathbf{u}, \mathbf{u} \right\rangle\right) = \ln\left(\|\mathbf{u}\|_{\mathbf{x}+t\mathbf{u}}^2\right).$$

Clearly, it is easy to show that $\phi'(t) = \frac{\left\langle \nabla^3 f(\mathbf{x}+t\mathbf{u})[\mathbf{u}]\mathbf{u}, \mathbf{u} \right\rangle}{\left\langle \nabla^2 f(\mathbf{x}+t\mathbf{u})\mathbf{u}, \mathbf{u} \right\rangle} = \frac{\left\langle \nabla^3 f(\mathbf{x}+t\mathbf{u})[\mathbf{u}]\mathbf{u}, \mathbf{u} \right\rangle}{\|\mathbf{u}\|_{\mathbf{x}+t\mathbf{u}}^2}$. Using again Definition 3.1 with $\mathbf{u} = \mathbf{v}$ and $\mathbf{x} + t\mathbf{u}$ instead of $\mathbf{x}$, we obtain $|\phi'(t)| \leq M_f \|\mathbf{u}\|_2$.
Since $\left| \int_0^1 \phi'(t)\mathrm{d}t \right| \leq \int_0^1 |\phi'(t)|\mathrm{d}t$, integrating $\phi'(t)$ over the interval $[0, 1]$ we get

$$\left| \ln\left(\|\mathbf{u}\|_{\mathbf{x}+\mathbf{u}}^2\right) - \ln\left(\|\mathbf{u}\|_{\mathbf{x}}^2\right) \right| \leq M_f \|\mathbf{u}\|_2.$$

Substituting $\mathbf{u} = \mathbf{y} - \mathbf{x}$ into this inequality, we get $\left| \ln \|\mathbf{y}-\mathbf{x}\|_{\mathbf{y}} - \ln \|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} \right| \leq \frac{M_f}{2} \|\mathbf{y}-\mathbf{x}\|_2$. Hence, $\ln \|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} - \frac{M_f}{2}\|\mathbf{y}-\mathbf{x}\|_2 \leq \ln \|\mathbf{y}-\mathbf{x}\|_{\mathbf{y}} \leq \ln \|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} + \frac{M_f}{2}\|\mathbf{y}-\mathbf{x}\|_2$. This inequality leads to (3.10) for the case $\nu = 2$. □

Next, we develop new bounds for the Hessian of f in the following proposition.

Proposition 3.7.2 (Bounds of Hessian)**.** For any $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$, let $d_\nu(\mathbf{x}, \mathbf{y})$ be defined by (3.8), and $\bar{\bar{\omega}}_\nu(\cdot)$ be defined by (3.9). Then, we have

$$\bar{\bar{\omega}}_\nu\left(d_\nu(\mathbf{x}, \mathbf{y})\right)^{-1} \nabla^2 f(\mathbf{x}) \preceq \nabla^2 f(\mathbf{y}) \preceq \bar{\bar{\omega}}_\nu\left(d_\nu(\mathbf{x}, \mathbf{y})\right) \nabla^2 f(\mathbf{x}), \tag{3.11}$$

where $d_\nu(\mathbf{x}, \mathbf{y}) < \frac{2}{\nu-2}$ is required for the case $\nu > 2$.

Proof. Let $\nu > 2$ and $0 \neq \mathbf{u} \in \mathbb{R}^n$. Consider the following univariate function on $[0, 1]$:

$$\psi(t) := \left\langle \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))\mathbf{u}, \mathbf{u} \right\rangle, \quad t \in [0, 1].$$

If we denote by $\mathbf{y}_t := \mathbf{x} + t(\mathbf{y} - \mathbf{x})$, then $\mathbf{y}_t - \mathbf{x} = t(\mathbf{y} - \mathbf{x})$, $\psi(t) = \|\mathbf{u}\|_{\mathbf{y}_t}^2$, and $\psi'(t) = \left\langle \nabla^3 f(\mathbf{y}_t)[\mathbf{y} - \mathbf{x}]\mathbf{u}, \mathbf{u} \right\rangle$. By Definition 3.1, we have

$$|\psi'(t)| \ \leq M_f \|\mathbf{u}\|_{\mathbf{y}_t}^2 \|\mathbf{y} - \mathbf{x}\|_{\mathbf{y}_t}^{\nu-2} \|\mathbf{y} - \mathbf{x}\|_2^{3-\nu} = M_f \psi(t) \left[\tfrac{\|\mathbf{y}_t - \mathbf{x}\|_{\mathbf{y}_t}}{t}\right]^{\nu-2} \|\mathbf{y} - \mathbf{x}\|_2^{3-\nu},$$

which implies

$$|\frac{d \ln \psi(t)}{dt}| \leq M_f \left[\tfrac{\|\mathbf{y}_t - \mathbf{x}\|_{\mathbf{y}_t}}{t}\right]^{\nu-2} \|\mathbf{y} - \mathbf{x}\|_2^{3-\nu}. \tag{3.12}$$

Assume that $d_\nu(\mathbf{x}, \mathbf{y}) < \frac{2}{\nu-2}$. Then, by the definition of $\mathbf{y}_t$ and $d_\nu(\cdot)$, we have $d_\nu(\mathbf{x}, \mathbf{y}_t) = t d_\nu(\mathbf{x}, \mathbf{y})$ and $\|\mathbf{y}_t - \mathbf{x}\|_{\mathbf{x}} = t\|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}}$. Using Proposition 3.7.1, we can derive

$$\begin{aligned} \tfrac{1}{t}\|\mathbf{y}_t - \mathbf{x}\|_{\mathbf{y}_t} &\leq \tfrac{1}{t}\left[1 - \tfrac{\nu-2}{2} d_\nu(\mathbf{x}, \mathbf{y}_t)\right]^{-\frac{1}{\nu-2}} \|\mathbf{y}_t - \mathbf{x}\|_{\mathbf{x}} \\ &= \left[1 - \tfrac{\nu-2}{2} d_\nu(\mathbf{x}, \mathbf{y}) t\right]^{-\frac{1}{\nu-2}} \|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}}. \end{aligned}$$

Hence, we can further derive

$$\left[\frac{1}{t}\|\mathbf{y}_t - \mathbf{x}\|_{\mathbf{y}_t}\right]^{\nu-2} \leq \frac{\|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}}^{\nu-2}}{1 - \frac{\nu-2}{2} d_\nu(\mathbf{x}, \mathbf{y}) t}$$

Integrating $\frac{d \ln \psi(t)}{dt}$ with respect to t on $[0, 1]$ and using the last inequality and (3.12), we get

$$|\int_0^1 \frac{d \ln \psi(t)}{dt} dt| \leq \int_0^1 |\frac{d \ln \psi(t)}{dt}| dt \leq \|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}}^{\nu-2} \|\mathbf{y} - \mathbf{x}\|_2^{3-\nu} \int_0^1 \frac{dt}{1 - \frac{\nu-2}{2} d_\nu(\mathbf{x}, \mathbf{y}) t}.$$

Clearly, we can compute this integral explicitly as

$$|\ln\left[\frac{\|\mathbf{u}\|_{\mathbf{y}}^2}{\|\mathbf{u}\|_{\mathbf{x}}^2}\right]| = |\ln\left[\frac{\psi(1)}{\psi(0)}\right]| \leq \frac{-2d_\nu(\mathbf{x},\mathbf{y})}{(\nu-2)d_\nu(\mathbf{x},\mathbf{y})}\ln\left[1-\frac{\nu-2}{2}d_\nu(\mathbf{x},\mathbf{y})\right] = \ln\left[\left(1-\frac{\nu-2}{2}d_\nu(\mathbf{x},\mathbf{y})\right)^{\frac{-2}{\nu-2}}\right].$$

Rearranging this inequality, we obtain

$$\left[1-\frac{\nu-2}{2}d_\nu(\mathbf{x},\mathbf{y})\right]^{\frac{2}{\nu-2}} \leq \frac{\|\mathbf{u}\|_{\mathbf{y}}^2}{\|\mathbf{u}\|_{\mathbf{x}}^2} \equiv \frac{\langle\nabla^2 f(\mathbf{y})\mathbf{u},\mathbf{u}\rangle}{\langle\nabla^2 f(\mathbf{x})\mathbf{u},\mathbf{u}\rangle} \leq \left[1-\frac{\nu-2}{2}d_\nu(\mathbf{x},\mathbf{y})\right]^{\frac{-2}{\nu-2}}.$$

Since this inequality holds for any $0 \neq \mathbf{u} \in \mathbb{R}^p$, it implies (3.11). If $\mathbf{u} = 0$, then (3.11) obviously holds.

Now, we consider the case $\nu = 2$. It follows from (3.12) that

$$|\ln\left[\frac{\|\mathbf{u}\|_{\mathbf{y}}^2}{\|\mathbf{u}\|_{\mathbf{x}}^2}\right]| = |\int_0^1 \frac{d\ln\psi(t)}{dt}dt| \leq \int_0^1 |\frac{d\ln\psi(t)}{dt}|dt \leq M_f\int_0^1 \|\mathbf{y}-\mathbf{x}\|_2 dt = M_f\|\mathbf{y}-\mathbf{x}\|_2.$$

Since this inequality holds for any $\mathbf{u} \in \mathbb{R}^p$, it implies (3.11). □

The following corollary provides a bound on the mean of the Hessian $G(\mathbf{x},\mathbf{y}) := \int_0^1 \nabla^2 f(\mathbf{x}+\tau(\mathbf{y}-\mathbf{x}))d\tau$, whose proof is moved to Appendix A.1.2.

Corollary 3.7.3. For any $\mathbf{x},\mathbf{y} \in \mathrm{dom}(f)$, let $d_\nu(\mathbf{x},\mathbf{y})$ be defined by (3.8). Then, we have

$$\underline{\kappa}_\nu(d_\nu(\mathbf{x},\mathbf{y}))\nabla^2 f(\mathbf{x}) \preceq \int_0^1 \nabla^2 f(\mathbf{x}+\tau(\mathbf{y}-\mathbf{x}))d\tau \preceq \overline{\kappa}_\nu(d_\nu(\mathbf{x},\mathbf{y}))\nabla^2 f(\mathbf{x}), \tag{3.13}$$

where

$$\underline{\kappa}_\nu(t) := \begin{cases} \frac{1-e^{-t}}{t} & \text{if } \nu = 2 \\ \frac{2}{\nu t}\left[1-\left(1-\frac{\nu-2}{2}t\right)^{\frac{\nu}{\nu-2}}\right] & \text{if } \nu > 2 \end{cases} \quad \text{and}$$

$$\overline{\kappa}_\nu(t) := \begin{cases} \frac{e^t-1}{t} & \text{if } \nu = 2 \\ -\frac{\ln(1-t)}{t} & \text{if } \nu = 4 \\ \frac{2}{(\nu-4)t}\left[1-\left(1-\frac{\nu-2}{2}t\right)^{\frac{\nu-4}{\nu-2}}\right] & \text{if } \nu > 2,\ \nu \neq 4. \end{cases}$$

Here, if $\nu > 2$, then we require the condition $d_\nu(\mathbf{x},\mathbf{y}) < \frac{2}{\nu-2}$ in (3.13).

Remark 3.7.1. In the above proposition, $\underline{\kappa}_\nu$ and $\overline{\kappa}_\nu$ are always non-negative and well-defined on their domains, respectively.

We prove a bound on the gradient inner product of $f \in \widetilde{\mathcal{F}}_{M,\nu}$.

Proposition 3.7.4 (Bounds of gradient map)**.** For any $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$, we have

$$\overline{\kappa}_\nu\left(-d_\nu(\mathbf{x},\mathbf{y})\right)\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^2 \leq \langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \leq \overline{\kappa}_\nu\left(d_\nu(\mathbf{x},\mathbf{y})\right)\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^2, \tag{3.14}$$

where, if $\nu > 2$, then the right-hand side inequality of (3.14) holds if $d_\nu(\mathbf{x},\mathbf{y}) < \frac{2}{\nu-2}$.

Proof. Let $\mathbf{y}_t := \mathbf{x} + t(\mathbf{y}-\mathbf{x})$. By the mean-value theorem, we have

$$\langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle = \int_0^1 \left\langle \nabla^2 f(\mathbf{y}_t)(\mathbf{y}-\mathbf{x}), \mathbf{y}-\mathbf{x}\right\rangle \mathrm{d}t = \int_0^1 \frac{1}{t^2}\|\mathbf{y}_t-\mathbf{x}\|_{\mathbf{y}_t}^2 \mathrm{d}t. \tag{3.15}$$

We consider the function $\bar{\bar{\omega}}_\nu$ defined by (3.9). It follows from Proposition 3.7.1 that

$$\bar{\bar{\omega}}_\nu\left(-d_\nu(\mathbf{x},\mathbf{y}_t)\right)\|\mathbf{y}_t-\mathbf{x}\|_{\mathbf{x}}^2 \leq \|\mathbf{y}_t-\mathbf{x}\|_{\mathbf{y}_t}^2 \leq \bar{\bar{\omega}}_\nu\left(d_\nu(\mathbf{x},\mathbf{y}_t)\right)\|\mathbf{y}_t-\mathbf{x}\|_{\mathbf{x}}^2.$$

Note that $d_\nu(\mathbf{x},\mathbf{y}_t) = t d_\nu(\mathbf{x},\mathbf{y})$ and $\|\mathbf{y}_t-\mathbf{x}\|_{\mathbf{x}} = t\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}$, the last estimate leads to

$$\bar{\bar{\omega}}_\nu\left(-t d_\nu(\mathbf{x},\mathbf{y})\right)\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^2 \leq \frac{1}{t^2}\|\mathbf{y}_t-\mathbf{x}\|_{\mathbf{y}_t}^2 \leq \bar{\bar{\omega}}_\nu\left(t d_\nu(\mathbf{x},\mathbf{y})\right)\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^2.$$

Substituting this estimate into (3.15), we obtain

$$\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^2 \int_0^1 \bar{\bar{\omega}}_\nu\left(-t d_\nu(\mathbf{x},\mathbf{y})\right) dt \leq \langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \leq \|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^2 \int_0^1 \bar{\bar{\omega}}_\nu\left(t d_\nu(\mathbf{x},\mathbf{y})\right) dt.$$

Using the function $\bar{\bar{\omega}}_\nu(\tau)$ from (3.9) to compute the left-hand side and the right-hand side integrals, we obtain (3.14). □

Finally, we prove a bound on function values of $f \in \widetilde{\mathcal{F}}_{M,\nu}$ in the following proposition.

Proposition 3.7.5 (Bounds of function values)**.** For any $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$, we have

$$\omega_\nu\left(-d_\nu(\mathbf{x},\mathbf{y})\right)\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^2 \leq f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \leq \omega_\nu\left(d_\nu(\mathbf{x},\mathbf{y})\right)\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^2, \tag{3.16}$$

where, if $\nu > 2$, then the right-hand side inequality holds if $d_\nu(\mathbf{x}, \mathbf{y}) < \frac{2}{\nu-2}$. Here, $d_\nu(\mathbf{x}, \mathbf{y})$ is defined by (3.8) and ω_ν is defined by

$$\omega_\nu(\tau) := \begin{cases} \frac{e^\tau - \tau - 1}{\tau^2} & \text{if } \nu = 2 \\ \frac{-2\tau - 4\ln(1-\frac{\tau}{2})}{\tau^2} & \text{if } \nu = 3 \\ \frac{(1-\tau)\ln(1-\tau)+\tau}{\tau^2} & \text{if } \nu = 4 \\ \left(\frac{2}{4-\nu}\right)\frac{1}{\tau}\left[\frac{1}{(3-\nu)\tau}\left(\left(1-\frac{\nu-2}{2}\tau\right)^{\frac{2(3-\nu)}{2-\nu}} - 1\right) - 1\right] & \text{otherwise.} \end{cases} \tag{3.17}$$

Note that $\omega_\nu(\tau) \geq 0$ for all $\tau \in \mathrm{dom}(\omega_\nu)$.

Proof. For any $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$, let $\mathbf{y}_t := \mathbf{x} + t(\mathbf{y} - \mathbf{x})$. Then, $\mathbf{y}_t - \mathbf{x} = t(\mathbf{y} - \mathbf{x})$. By the mean-value theorem, we have

$$f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle = \int_0^1 \tfrac{1}{t} \langle \nabla f(\mathbf{y}_t) - \nabla f(\mathbf{x}), \mathbf{y}_t - \mathbf{x} \rangle \, dt.$$

Now, by Proposition 3.7.4, we have

$$\overline{\kappa}_\nu\left(-d_\nu(\mathbf{x}, \mathbf{y}_t)\right) \|\mathbf{y}_t - \mathbf{x}\|_{\mathbf{x}}^2 \leq \langle \nabla f(\mathbf{y}_t) - \nabla f(\mathbf{x}), \mathbf{y}_t - \mathbf{x} \rangle \leq \overline{\kappa}_\nu\left(d_\nu(\mathbf{x}, \mathbf{y}_t)\right) \|\mathbf{y}_t - \mathbf{x}\|_{\mathbf{x}}^2.$$

Clearly, by the definition (3.8), we have $d_\nu(\mathbf{x}, \mathbf{y}_t) = t d_\nu(\mathbf{x}, \mathbf{y})$ and $\|\mathbf{y}_t - \mathbf{x}\|_{\mathbf{x}} = t\|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}}$. Combining these relations, and the above two inequalities, we can show that

$$\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^2 \int_0^1 t\overline{\kappa}_\nu\left(-t d_\nu(\mathbf{x}, \mathbf{y})\right) dt \leq f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^2 \int_0^1 t\overline{\kappa}_\nu\left(t d_\nu(\mathbf{x}, \mathbf{y})\right) dt.$$

By integrating the left- and right-hand side of the above inequality, we obtain (3.16). □

3.8 Conclusion

We have generalized the self-concordance notion in [75] to a more general class of smooth and convex functions. Such a function class covers several well-known examples, including logistic, exponential, reciprocal and standard self-concordant functions. We developed a unified

theory to reveal the smoothness structures of this functional class and discussed the behavior in the dual space. We also obtained some fundamental properties incorporating generalized self-concordance with Lipschitz gradient and strong convexity. We provided several key bounds on the Hessian, gradient and function value of this function class. For our reference convenience, we provide a short summary on the main properties of *generalized self-concordant* functions in Table 3.2 below.

Table 3.2: Summary of gsc properties and the corresponding range of ν

Result	**Property**	**Range of ν**
Definitions 3.2.1 and 3.1	definitions of gsc functions	$\nu > 0$
Proposition 3.3.1	sum of gsc functions	$\nu \geq 2$
Proposition 3.3.2	affine transformation of gsc functions with $\mathcal{A}(x) = Ax + b$	$\nu \in (0, 3]$ for general A $\nu > 3$ for over-completed A
Proposition 3.3.4(a)	non-degenerate property	$\nu \geq 2$
Proposition 3.3.4(b)	unboundedness	$\nu > 0$
Proposition 3.4.1(a)	gsc and strong convexity	$\nu \in (0, 3]$
Proposition 3.4.1(b)	gsc and Lipschitz gradient continuity	$\nu \geq 2$
Proposition 3.5.1	if f^* is the conjugate of a gsc function f, then $\nu + \nu_* = 6$	$\nu_* \in (0, 6)$ if $p = 1$ (univariate) $\nu_* \in [3, 6)$ if $p > 1$ (multivariate)
Propositions 3.7.1, 3.7.2, 3.7.4, and 3.7.5	local norm, Hessian, gradient, and function value bounds	$\nu \geq 2$

Although several results hold for a different range of ν, the complete theory only holds for $\nu \in [2, 3]$. However, this is sufficient to cover two important cases: $\nu = 2$ in [1, 2] and $\nu = 3$ in [75]. We will further illustrate our main theory and algorithm in Chapter 4.

CHAPTER 4

Generalized self-concordant minimization

4.1 Introduction

In this chapter, we apply the theory developed in the Chapter 3 to design new Newton-type methods to minimize a *generalized self-concordant* function. As stated in Chapter 1, we can prove both local and global convergence for composite optimization by using our new concept and theory, without additional smoothness assumptions.

In the rest of this chapter, Section 4.2 is devoted to studying a full-step and damped-step Newton schemes to minimize a generalized self-concordant function including their convergence guarantee. Section 4.3 extends to the composite setting (2.7) and studies proximal Newton-type methods, and investigates their convergence guarantees. Numerical examples are provided in Section 4.4 to illustrate the advantages of our theory. Section 4.5 summarizes our conclusion. Besides, several technical results and proofs are moved to the appendix.

4.2 Generalized self-concordant minimization

We apply the theory developed in Chapter 3 to design new Newton-type methods to minimize a *generalized self-concordant* function. More precisely, we consider the following non-composite convex problem formulation (equation (2.1) of Chapter 2):

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}),$$

where $f \in \widetilde{\mathcal{F}}^p_{M_f,\nu} : \mathbb{R}^p \to \mathbb{R}$ in the sense of Definition 3.1 with $\nu \in [2,3]$ and $M_f \geq 0$. Since f is smooth and convex, the optimality condition $\nabla f(\mathbf{x}^\star_f) = 0$ is necessary and sufficient for $\mathbf{x}^\star_f$ to be an optimal solution of (2.1).

The following theorem shows the existence and uniqueness of the solution $\mathbf{x}_f^\star$ of (2.1). It can be considered as a special case of Theorem 4.3.1 below with $g \equiv 0$.

Theorem 4.2.1. Suppose that $f \in \widetilde{\mathcal{F}}_{M_f,\nu}(\mathrm{dom}(f))$ for given parameters $M_f > 0$ and $\nu \in [2, 3]$. Denote by $\sigma_{\min}(\mathbf{x}) := \lambda_{\min}(\nabla^2 f(\mathbf{x}))$ and $\lambda(\mathbf{x}) := \|\nabla f(\mathbf{x})\|_{\mathbf{x}}^*$ for $\mathbf{x} \in \mathrm{dom}(f)$. Suppose further that there exists $\mathbf{x} \in \mathrm{dom}(f)$ such that $\sigma_{\min}(\mathbf{x}) > 0$ and

$$\lambda(\mathbf{x}) < \frac{2\left[\sigma_{\min}(\mathbf{x})\right]^{\frac{3-\nu}{2}}}{(4-\nu)M_f}.$$

Then, problem (2.1) has a unique solution $\mathbf{x}_f^\star$ in $\mathrm{dom}(f)$.

We say that the unique solution $\mathbf{x}_f^\star$ of (2.1) is *strongly regular* if $\nabla^2 f(\mathbf{x}_f^\star) \succ 0$. The strong regularity of $\mathbf{x}_f^\star$ for (2.1) is equivalent to the strong second order optimality condition. Theorem 4.2.1 covers [74, Theorem 4.1.11] for standard self-concordant functions as a special case.

We consider the following Newton-type scheme to solve (2.1). Starting from an arbitrary initial point $\mathbf{x}^0 \in \mathrm{dom}(f)$, we generate a sequence $\{\mathbf{x}^k\}_{k\geq 0}$ as follows:

$$\mathbf{x}^{k+1} := \mathbf{x}^k + \tau_k n_{\mathrm{nt}}^k, \qquad \text{where } \; n_{\mathrm{nt}}^k := -\nabla^2 f(\mathbf{x}^k)^{-1}\nabla f(\mathbf{x}^k), \tag{4.1}$$

and $\tau_k \in (0, 1]$ is a given step-size. We call n_{nt}^k a Newton direction.

- If $\tau_k = 1$ for all $k \geq 0$, then we call (4.1) a *full-step* Newton scheme.
- Otherwise, i.e., $\tau_k \in (0, 1)$, we call (4.1) a *damped-step* Newton scheme.

Clearly, computing the Newton direction n_{nt}^k requires to solve the following linear system:

$$\nabla^2 f(\mathbf{x}^k) n_{\mathrm{nt}}^k = -\nabla f(\mathbf{x}^k). \tag{4.2}$$

Next, we define a *Newton decrement* λ_k and a quantity β_k, respectively as

$$\lambda_k := \|n_{\mathrm{nt}}^k\|_{\mathbf{x}^k} \quad \text{and} \quad \beta_k := \|n_{\mathrm{nt}}^k\|_2. \tag{4.3}$$

With λ_k and β_k given by (4.3), we also define

$$d_k := M_f \lambda_k^{\nu-2} \beta_k^{3-\nu}, \text{ for } \nu \in [2,3], \tag{4.4}$$

then $\tau_k d_k = d_\nu(\mathbf{x}^k, \mathbf{x}^{k+1})$ by the definition of d_ν in (3.8). Let us first show how to choose a suitable step-size τ_k in the damped-step Newton scheme and prove its convergence properties in the following theorem, whose proof can be found in Appendix A.2.2.

Theorem 4.2.2. Let $\{\mathbf{x}^k\}$ be the sequence generated by the damped-step Newton scheme (4.1) with the following step-size:

$$\tau_k := \begin{cases} \frac{1}{d_k}\ln(1+d_k) & \text{if } \nu = 2 \\ \frac{2}{(\nu-2)d_k}\left[1 - \left(1 + \frac{4-\nu}{2}d_k\right)^{-\frac{\nu-2}{4-\nu}}\right] & \text{if } \nu \in (2,3], \end{cases} \tag{4.5}$$

where d_k is defined by (4.4). Then, $\tau_k \in (0,1]$, $\{\mathbf{x}^k\}$ in $\mathrm{dom}(f)$, and this step-size guarantees the following descent property

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \Delta_k, \tag{4.6}$$

where $\Delta_k := \lambda_k^2 \tau_k - \omega_\nu (\tau_k d_k) \tau_k^2 \lambda_k^2 > 0$ with ω_ν defined by (3.17).

Assume that the unique solution $\mathbf{x}_f^\star$ of (2.1) exists. Then, there exists a neighborhood $\mathcal{N}(\mathbf{x}_f^\star)$ such that if we initialize the Newton scheme (4.1) at $\mathbf{x}^0 \in \mathcal{N}(\mathbf{x}_f^\star) \cap \mathrm{dom}(f)$, then the whole sequence $\{\mathbf{x}^k\}$ converges to $\mathbf{x}_f^\star$ at a quadratic rate.

Example 4.1 Better step-size for regularized logistic and exponential models. Consider the minimization problem (2.1) with the objective function $f(\cdot) := \phi(\cdot) + \frac{\gamma}{2}\|\cdot\|_2^2$, where ϕ is defined as in (3.5) with $\varphi_i(t) = \ln(1+e^{-t})$ being the logistic loss. That is

$$f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n \ln(1 + e^{-\mathbf{a}_i^\top \mathbf{x}}) + \frac{\gamma}{2}\|\mathbf{x}\|_2^2.$$

As we shown in Chapter 3 that f is either *generalized self-concordant* with $\nu = 2$ or *generalized self-concordant* with $\nu = 3$ but with different constant M_f.

Let us define $R_A := \max\{\|\mathbf{a}_i\|_2 \mid 1 \leq i \leq n\}$. Then, if we consider $\nu = 2$, then we have $M_f^{(2)} = R_A$ due to Corollary 3.4.4, while if we choose $\nu = 3$, then $M_f^{(3)} = \frac{1}{\sqrt{\gamma}} R_A$ due to

Proposition 3.4.1. By the definition of f, we have $\nabla^2 f(\mathbf{x}) \succeq \gamma \mathbb{I}$. Hence, using this inequality and the definition of λ_k and d_k from (4.3), we can show that

$$d_k = M_f^{(2)} \|\nabla^2 f(\mathbf{x}^k)^{-1} \nabla f(\mathbf{x}^k)\|_2 \leq \tfrac{R_A}{\sqrt{\gamma}} \lambda_k = M_f^{(3)} \lambda_k. \tag{4.7}$$

For any $\tau > 0$, we have $\frac{\ln(1+\tau)}{\tau} > \frac{1}{1+0.5\tau}$. Using this elementary result and (4.7), we obtain

$$\tau_k^{(2)} = \tfrac{\ln(1+d_k)}{d_k} > \tfrac{1}{1+0.5d_k} \geq \tfrac{1}{1+0.5M_f^{(3)}\lambda_k} = \tau_k^{(3)}.$$

This inequality has shown that the step-size τ_k given by Theorem 4.2.2 satisfies $\tau_k^{(2)} > \tau_k^{(3)}$, where $\tau_k^{(\nu)}$ is a given step-size computed by (4.5) for $\nu = 2$ and 3, respectively. Such a statement confirms that the damped-step Newton method using $\tau_k^{(2)}$ is theoretically better than using $\tau_k^{(3)}$. This result will empirically be confirmed by our experiments in Section 4.4. $\diamond$

Next, we study the full-step Newton scheme derived from (4.1) by setting the step-size $\tau_k = 1$ for all $k \geq 0$ as a full-step. Let $\underline{\sigma}_k := \lambda_{\min}\left(\nabla^2 f(\mathbf{x}^k)\right)$ be the smallest eigenvalue of $\nabla^2 f(\mathbf{x}^k)$. Since $\nabla^2 f(\mathbf{x}^k) \succ 0$, we have $\underline{\sigma}_k > 0$. The following theorem shows a local quadratic convergence of the full-step Newton scheme (4.1) for solving (2.1), whose proof can be found in Appendix A.2.3.

Theorem 4.2.3. Let $\{\mathbf{x}^k\}$ be the sequence generated by the full-step Newton scheme (4.1) by setting the step-size $\tau_k = 1$ for $k \geq 0$. Let d_k and λ_k be defined by (4.3). Then, the following statements hold:

(a) If $\nu = 2$ and the starting point $\mathbf{x}^0$ satisfies $\underline{\sigma}_0^{-1/2}\lambda_0 < \frac{d_2^\star}{M_f}$, then both sequences $\{\underline{\sigma}_k^{-1/2}\lambda_k\}$ and $\{d_k\}$ decrease and quadratically converge to zero, where $d_2^\star \approx 0.12964$.

(b) If $2 < \nu < 3$, and the starting point $\mathbf{x}^0$ satisfies $\underline{\sigma}_0^{-\frac{3-\nu}{2}}\lambda_0 < \frac{1}{M_f}\min\{d_\nu^\star, 0.5\}$, then both sequences $\{\underline{\sigma}_k^{-\frac{3-\nu}{2}}\lambda_k\}$ and $\{d_k\}$ decrease and quadratically converge to zero, where $d_\nu^\star$ is the unique solution of the equation $R_\nu(t) = 2\left(1 - \frac{\nu-2}{2}t\right)^{\frac{4-\nu}{\nu-2}}$ with $R_\nu(\cdot)$ given by (A.7).

(c) If $\nu = 3$ and the starting point $\mathbf{x}^0$ satisfies $\lambda_0 < \frac{1}{2M_f}$, then the sequence $\{\lambda_k\}$ decreases and quadratically converges to zero.

As a consequence, if $\{d_k\}$ locally converges to zero at a quadratic rate, then $\left\{\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{H}_k}\right\}$ also locally converges to zero at a quadratic rate, where $\mathbf{H}_k = \mathbb{I}$, the identity matrix, if $\nu = 2$; and $\mathbf{H}_k = \nabla^2 f(\mathbf{x}^k)^{\nu-2}$ if $2 < \nu \leq 3$. Hence, $\{\mathbf{x}^k\}$ locally converges to $\mathbf{x}_f^\star$, the unique solution of (2.1), at a quadratic rate.

If we combine the results of Theorem 4.2.2 and Theorem 4.2.3, then we can design a two-phase Newton algorithm for solving (2.1) as follows:

- *Phase 1:* Starting from an arbitrary initial point $\mathbf{x}^0 \in \mathrm{dom}(f)$, we perform the damped-step Newton scheme (4.1) until the condition in Theorem 4.2.3 is satisfied.

- *Phase 2:* Using the output $\mathbf{x}^j$ of *Phase 1* as an initial point for the full-step Newton scheme (4.1) with $\tau_k = 1$, and perform this scheme until it achieves an ε-solution $\mathbf{x}^k$ to (2.1).

We also note that the damped-step Newton scheme (4.1) can also achieve a local quadratic convergence as shown in Theorem 4.2.2. Hence, we combine this fact and the above two-phase scheme to derive the Newton algorithm as shown in Algorithm 1 below.

Algorithm 1 (*Newton algorithm for generalized self-concordant minimization*)

Inputs: Choose an arbitrary initial point $\mathbf{x}^0 \in \mathrm{dom}(f)$ and a desired accuracy $\varepsilon > 0$.
Output: An ε-solution $\mathbf{x}^k$ of (2.1).
Initialization: Compute $d_\nu^\star$ according to Theorem 4.2.3 if needed.
For $k = 0, \ldots, k_{\max}$, **perform:**
  Compute the Newton direction n_{nt}^k by solving $\nabla^2 f(\mathbf{x}^k) n_{\mathrm{nt}}^k = -\nabla f(\mathbf{x}^k)$.
  Compute $\lambda_k := \|n_{\mathrm{nt}}^k\|_{\mathbf{x}^k}^*$, and compute $\beta_k := \|n_{\mathrm{nt}}^k\|_2$ if $\nu \neq 3$.
  If $\lambda_k \leq \varepsilon$, then TERMINATE and return $\mathbf{x}^k$.
  If *Phase 2 is used*, then compute $\underline{\sigma}_k = \lambda_{\min}(\nabla^2 f(\mathbf{x}^k))$ if $2 \leq \nu < 3$.
  If *Phase 2 is used* and $(\lambda_k, \underline{\sigma}_k)$ satisfies Theorem 4.2.3, then set $\tau_k := 1$ (**full-step**). Otherwise, compute the step-size τ_k by (4.5) (**damped-step**)
  Update $\mathbf{x}^{k+1} := \mathbf{x}^k + \tau_k n_{\mathrm{nt}}^k$.
End for

Per-iteration complexity: The main step of Algorithm 1 is the solution of the symmetric positive definite linear system (4.2). This system can be solved by using either Cholesky fac-

torization or conjugate gradient methods, which, in the worst case, requires $\mathcal{O}(p^3)$ operations. Computing λ_k requires the inner product $\langle n_{\text{nt}}^k, \nabla f(\mathbf{x}^k)\rangle$ which needs $\mathcal{O}(p)$ operations.

Conceptually, the two-phase option of Algorithm 1 requires the smallest eigenvalue of $\nabla^2 f(\mathbf{x}^k)$ to terminate Phase 1. However, switching from Phase 1 to Phase 2 can be done automatically allowing some tolerance in the step-size τ_k. Indeed, the step-size τ_k given by (4.5) converges to 1 as $k \to \infty$. Hence, when τ_k is closed to 1, e.g., $\tau_k \geq 0.9$, we can automatically set it to 1 and remove the computation of λ_k to reduce the computational time.

In the one-phase option, we can always perform only Phase 1 until achieving an ε-optimal solution as shown in Theorem 4.2.2. Therefore, the per-iteration complexity of Algorithm 1 is $\mathcal{O}(p^3) + \mathcal{O}(p)$ in the worst case. A careful implementation of conjugate gradient methods with a warm-start can significantly reduce this per-iteration computation complexity.

Remark 4.2.1 Inexact Newton methods. We can allow Algorithm 1 to compute the Newton direction n_{nt}^k approximately. In this case, we approximately solve the symmetric positive definite system (4.2). By an appropriate choice of stopping criterion, we can still prove convergence of Algorithm 1 under inexact computation of n_{nt}^k. For instance, the following criterion is often used in inexact Newton methods [27], but defined via the local dual norm of f:

$$\|\nabla^2 f(\mathbf{x}^k) n_{\text{nt}}^k + \nabla f(\mathbf{x}^k)\|_{\mathbf{x}^k}^* \leq \kappa \|\nabla f(\mathbf{x}^k)\|_{\mathbf{x}^k}^*,$$

for a given relaxation parameter $\kappa \in [0, 1)$. This extension can be found in Chapter 5.

4.3 Composite *generalized self-concordant* minimization

Let $f \in \widetilde{\mathcal{F}}_{M_f,\nu}(\text{dom}(f))$, and g be a proper, closed, and convex function. We consider the composite convex minimization problem (2.7) in Section 2.4:

$$F^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \Big\{ F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x}) \Big\}.$$

Note that $\mathrm{dom}(f) := \mathrm{dom}(f) \cap \mathrm{dom}(g)$ may be empty. To make this problem nontrivial, we assume that $\mathrm{dom}(f)$ is nonempty. The optimality condition for (2.7) can be written as follows:

$$0 \in \nabla f(\mathbf{x}^\star) + \partial g(\mathbf{x}^\star). \tag{4.8}$$

Under the qualification condition $0 \in \mathrm{ri}(\mathrm{dom}(g) - \mathrm{dom}(f))$, (4.8) is necessary and sufficient for $\mathbf{x}^\star$ to be an optimal solution of (2.7), where $\mathrm{ri}(\mathcal{X})$ is the relative interior of $\mathcal{X}$.

4.3.1 Existence, uniqueness, and regularity of optimal solutions

Assume that $\nabla^2 f(\mathbf{x})$ is positive definite (i.e., nonsingular) at some point $\mathbf{x} \in \mathrm{dom}(f)$. We prove in the following theorem that problem (2.7) has a unique solution $\mathbf{x}^\star$. The proof can be found in Appendix A.2.4. This theorem can also be considered as a generalization of [74, Theorem 4.1.11] and [102, Lemma 4] in standard self-concordant settings in [74, 102].

Theorem 4.3.1. Suppose that the function f of (2.7) belongs to $\widetilde{\mathcal{F}}_{M_f,\nu}$ with $M_f > 0$ and $\nu \in [2,3]$. Denote by $\sigma_{\min}(\mathbf{x}) := \lambda_{\min}(\nabla^2 f(\mathbf{x}))$ and $\lambda(\mathbf{x}) := \|\nabla f(\mathbf{x}) + \mathbf{v}\|_{\mathbf{x}}^*$ for $\mathbf{x} \in \mathrm{dom}(f)$ and $\mathbf{v} \in \partial g(\mathbf{x})$. Suppose further that there exists $\mathbf{x} \in \mathrm{dom}(f)$ such that $\sigma_{\min}(\mathbf{x}) > 0$ and

$$\lambda(\mathbf{x}) < \frac{2\left[\sigma_{\min}(\mathbf{x})\right]^{\frac{3-\nu}{2}}}{(4-\nu)M_f}.$$

Then, problem (2.7) has a unique solution $\mathbf{x}^\star$ in $\mathrm{dom}(F)$.

Now, we recall a condition such that the solution $\mathbf{x}^\star$ of (2.7) is strongly regular in the following Robinson's sense [90]. We say that the optimal solution $\mathbf{x}^\star$ of (2.7) is *strongly regular* if there exists a neighborhood $\mathcal{U}(\mathbf{0})$ of zero such that for any $\delta \in \mathcal{U}(\mathbf{0})$, the following perturbed problem

$$\min_{\mathbf{x}\in\mathbb{R}^p}\{\langle \nabla f(\mathbf{x}^\star) - \delta, \mathbf{x} - \mathbf{x}^\star\rangle + \tfrac{1}{2}\left\langle \nabla^2 f(\mathbf{x}^\star)(\mathbf{x} - \mathbf{x}^\star), \mathbf{x} - \mathbf{x}^\star\right\rangle + g(\mathbf{x})\}$$

has a unique solution $\mathbf{x}^*(\delta)$, and this solution is Lipschitz continuous on $\mathcal{U}(\mathbf{0})$.

If $\nabla^2 f(\mathbf{x}^\star) \succ 0$, then $\mathbf{x}^\star$ is strongly regular. While the strong regularity of the solution $\mathbf{x}^\star$ requires a weaker condition than $\nabla^2 f(\mathbf{x}^\star) \succ 0$. For further details of the regularity theory, we refer the reader to [90].

4.3.2 Proximal Newton methods

In this section, we develop a proximal Newton algorithm to solve the composite convex minimization problem (2.7) where f is a *generalized self-concordant* function. This problem covers [101, 102] as special cases.

Given $\mathbf{x}^k \in \mathrm{dom}(f)$, we first approximate f at $\mathbf{x}^k$ by the following convex quadratic surrogate:

$$Q_f(\mathbf{x}; \mathbf{x}^k) := f(\mathbf{x}^k) + \left\langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \right\rangle + \tfrac{1}{2} \left\langle \nabla^2 f(\mathbf{x}^k)(\mathbf{x} - \mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \right\rangle.$$

Next, the main step of the proximal Newton method requires to solve the following subproblem, which is the first step of (2.8):

$$\mathbf{z}^k := \operatorname*{argmin}_{\mathbf{x} \in \mathrm{dom}(g)} \Big\{ Q_f(\mathbf{x}; \mathbf{x}^k) + g(\mathbf{x}) \Big\} = \mathrm{prox}_{\nabla^2 f(\mathbf{x}^k)^{-1} g}\Big(\mathbf{x}^k - \nabla^2 f(\mathbf{x}^k)^{-1} \nabla f(\mathbf{x}^k) \Big). \tag{4.9}$$

The optimality condition for this subproblem is the following linear monotone inclusion:

$$0 \in \nabla f(\mathbf{x}^k) + \nabla^2 f(\mathbf{x}^k)(\mathbf{z}^k - \mathbf{x}^k) + \partial g(\mathbf{z}^k). \tag{4.10}$$

Here, we note that $\mathrm{dom}(Q_f(\cdot; \mathbf{x}^k)) = \mathbb{R}^p$. Hence, $\mathrm{dom}(Q_f(\cdot; \mathbf{x}^k) + g(\cdot)) = \mathrm{dom}(g)$. In the setting (2.7), $\mathbf{z}^k$ may not be in $\mathrm{dom}(f)$. Our next step is to update the next iteration $\mathbf{x}^{k+1}$ as

$$\mathbf{x}^{k+1} := \mathbf{x}^k + \tau_k n_{\mathrm{pnt}}^k = (1 - \tau_k)\mathbf{x}^k + \tau_k \mathbf{z}^k, \tag{4.11}$$

where $n_{\mathrm{pnt}}^k := \mathbf{z}^k - \mathbf{x}^k$ is the proximal Newton direction, and $\tau_k \in (0, 1]$ is a given step-size.

Associated with the proximal Newton direction n_{pnt}^k, we define the following proximal Newton decrement and the ℓ_2-norm quantity of n_{pnt}^k as

$$\lambda_k := \|n_{\mathrm{pnt}}^k\|_{\mathbf{x}^k} \qquad \text{and} \qquad \beta_k := \|n_{\mathrm{pnt}}^k\|_2. \tag{4.12}$$

Our first goal is to show that we can explicitly compute the step-size τ_k in (4.11) using λ_k and β_k such that we obtain a descent property for F. This statement is presented in the following theorem, whose proof is deferred to Appendix A.2.5.

Theorem 4.3.2. Let $\{\mathbf{x}^k\}$ be the sequence generated by the proximal Newton scheme (4.11) starting from $\mathbf{x}^0 \in \mathrm{dom}(f)$. If we choose the step-size τ_k as in (4.5) of Theorem 4.2.2, then $\tau_k \in (0, 1]$, $\{\mathbf{x}^k\}$ in $\mathrm{dom}(f)$ and

$$F(\mathbf{x}^{k+1}) \leq F(\mathbf{x}^k) - \Delta_k, \tag{4.13}$$

where $\Delta_k := \lambda_k^2 \tau_k - \omega_\nu(\tau_k d_k)\tau_k^2 \lambda_k^2 > 0$ for $\tau_k > 0$ and d_k as defined in Theorem 4.2.2.

There exists a neighborhood $\mathcal{N}(\mathbf{x}^\star)$ of the unique solution $\mathbf{x}^\star$ of (2.7) such that if we initialize the scheme (4.11) at $\mathbf{x}^0 \in \mathcal{N}(\mathbf{x}^\star) \cap \mathrm{dom}(f)$, then $\{\mathbf{x}^k\}$ quadratically converges to $\mathbf{x}^\star$.

Next, we prove a local quadratic convergence of the full-step proximal Newton method (4.11) with the unit step-size $\tau_k = 1$ for all $k \geq 0$. The proof is given in Appendix A.2.6.

Theorem 4.3.3. Suppose that the sequence $\{\mathbf{x}^k\}$ is generated by (4.11) with full-step, i.e., $\tau_k = 1$ for $k \geq 0$. Let $d_k := d_\nu(\mathbf{x}^k, \mathbf{x}^{k+1})$ be defined by (3.8) and λ_k be defined by (4.12). Then, the following statements hold:

(a) If $\nu = 2$ and the starting point $\mathbf{x}^0$ satisfies $\underline{\sigma}_0^{-1/2}\lambda_0 < d_2^\star/M_f$, then both sequences $\{\underline{\sigma}_k^{-1/2}\lambda_k\}$ and $\{d_2^k\}$ decrease and quadratically converge to zero, where $d_2^\star \approx 0.35482$.

(b) If $2 < \nu < 3$, and the starting point $\mathbf{x}^0$ satisfies $\underline{\sigma}_0^{-\frac{3-\nu}{2}}\lambda_0 < \frac{1}{M_f}\min\{d_\nu^\star, 0.5\}$, then both sequences $\{\underline{\sigma}_k^{-\frac{3-\nu}{2}}\lambda_k\}$ and $\{d_\nu^k\}$ decrease and quadratically converge to zero, where $d_\nu^\star$ is the unique solution to the equation $\frac{R_\nu(t)}{2-\left(1-\frac{\nu-2}{2}t\right)^{\frac{-2}{\nu-2}}} = 2\left(1-\frac{\nu-2}{2}t\right)^{\frac{4-\nu}{\nu-2}}$. in t with $R_\nu(\cdot)$ given in (A.7).

(c) If $\nu = 3$ and the starting point $\mathbf{x}^0$ satisfies $\lambda_0 < \frac{d_3^\star}{M_f}$, then the sequence $\{\lambda_k\}$ decreases and quadratically converges to zero, where $d_3^\star \approx 0.41886$.

As a consequence, if $\{d_k\}$ locally converges to zero at a quadratic rate, then $\left\{\|\mathbf{x}^k - \mathbf{x}^\star\|_{\mathbf{H}_k}\right\}$ also locally converges to zero at a quadratic rate, where $\mathbf{H}_k = \mathbb{I}$, the identity matrix, if $\nu = 2$; and $\mathbf{H}_k = \nabla^2 f(\mathbf{x}^k)^{\nu-2}$ if $2 < \nu \leq 3$. Hence, $\{\mathbf{x}^k\}$ locally converges to $\mathbf{x}^\star$, the unique solution of (2.7), at a quadratic rate.

Similar to Algorithm 1, we can also combine the results of Theorems 4.3.2 and 4.3.3 to design a proximal Newton algorithm for solving (2.7). This algorithm is described in Algorithm 2 below.

Algorithm 2 (*Proximal Newton algorithm for composite generalized self-concordant minimization*)

Inputs: Choose an arbitrary initial point $\mathbf{x}^0 \in \mathrm{dom}(f)$ and a desired accuracy $\varepsilon > 0$.
Output: An ε-solution $\mathbf{x}^k$ of (2.7).
Initialization: Compute $d_\nu^\star$ according to Theorem 4.3.3 if needed.
For $k = 0, \ldots, k_{\max}$, **perform:**
  Compute the proximal Newton direction n_{pnt}^k by solving (4.9).
  Compute $\lambda_k := \|n_{\mathrm{pnt}}^k\|_{\mathbf{x}^k}^*$, and compute $\beta_k := \|n_{\mathrm{pnt}}^k\|_2$ if $\nu \neq 3$.
  If $\lambda_k \leq \varepsilon$, then TERMINATE.
  If *Phase 2 is used*, then compute $\underline{\sigma}_k = \lambda_{\min}(\nabla^2 f(\mathbf{x}^k))$ if $2 \leq \nu < 3$.
  If *Phase 2 is used* and $(\lambda_k, \underline{\sigma}_k)$ satisfies Theorem 4.3.3, then set $\tau_k := 1$ (**full-step**). Otherwise, compute the step-size τ_k by (4.5) (**damped-step**).
  Update $\mathbf{x}^{k+1} := \mathbf{x}^k + \tau_k n_{\mathrm{pnt}}^k$.
End for

Implementation remarks: The main step of Algorithm 2 is the computation of the proximal Newton step n_{pnt}^k, or the trial point $\mathbf{z}^k$ in (4.9). This step requires to solve a composite quadratic convex minimization problem (4.9) with strongly convex objective function. If g is proximally tractable, then we can apply proximal-gradient methods or splitting techniques [3, 4, 73] to solve this problem. We can also combine accelerated proximal-gradient methods with a restarting strategy [36, 41, 79] to accelerate the performance of these algorithms. These methods will be used in our numerical experiments in Section 4.4.

As noticed in Remark 4.2.1, we can also develop an inexact proximal Newton variant for Algorithm 2 by approximately solving the subproblem (4.9). We leave this extension to Chapter 5.

4.4 Numerical experiments

We provide four examples to verify our theoretical results and compare our methods with existing methods in the leterature. Our algorithms are implemented in Matlab 2014b running on a MacBook Pro. Retina, 2.7 GHz Intel Core i5 with 16Gb 1867 MHz DDR3 memory.

4.4.1 Comparison with [109] on regularized logistic regression

In this example, we empirically show that our theory provides a better step-size for logistic regression compared to [109] as theoretically shown in Example 4.1. In addition, our step-size can be used to guarantee a global convergence of Newton method without linesearch. It can also be used as a lower bound for backtracking or forward linesearch to enhance the performance of Algorithm 1.

To illustrate these aspects, we consider the following regularized logistic regression problem:

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \Big\{ f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} \ell(y_i(\mathbf{a}_i^\top \mathbf{x} + \mu)) + \frac{\gamma}{2}\|\mathbf{x}\|_2^2 \Big\}, \tag{4.14}$$

where $\ell(s) = \ln(1 + e^{-s})$ is a logistic loss, μ is a given intercept, $y_i \in \{-1, 1\}$ and $\mathbf{a}_i \in \mathbb{R}^p$ are given as input data for $i = 1, \ldots, n$, and $\gamma > 0$ is a given regularization parameter.

As shown previously in Proposition 3.4.5, $f \in \widetilde{\mathcal{F}}_{M_f^{(3)},3}$ with $M_f^{(3)} = \frac{1}{\sqrt{\gamma}} \max\{\|\mathbf{a}_i\|_2 \mid 1 \le i \le n\}$. On the other hand, $f \in \widetilde{\mathcal{F}}_{M_f^{(2)},2}$ with $M_f^{(2)} := \max\{\|\mathbf{a}_i\|_2 \mid 1 \le i \le n\}$.

We implement Algorithm 1 using two different step-sizes $\tau_k^{(2)} = \frac{\ln(1+d_k)}{d_k}$ and $\tau_k^{(3)} := \frac{1}{1+0.5M_f^{(3)}\lambda_k}$ as suggested by Theorem 4.2.2 for $\nu = 2$ and $\nu = 3$, respectively. We terminate Algorithm 1 if $\|\nabla f(\mathbf{x}^k)\|_2 \le 10^{-8}\max\{1, \|\nabla f(\mathbf{x}^0)\|_2\}$, where $\mathbf{x}^0 = \mathbf{0}$ is an initial point. To solve the linear system (4.2), we apply a conjugate gradient method to avoid computing the inverse $\nabla^2 f(\mathbf{x}^k)^{-1}$ of the Hessian matrix $\nabla^2 f(\mathbf{x}^k)$ in large-scale problems. We also compare our algorithms with the fast gradient method in [74] using the optimal step-size for strongly convex functions, which has the optimal linear convergence rate.

We test all algorithms on a binary classification dataset downloaded from [17][1]. As suggested in [109], we normalize the data such that each row $\mathbf{a}_i$ has $\|\mathbf{a}_i\|_2 = 1$ for $i = 1, \ldots, n$. The parameter is set to $\gamma := 10^{-5}$ as in [109].

The convergence behavior of Algorithm 1 for $\nu = 2$ and $\nu = 3$ is plotted in Figure 4.1 for the `news20` problem. As we can see from this figure that Algorithm 1 with $\nu = 2$ outperforms

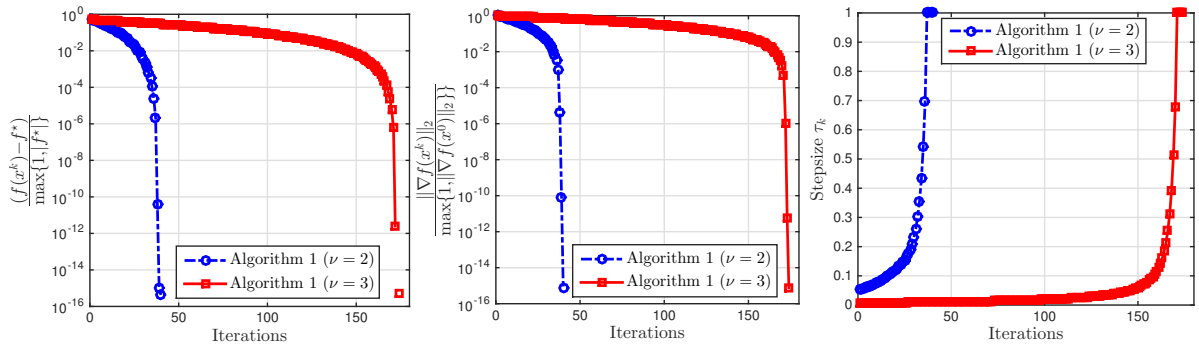

Figure 4.1: The convergence of Algorithm 1 for `news20.binary` (Left: Relative objective residuals, Middle: Relative norms of gradient, and Right: step-sizes).

the case $\nu = 3$. The right-most plot reveals the relative objective residual $\frac{f(\mathbf{x}^k)-f^\star}{\max\{1,|f^\star|\}}$, the middle one shows the relative gradient norm $\frac{\|\nabla f(\mathbf{x}^k)\|_2}{\max\{1,\|\nabla f(\mathbf{x}^0)\|_2\}}$, and the left-most figure displays the step-size $\tau_k^{(2)}$ and $\tau_k^{(3)}$. Note that the step-size $\tau_k^{(3)}$ of Algorithm 1 depends on the regularization parameter γ. If γ is small, then $\tau_k^{(3)}$ is also small. In contrast, the step-size $\tau_k^{(2)}$ of Algorithm 1 is independent of γ.

Our second test is performed on six problems with different sizes. Table 4.1 shows the performance and results of the 3 algorithms: Algorithm 1 with $\nu = 2$, Algorithm 1 with $\nu = 3$, and the fast-gradient method in [74]. Here, n is the number of data points, p is the number of variables, `iter` is the number of iterations, `error` is the training error measured by $\frac{1}{2n}\sum_{i=1}^{n}(1 - \text{sign}(y_i(\mathbf{a}_i^\top \mathbf{x} + \mu)))$, and $f(\mathbf{x}^k)$ is the objective value achieved by these three algorithms.

[1] `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`

Table 4.1: The results of the three algorithms for solving the logistic regression problem (4.14).

Problem			Algorithm 1 ($\nu = 2$)				Algorithm 1 ($\nu = 3$)				Fast gradient method [74]			
Name	p	n	iter	time[s]	$f(\mathbf{x}^k)$	error	iter	time[s]	$f(\mathbf{x}^k)$	error	iter	time[s]	$f(\mathbf{x}^k)$	error
a4a	122	4781	22	0.57	3.250e-01	0.150	177	4.99	3.250e-01	0.150	1396	2.13	3.250e-01	0.150
w4a	300	6760	27	1.14	5.297e-02	0.013	246	8.41	5.297e-02	0.013	863	1.71	5.297e-02	0.013
covtype	54	581012	23	17.22	7.034e-04	0.488	272	235.40	7.034e-04	0.488	1896	318.32	7.034e-04	0.488
rcv1	47236	20242	39	12.45	1.085e-01	0.009	218	60.80	1.085e-01	0.009	366	9.69	1.085e-01	0.009
gisette	5000	6000	40	109.23	1.090e-01	0.008	220	507.03	1.090e-01	0.008	2180	1183.67	1.090e-01	0.008
real-sim	20958	72201	39	22.69	1.287e-01	0.016	218	124.37	1.287e-01	0.016	271	24.74	1.287e-01	0.016
news20	1355191	19954	42	86.47	1.602e-01	0.005	197	420.87	1.602e-01	0.005	623	153.22	1.602e-01	0.005

We observe that our step-size $\tau_k^{(2)}$ using $\nu = 2$ works much better than $\tau_k^{(3)}$ using $\nu = 3$ as in [109]. This confirms the theoretical analysis in Example 4.1. This step-size is useful for parallel and distributed implementation, where evaluating the objective values often requires high computational effort due to communication and data transferring. Note that the computation of the step-size $\tau_k^{(2)}$ in Algorithm 1 only needs $\mathcal{O}(p)$ operations, and do not require to pass over all data points. Algorithm 1 with $\nu = 2$ also works better than the fast gradient method [74] in this experiment, especially for the case $n \gg 1$. Note that the fast gradient method uses the optimal step-size and has a linear convergence rate in this case.

Finally, we show that our step-size $\tau_k^{(2)}$ can be used as a lower bound to enhance a backtracking linesearch procedure in Newton methods. The Armijo linesearch condition is given as

$$f(\mathbf{x}^k + \tau_k n_{\text{nt}}^k) \leq f(\mathbf{x}^k) - c_1 \tau_k \nabla f(\mathbf{x}^k)^\top n_{\text{nt}}^k, \tag{4.15}$$

where $c_1 \in (0, 1)$ is a given constant. Here, we use $c_1 = 10^{-6}$, which is sufficiently small.

- In our backtracking linesearch variant, we search for the best step-size $\tau \in [\tau_k^{(2)}, 1]$. This variant requires to compute $\tau_k^{(2)}$, which needs $\mathcal{O}(p)$ operations.

- In the standard backtracking linesearch routine, we search for the best step-size $\tau \in (0, 1]$.

Both strategies use a bisection section rule as $\tau \leftarrow \tau/2$ starting from $\tau \leftarrow 1$. The results on 3 problems are reported in Table 4.2.

As shown in Table 4.2, using the step-size $\tau_k^{(2)}$ as a lower bound for backtracking linesearch also reduces the number of function evaluations in these three problems. Note that the number of function evaluations depends on the starting point $\mathbf{x}^0$ as well as the factor c_1 in (4.15). If

Table 4.2: The performance and results of the two linesearch variants of Algorithm 1 for solving (4.14).

Problem			Algorithm 1 (Standard linesearch)						Algorithm 1 (Linesearch with $\tau_k^{(2)}$)					
Name	p	n	iter	nfval	time[s]	$\frac{\|\nabla f(\mathbf{x}^k)\|_2}{\|\nabla f(\mathbf{x}^0)\|_2}$	$f(\mathbf{x}^k)$	error	iter	nfval	time[s]	$\frac{\|\nabla f(\mathbf{x}^k)\|_2}{\|\nabla f(\mathbf{x}^0)\|_2}$	$f(\mathbf{x}^k)$	error
covtype	54	581012	25	68	14.99	5.8190e-09	7.034e-04	0.488	14	31	9.89	1.3963e-11	7.034e-04	0.488
rcv1	47236	20242	9	21	1.85	1.3336e-11	1.085e-01	0.009	9	19	1.88	1.3336e-11	1.085e-01	0.009
gisette	5000	6000	8	22	18.28	1.2088e-09	1.090e-01	0.008	8	17	19.68	1.2088e-09	1.090e-01	0.008

we set c_1 too small, then the decrease on f can be small. Otherwise, if we set c_1 too high, then our decrement $c_1\tau_k\nabla f(\mathbf{x}^k)^\top n_{\text{nt}}^k$ may never be achieved, and the linesearch condition fails to hold. If we change the starting point $\mathbf{x}^0$, the number of function evaluations can significantly be increased.

4.4.2 The case $\nu = 2$: Matrix balancing

We consider the following convex optimization problem originated from matrix balancing [21]:

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \Big\{ f(\mathbf{x}) := \sum_{1\le i,j\le p} a_{ij} e^{x_i - x_j} \Big\}, \tag{4.16}$$

where $\mathbf{A} = (a_{ij})_{p\times p}$ is a nonnegative square matrix in $\mathbb{R}^{p\times p}$. Although (4.16) is a smooth unconstrained problem, its objective function f is not strongly convex and does not have Lipschitz gradient. Existing gradient-type methods do not have a theoretical convergence guarantee as well as a rule to compute step-sizes. However, (4.16) is an important problem in scientific computing.

By Proposition 3.3.1 and Corollary 3.4.4, $f \in \widetilde{\mathcal{F}}_{\sqrt{2},2}$. We implement Algorithm 1 and the most recent method proposed in [21] (called Boxed-constrained Newton method (BCNM)) to solve (4.16). Note that [21] is not directly applicable to (4.16), but it solves a regularization of this problem. Since $\nabla^2 f(\mathbf{x})$ is not positive definite, we use a projected conjugate gradient gradient (CG) method to solve the linear system in Algorithm 1. We use an accelerated projected gradient method (FISTA) [4] to solve the subproblem for the method in [21]. We terminate these subsolvers using either a tolerance 10^{-9} or a maximum 200 iterations. For the outer loop, we terminate Algorithm 1 and BCNM using the same stopping criterion: $\delta f_k' := \|\nabla f(\mathbf{x}^k)\|_2 / \max\{1, \|\nabla f(\mathbf{x}^0)\|_2\} \le 10^{-8}$. We choose $\mathbf{x}^0 := \mathbf{0}^p$ as an initial point.

We test both algorithms on several synthetic and real datasets. The synthetic data is generated as in [85] with different structures. The basic matrix $\mathbf{H} = (H_{ij})_{p\times p}$ is an $n \times n$ upper Hessenberg matrix defined as $H_{ij} = 0$ if $j < i - 1$, and $H_{ij} = 1$ otherwise. $\mathbf{H}_1$ differs from $\mathbf{H}$ only in that H_{11} is replaced by p^2; $\mathbf{H}_2$ differs from $\mathbf{H}$ only in that H_{12} is replaced by p^2; and $\mathbf{H}_3 = \mathbf{H} + (p^2 - 1)\mathbb{I}_p$. We use these matrices for A in (4.16). We take $p = 1000$, 5000, 10000, and 15000. We name each problem instance by "`Hdy`", where `H` stands for Hessenberg, and $\mathtt{y} = 10^{-3}p$.

The real data[2] has different structures from different application fields, suggested by [19]. Since we require the matrix $\mathbf{A}$ to be nonnegative, we take $\mathbf{A}_0 := \max\{0, \mathbf{A}\}$ (entry-wise). For the real data, if $\mathbf{A}$ is high ill-conditioned, then we add uniform noise $\mathcal{U}[0, \sigma]$ to $\mathbf{A}$, where $\sigma = 10^{-5} \max_{ij} A_{ij}$.

The final results of both algorithms are reported in Table 4.3, where p is the size of matrix $\mathbf{A}$; `iter/siter` is the maximum number of Newton-type iterations / CG or FISTA iterations; `time[s]` is the computational time in second; $\delta f'_k$ is the relative gradient norm defined above; t_{rat} is the ratio of the computational time between Algorithm 1 and BCNM; and $\delta \mathbf{x}^k$ is the relative difference between $\mathbf{x}^k$ given by Algorithm 1 and BCNM.

As we can see from our experiment, both methods give almost the same result in terms of the objective values $f(\mathbf{x}^k)$ and approximate solutions $\mathbf{x}^k$. Given the same stopping criteria and solution quality, Algorithm 1 outperforms BCNM in all datasets in terms of average computational time, which is specified by $t_{\text{rat}} = \frac{\text{time}_{\text{BCNM}}}{\text{time}_{\text{Alg1}}}$. In particular, for many asymmetric and/or ill-conditioned datasets (e.g., `H2d5`, or `bwm`), Algorithm 1 is approximately from 8 to 17 times faster than BCNM.

4.4.3 The case $\nu \in (2, 3)$: Distance-weighted discrimination regression.

In this example, we test the performance of Algorithm 1 on the distance-weighted discrimination (DWD) problem introduced in [67]. In order to directly use Algorithm 1, we slightly

[2] `https://math.nist.gov/MatrixMarket/searchtool.html`

Table 4.3: Summary of the results of Algorithm 1 and BCNM on 10 synthetic and 30 real problem instances

Datasets		Algorithm 1				BCNM				Comparison	
Name	p	iter/siter	time[s]	$f(\mathbf{x}^k)$	$\delta f'_k$	iter/siter	time[s]	$f(\mathbf{x}^k)$	$\delta f'_k$	t_{rat}	$\delta \mathbf{x}^k$
Synthetic datasets											
H1d1	1000	8/77	0.32	5.07e+05	3.52e-09	8/1028	1.55	5.07e+05	1.82e-10	4.88	4.0e-07
H1d5	5000	7/66	2.54	1.45e+07	2.50e-10	7/648	24.99	1.45e+07	1.73e-10	9.84	3.8e-08
H1d10	10000	7/64	8.74	6.24e+07	8.62e-14	6/461	61.61	6.24e+07	4.82e-09	7.05	7.6e-07
H1d15	15000	7/63	18.63	1.48e+08	3.55e-14	6/395	120.41	1.48e+08	3.66e-10	6.47	2.1e-08
H2d5	5000	7/62	2.53	1.45e+07	7.34e-10	7/640	20.36	1.45e+07	1.88e-10	8.04	1.1e-07
H2d10	10000	7/64	9.16	6.24e+07	2.07e-13	6/467	61.44	6.24e+07	4.75e-09	6.71	7.6e-07
H2d15	15000	7/63	19.66	1.48e+08	3.18e-14	6/395	119.16	1.48e+08	3.52e-10	6.06	1.9e-08
H3d5	5000	4/32	1.34	1.25e+11	1.22e-11	3/15	2.28	1.25e+11	2.47e-11	1.70	6.7e-11
H3d10	10000	4/32	4.52	1.00e+12	1.79e-11	3/14	8.21	1.00e+12	2.29e-11	1.82	2.6e-11
H3d15	15000	4/28	8.72	3.38e+12	1.15e-11	3/12	18.06	3.38e+12	2.59e-10	2.07	4.9e-10
Real datasets											
bcs	10974	4/362	43.95	2.28e+12	2.39e-12	9/438	87.89	2.28e+12	9.83e-09	2.00	2.1e-08
bcs	11948	4/204	31.23	9.30e+12	1.85e-12	14/305	91.19	9.30e+12	8.76e-09	2.92	4.8e-08
bcs	15439	4/36	11.89	1.53e+16	1.21e-12	3/16	19.13	1.53e+16	1.13e-10	1.61	4.4e-11
bcsm	15439	4/28	9.86	2.18e+11	1.98e-12	3/12	18.06	2.18e+11	2.52e-10	1.83	3.3e-10
bwm	2000	4/800	4.06	9.13e+07	2.62e-11	500/1680	72.15	9.13e+07	1.05e-08	17.77	7.3e-09
e40r01	17281	5/178	59.65	9.86e+04	3.49e-12	4/230	92.36	9.86e+04	1.20e-09	1.55	4.6e-08
e40r05	17281	6/279	92.71	1.02e+05	5.09e-13	5/476	170.58	1.02e+05	7.07e-10	1.84	3.0e-08
e40r20	17281	7/489	160.63	1.48e+05	7.86e-14	6/751	278.32	1.48e+05	1.14e-09	1.73	1.6e-09
e40r30	17281	7/492	159.09	1.90e+05	6.21e-14	6/759	260.82	1.90e+05	1.11e-09	1.64	2.0e-09
e40r40	17281	7/486	152.54	2.36e+05	6.09e-14	6/726	247.59	2.36e+05	3.15e-09	1.62	3.8e-09
fid011	16614	4/434	122.21	4.55e+11	7.23e-12	21/465	268.17	4.55e+11	9.56e-09	2.19	3.6e-09
fid019	12005	4/241	37.62	1.69e+10	2.06e-12	13/306	84.94	1.69e+10	9.18e-09	2.26	5.3e-08
fid035	19716	4/261	116.65	2.78e+10	5.24e-12	4/295	164.79	2.78e+10	3.67e-09	1.41	1.1e-08
fidm09	4683	4/685	16.14	1.65e+05	2.60e-12	93/829	67.09	1.65e+05	9.85e-09	4.16	2.5e-08
fidm11	22294	3/222	118.68	4.63e+03	2.93e-09	3/299	178.42	4.63e+03	9.16e-10	1.50	1.3e-07
fidm13	3549	4/667	9.17	8.73e+02	9.86e-14	5/653	9.49	8.73e+02	1.68e-09	1.03	2.7e-08
fidm15	9287	3/231	21.43	2.23e+03	7.48e-09	3/321	32.61	2.23e+03	2.03e-09	1.52	6.7e-07
fidm29	13668	4/451	82.61	1.07e+04	1.51e-12	12/452	135.98	1.07e+04	9.67e-09	1.65	1.8e-08
fidm33	2353	4/397	2.62	9.70e+03	1.31e-12	5/585	3.99	9.70e+03	9.88e-09	1.53	2.4e-08
fidm37	9152	4/483	44.73	1.61e+10	1.23e-11	70/614	212.39	1.61e+10	9.84e-09	4.75	2.3e-08
gre	1107	6/595	1.23	1.07e+03	4.27e-10	6/927	1.93	1.07e+03	4.72e-09	1.57	5.6e-08
lnsp	3937	8/402	7.43	2.56e+12	4.03e-14	7/669	13.60	2.56e+12	3.10e-10	1.83	1.5e-08
mah	1258	8/77	0.45	4.57e+05	1.97e-11	8/1001	3.00	4.57e+05	7.25e-11	6.63	4.7e-09
mem	17758	4/32	14.51	4.57e+02	1.53e-13	3/15	26.57	4.57e+02	1.19e-11	1.83	4.8e-11
mhd	3200	4/165	2.22	5.09e+01	2.39e-14	4/437	6.26	5.09e+01	1.94e-09	2.82	1.7e-07
mhd	4800	4/136	3.97	5.30e+01	4.79e-14	3/423	11.88	5.30e+01	3.30e-09	2.99	1.3e-07
olm	2000	8/640	3.27	2.94e+07	2.05e-15	7/846	4.80	2.94e+07	1.30e-10	1.47	2.7e-09
olm	5000	7/426	11.42	5.41e+08	9.14e-11	6/651	20.75	5.41e+08	4.85e-10	1.82	3.5e-09
ora678	2529	9/898	6.95	3.16e+02	9.95e-11	8/1512	11.92	3.16e+02	8.06e-09	1.71	1.1e-06
pde	2961	6/197	2.56	1.05e+04	5.65e-13	5/311	4.17	1.05e+04	6.14e-10	1.63	8.4e-09

modify the setting in [67] to obtain the following form:

$$f^\star := \min_{\mathbf{x}=[\mathbf{w},\xi,\mu]^\top \in \mathbb{R}^p} \left\{ f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(\mathbf{a}_i^\top \mathbf{w} + \mu y_i + \xi_i)^q} + \mathbf{c}^\top \xi + \frac{1}{2}\left(\gamma_1 \|\mathbf{w}\|_2^2 + \gamma_2 \mu^2 + \gamma_3 \|\xi\|_2^2\right) \right\}, \tag{4.17}$$

where $q > 0$, $\mathbf{a}_i, y_i$ $(i = 1, \ldots, n)$ and $\mathbf{c}$ are given, and $\gamma_s > 0$ $(s = 1, 2, 3)$ are three regularization parameters for $\mathbf{w}$, μ and ξ, respectively. Here, the variable $\mathbf{x}$ consists of the support vector $\mathbf{w}$, the intercept μ, and the slack variable ξ as used in [67]. Here, we penalize these variables by

using least squares terms instead of the ℓ_1-penalty term as in [67]. Note that the setting (4.17) is not just limited to the DWD application above, but can also be used to formulate other practical models such as time optimal path planning problems in robotics [105] if we choose an appropriate parameter q.

Since $\varphi(t) := \frac{1}{t^q} \in \widetilde{\mathcal{F}}_{M_\varphi,\nu}$ with $M_\varphi := \frac{q+2}{\sqrt[(q+2)]{q(q+1)}} n^{\frac{1}{q+2}}$ and $\nu := \frac{2(q+3)}{q+2} \in (2,3)$, using Proposition 3.3.1, we can show that $f \in \widetilde{\mathcal{F}}_{M_f,\nu}$ with $M_f := \frac{q+2}{\sqrt[(q+2)]{q(q+1)}} n^{\frac{1}{q+2}} \max\{\|(\mathbf{a}_i^\top, y_i, \mathbf{e}_i^\top)^\top\|_2^{q/(q+2)} \mid 1 \le i \le n\}$ and the same ν as φ (here, $\mathbf{e}_i$ is the i-th unit vector). Problem (4.17) can be transformed into a second-order cone program [44], and can be solved by interior-point methods. For instance, if we choose $q = 1$, then, by introducing intermediate variables s_i and r_i, we can transform (4.17) into a second-order cone program using the fact that $\frac{1}{r_i} \le s_i$ is equivalent to $\sqrt{(r_i - s_i)^2 + 2^2} \le (r_i + s_i)$.

We implement Algorithm 1 to solve (4.17) and compare it with the interior-point method implemented in commercial software: Mosek. We experienced that Mosek is much faster than other interior-point solvers such as SDPT3 [100] or SDPA [106] in this test. For instance, Mosek is from 52 to 125 times faster than SDPT3 in this example. Hence, we only present the results of Mosek.

We also incorporate Algorithm 1 with a backtracking linesearch using our step-size τ_k (LS with τ_k) as a lower bound. Note that since f does not have a Lipschitz gradient map, we cannot apply gradient-type methods to solve (4.17) due to the lack of a theoretical guarantee.

Since we cannot run Mosek on big data sets, we rather test our algorithms and this interior-point solvers on the 6 small and medium size problems using data from [17][3]. We choose the regularization parameters as $\gamma_1 = \gamma_2 = 10^{-5}$ and $\gamma_3 = 10^{-7}$. Note that if the data set has the size of (n, p), then number of variables in (4.17) becomes $p + n + 1$. Hence, we use a built-in Matlab conjugate gradient solver to compute the Newton direction $n_{\rm nt}^k$. The initial point $\mathbf{x}^0$ is chosen as $\mathbf{w}^0 := \mathbf{0}$, $\mu^0 := 0$ and $\xi^0 := \mathbf{1}$. In our algorithms, we use $\|\nabla f(\mathbf{x}^k)\|_2 \le 10^{-8} \max\{1, \|\nabla f(\mathbf{x}^0)\|_2\}$ as a stopping criterion.

Note that by defining $\gamma_{\min} := \min\{\gamma_1, \gamma_2, \gamma_3\} = 10^{-7} > 0$, the objective function of (4.17) is $\gamma_{\min}$-strongly convex. By Proposition 3.4.1(a), we can cast this function into $\widetilde{\mathcal{F}}_{\hat{M}_f,\hat{\nu}}$ class with

[3] `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`

$\hat{\nu} = 3$ and $\hat{M}_f := \gamma_{\min}^{\frac{-q}{2(q+2)}} M_f$, where M_f is given above. We also implement Algorithm 1 using $\hat{\nu} = 3$ to solve (4.17).

Table 4.4: The performance and results of the four methods for solving the DWD problem (4.17).

Problem			Algorithm 1			Algorithm 1 (LS with τ_k)			Algorithm 1 ($\nu = 3$)			Mosek	
Name	n	p	iter	time[s]	$\frac{\|\nabla f(\mathbf{x}^k)\|_2}{\|\nabla f(\mathbf{x}^0)\|_2}$	iter	time[s]	$\frac{\|\nabla f(\mathbf{x}^k)\|_2}{\|\nabla f(\mathbf{x}^0)\|_2}$	iter	time[s]	$\frac{\|\nabla f(\mathbf{x}^k)\|_2}{\|\nabla f(\mathbf{x}^0)\|_2}$	time[s]	$\frac{\|\nabla f(\mathbf{x}^k)\|_2}{\|\nabla f(\mathbf{x}^0)\|_2}$
$q = 1$													
a1a	1605	119	170	1.35	9.038e-12	13	0.12	4.196e-13	574	5.77	7.031e-14	0.49	1.806e-08
a2a	2265	119	192	2.71	1.661e-13	12	0.15	8.549e-09	633	7.67	8.903e-09	0.50	2.858e-08
a4a	4781	122	247	5.60	1.180e-13	12	0.27	5.380e-10	790	21.06	3.171e-13	0.94	1.740e-08
leu	38	7129	54	2.71	2.214e-10	15	0.58	3.995e-13	193	10.64	5.275e-12	0.72	2.828e-07
w1a	2270	300	169	2.88	9.752e-09	13	0.17	4.968e-09	676	10.44	8.678e-09	0.50	1.561e-08
w2a	3184	300	193	3.32	4.532e-13	13	0.27	1.428e-09	751	15.02	7.662e-14	0.61	1.793e-08
$q = 2$													
a1a	1605	119	166	2.28	6.345e-12	14	0.15	5.185e-13	1372	13.62	3.299e-09	0.48	1.617e-09
a2a	2265	119	186	2.63	3.028e-12	13	0.22	5.015e-09	1484	16.65	5.325e-09	0.56	3.070e-09
a4a	4781	122	235	5.03	8.676e-13	13	0.31	4.347e-10	1764	53.92	2.662e-09	1.25	4.039e-09
leu	38	7129	57	3.08	1.631e-10	16	0.63	2.754e-12	574	39.20	2.076e-12	0.73	6.436e-08
w1a	2270	300	146	2.15	1.311e-12	14	0.22	4.057e-09	1533	27.26	1.110e-09	0.59	1.295e-09
w2a	3184	300	165	3.43	3.397e-09	14	0.29	1.187e-09	1661	30.63	8.004e-09	0.71	1.653e-09

The results and performance of the four algorithms are reported in Table 4.4 for two cases: $q = 1$ and $q = 2$. We can see that Algorithm 1 with $\nu = 2$ outperforms the case $\hat{\nu} = 3$ in terms of iterations. The case $\nu = 2$ is approximately from 3 to 13 times faster than the case $\hat{\nu} = 3$. This is not surprising since $\hat{M}_f$ depends on $\gamma_{\min}$, and it is large since $\gamma_{\min}$ is small. Hence, the step-size $\tau_k^{(3)}$ computed by using $\hat{M}_f$ is smaller than $\tau_k^{(2)}$ computed from M_f as we have seen in the first example. Mosek works really well in this example and it is slightly better than Algorithm 1 with $\nu = 2$. If we combine Algorithm 1 with a backtracking linesearch, then this variant outperforms Mosek. All the algorithms achieve a very high accuracy in terms of the relative norm of the gradient $\frac{\|\nabla f(\mathbf{x}^k)\|_2}{\|\nabla f(\mathbf{x}^0)\|_2}$, which is up to 10^{-8}. We emphasize that our methods are highly parallelizable and their performance can be improved by exploiting this structure as studied in [109] for the logistic case.

4.4.4 The case $\nu = 3$: Portfolio optimization with logarithmic utility functions.

In this example, we aim at verifying Algorithm 2 for solving the composite *generalized self-concordant* minimization problem (2.7) with $\nu = 3$. We illustrate this algorithm on the following portfolio optimization problem with logarithmic utility functions [95] (scaled by a

factor of $\frac{1}{n}$):

$$f^\star = \min_{\mathbf{x}\in\mathbb{R}^p}\{f(\mathbf{x}) := -\sum_{i=1}^{n}\ln(\mathbf{w}_i^\top\mathbf{x}) \mid \mathbf{x}\geq 0, \quad \mathbf{1}^\top\mathbf{x} = 1\}, \tag{4.18}$$

where $\mathbf{w}_i \in \mathbb{R}^p_+$ for $i = 1,\ldots,n$ are given vectors presenting the returns at the i-th period of the assets considered in the portfolio data. More precisely, as indicated in [11], $\mathbf{w}_i$ measures the return as the ratio $w_{ij} = v_{i,j}/v_{i-1,j}$ between the closing prices $v_{i,j}$ and $v_{i-1,j}$ of the stocks on the current day i and on the previous day $i-1$, respectively; $\mathbf{1} \in \mathbb{R}^p$ is a vector of all ones. The aim is to find an optimal strategy to assign the proportion of the assets in order to maximize the expected return among all portfolios.

Note that problem (4.18) can be cast into an online optimization model [48]. The authors in [48] proposed an online Newton method to solve this problem. In this case, the regret of such an online algorithm showing the difference between the objective function of the online counterpart and the objective function of (4.18) converges to zero at a rate of $\frac{1}{\sqrt{n}}$ as $n \to \infty$. If n is relatively small (e.g., $n = 1000$), then the online Newton method does not provide a good approximation to (4.18).

Let $\Delta := \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{x} \geq 0, \quad \mathbf{1}^\top\mathbf{x} = 1\}$ be the standard simplex, and $g(\mathbf{x}) := \delta_\Delta(\mathbf{x})$ be the indicator function of Δ. Then, we can formulate (4.18) into (2.7). The function f defined in (4.18) is (M_f, ν)-generalized self-concordant with $\nu = 3$ and $M_f = 2$.

We implement Algorithm 2 using an accelerated projected gradient method [4, 74] to compute the proximal Newton direction. We also implement the Frank-Wolfe algorithm and its linesearch variant in [37, 54], and a projected gradient method using Barzilai and Borwein's step-size to solve (4.18). We name these algorithms by `FW`, `FW-LS`, and `PG-BB`, respectively.

We emphasize that both `PG-BB` and `FW`-LS do not have a theoretical guarantee when solving (4.18). `FW` has a theoretical guarantee as recently proved in [80], but the complexity bound is rather pessimistic. We terminate all the algorithms using $\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2 \leq \varepsilon\max\{1, \|\mathbf{x}^k\|_2\}$, where $\varepsilon = 10^{-8}$ in Algorithm 2, $\varepsilon = 10^{-6}$ in `PG-BB`, and $\varepsilon = 10^{-4}$ in `FW` and `FW-LS`. We choose different accuracies for these methods due to the limitation of first-order methods for attaining high accuracy solutions in the last three algorithms.

We test these algorithms on two categories of dataset: synthetic and real stock data. For the synthetic data, we generate matrix $\mathbf{W}$ with given price ratios as described above in Matlab.

More precisely, we generate $\mathbf{W} := \text{ones}(n, p) + \mathcal{N}(0, 0.1)$, which allows the closing prices to vary about 10% between two consecutive periods. We test with three instances, where $(n, p) = (1000, 800)$, $(1000, 1000)$, and $(1000, 1200)$, respectively. We name these three datasets by PortfSyn1, PortfSyn2, and PortfSyn3, respectively. For the real data, we download a US stock dataset using an excel tool[4]. This tool gives us the closing prices of the US stock market in a given period of time. We generate three datasets with different sizes using different numbers of stocks from 2005 to 2016 as described in [11]. We pre-processed the data by moving stocks that are empty or lacking of information in the time period we specified. We name these three datasets by Stock1, Stocks2, and Stocks3, respectively.

The results and the performance of the four algorithms are given in Table 4.5. Here, `iter` gives the number of iterations, `time` is the computational time in second, `error` measures the relative difference between the approximate solution $\mathbf{x}^k$ given by the algorithms and the interior-point solution provided by CVX [44] with the high precision configuration (up to 1.8×10^{-12}): $\|\mathbf{x}^k - \mathbf{x}^*_{\text{cvx}}\| / \max\{1, \|\mathbf{x}^*_{\text{cvx}}\|\}$.

Table 4.5: The performance and results of the four algorithms for solving the portfolio optimization problem (4.18).

Problem			Algorithm 2			PG-BB			FW			FW-LS		
Name	n	p	iter	time[s]	error	iter	time[s]	error	iter	time[s]	error	iter	time[s]	error
Synthetic Data														
PortfSyn1	1000	800	6	5.68	2.4e-04	645	3.98	2.3e-04	15530	96.47	2.3e-04	6509	47.88	2.3e-04
PortfSyn2	1000	1000	6	6.96	6.8e-05	1207	11.54	7.5e-05	17201	166.89	1.7e-04	6664	70.15	1.4e-04
PortfSyn3	1000	1200	7	12.91	3.2e-04	959	9.55	3.0e-04	16391	159.28	3.3e-04	5750	64.36	3.2e-04
Real Data														
Stocks1	473	500	8	1.22	7.1e-06	736	1.22	1.9e-06	16274	24.93	7.0e-05	2721	5.28	4.1e-04
Stocks2	625	723	8	3.71	2.7e-05	1544	4.37	8.0e-06	11956	34.35	3.1e-04	2347	9.33	5.2e-04
Stocks3	625	889	10	6.83	5.6e-05	1074	6.54	5.4e-06	13027	52.89	1.7e-04	2096	8.46	7.4e-04

From Table 4.5 we can see that Algorithm 2 has a comparable performance to the first-order methods: `FW-LS` and `PG-BB`. While our method has a rigorous convergence guarantee, these first-order methods remains lacking of a theoretical guarantee. Note that Algorithm 2 and `PG-BB` are faster than the `FW` method and its linesearch variant although the optimal solution $\mathbf{x}^\star$ of this problem is very sparse. We also note that `PG-BB` gives a smaller error to the CVX solution. This CVX solution is not the ground-truth $\mathbf{x}^\star$ but gives a high approximation to $\mathbf{x}^\star$.

[4] `http://www.excelclout.com/historical-stock-prices-in-excel/`

In fact, the CVX solution is dense. Hence, it is not clear if `PG-BB` produces a better solution than other methods.

4.5 Conclusion

We have illustrated our theory by applying it to solve a class of smooth convex minimization problems and its composite setting. We believe that our theory provides an appropriate approach to exploit the curvature of these problems and allows us to compute an explicit step-size in Newton-type methods that have a global convergence guarantee even for non-Lipschitz gradient/Hessian functions. While our theory is still valid for the case $\nu > 3$, we have not found yet a representative application in a high-dimensional space. We therefore limit our consideration to Newton and proximal Newton methods for $\nu \in [2, 3]$, but our key bounds in Section 3.7 remain valid for $\nu > 3$.

CHAPTER 5

Composite convex optimization with global and local inexact oracles

5.1 Introduction

In this chapter we introduce new global and local inexact second-order oracle concepts for a wide class of convex functions in composite optimization. In particular, we consider the following composite convex optimization problem:

$$F^\star = \min_{\mathbf{x}\in\mathbb{R}^p} \Big\{ F(\mathbf{x}) := f(\mathbf{x}) + R(\mathbf{x}) \Big\}, \tag{5.1}$$

where f and R are proper, closed, and convex from $\mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$. It is well-known that problem (5.1) covers various applications in machine learning, statistics, signal and image processing, and control. Very often in applications, f can be considered as a loss or a data fidelity function, while R is referred to as a regularizer that can promote desired structures of solutions. In particular, if R is the indicator of a convex set $\mathcal{X}$, then (5.1) also covers constrained settings.

Optimization methods for solving (5.1) often rely on a so-called "oracle" [70] to query information for generating an approximate solution. However, such an oracle may not be available in practice, but only its approximation can be accessed. We focus on inexact oracles to design numerical methods for solving (5.1). We first deal with a relatively general convex setting of (5.1) by equipping f with a global inexact oracle. Then, we limit our consideration to a class of self-concordant functions and introduce a local second-order inexact oracle.

The rest of this chapter is organized as follows. Section 5.2 introduces the concept of inexact oracle, which consists of both global and local inexact oracles. We then develop some key properties using such inexact oracles. Section 5.3 presents several examples of inexact oracles. Section 5.4 develops proximal Newton-type methods using inexact oracles. We show that the obtained algorithms achieve both global convergence and local convergence from linear

to quadratic rate. We also show that our methods cover some existing inexact methods in the literature as special cases. Section 5.5 shows an application to primal-dual methods, and the last section provides some representative examples to illustrate the theory.

5.2 Inexact second-order oracles

We introduce a global and a local inexact oracle concept for self-concordant function class in convex optimization. Utilizing this new notion, we develop several properties of self-concordant functions that are similar to [75] but using inexact oracles.

5.2.1 Inexact oracles for convex functions

Let f be a convex function with $\mathrm{dom}(f) \subseteq \mathbb{R}^p$. Given three mappings $\tilde{f}(\cdot) \in \mathbb{R}$, $g(\cdot) \in \mathbb{R}^p$, and $H(\cdot) \in \mathcal{S}^p_{++}$ defined on $\mathrm{dom}(f)$, similarly to the definition of local norm based on Hessian, we define the following weighted norm and its dual norm based on $H(\mathbf{x})$ for any $\mathbf{u}$ and $\mathbf{v}$ as

$$\|\mathbf{u}\|_{\mathbf{x}} := \|\mathbf{u}\|_{H(\mathbf{x})} = (\mathbf{u}^\top H(\mathbf{x})\mathbf{u})^{1/2} \text{ and } \|\mathbf{v}\|^*_{\mathbf{x}} := \|\mathbf{v}\|^*_{H(\mathbf{x})} = (\mathbf{v}^\top H(\mathbf{x})^{-1}\mathbf{v})^{1/2}.$$

We still have the relation $\langle \mathbf{u}, \mathbf{v} \rangle \leq \|\mathbf{u}\|_{\mathbf{x}} \|\mathbf{v}\|^*_{H(\mathbf{x})}$ for any $\mathbf{x} \in \mathrm{dom}(f)$.

Next, we introduce the following two types of inexact oracle[1] of f. Following [74], we define a strict convex increasing function $\omega(t) = t - \ln(1+t)$ and its conjugate $\omega_*(\tau) := \omega(-\tau) = -\tau - \ln(1-\tau)$. We also define a function $\tilde{\omega}(u, v) := -uv + \ln(1-u)$ similarly, which will be used later.

Definition 5.1 Global inexact oracle. For a general convex (possibly non-smooth) function f, a triple $(\tilde{f}, g, H)$ is called a (δ_0, δ_1)-global inexact oracle of f with accuracies $\delta_0 \in [0, 1]$ and $\delta_1 \geq 0$, if for any $\mathbf{x} \in \mathrm{dom}(f)$, we have

$$\omega\left((1-\delta_0)\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right) \leq f(\mathbf{y}) - \tilde{f}(\mathbf{x}) - \langle g(\mathbf{x}), \mathbf{y}-\mathbf{x} \rangle \leq \omega_*\left((1+\delta_0)\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right) + \delta_1, \qquad (5.2)$$

[1] As defined in [74]: Oracle is a process of collecting information of the triple $(\tilde{f}, g, H)$. However, for our convenience of presentation, we also call this triple an inexact oracle.

for all $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$, $H(\mathbf{x}) \succ 0$. $\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < \frac{1}{1+\delta_0}$ is required on the right-hand side. Moreover, for any $\mathbf{y} \in \mathbb{R}^p$ such that $\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < \frac{1}{1+\delta_0}$, we have $\mathbf{y} \in \mathrm{dom}(f)$.

This inexact oracle is defined at any $\mathbf{x} \in \mathrm{dom}(f)$. Hence, it is referred to as a global inexact oracle. Here $H(\cdot) \succ 0$ is only required for $\mathbf{x}$ in some level set of $\mathbf{x}^0$, which will be discussed in Section 5.4. Moreover, it does not require differentiability of f. However, for this inexact oracle, if f is twice differentiable, then $\tilde{f}$ gives an approximation to f, g is an approximation to ∇f, and H is an approximation to $\nabla^2 f$. δ_0 and δ_1 are not necessarily depended on $\mathbf{x}$ or $\mathbf{y}$. Clearly, from [74, Theorem 4.1.9], f is a self-concordant function if and only if it admits a $(0,0)$-global inexact oracle, namely $\tilde{f}(\mathbf{x}) = f(\mathbf{x})$, $g(\mathbf{x}) = \nabla f(\mathbf{x})$ and $H(\mathbf{x}) = \nabla^2 f(\mathbf{x})$ by setting $\delta_0 = 0$ and $\delta_1 = 0$.

The second condition "$\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < \frac{1}{1+\delta_0}$ implies $\mathbf{y} \in \mathrm{dom}(f)$" in Definition 5.1 automatically holds if f is self-concordant and $H(\mathbf{x}) = \nabla^2 f(\mathbf{x})$ with $\delta_0 = 0$. This condition is often referred to as Dinkin's ellipsoid in self-concordant functions, see [74]. If $\mathrm{dom}(f) = \mathbb{R}^p$, then this condition holds. However, when $\mathrm{dom}(f) \subset \mathbb{R}^p$ we need to impose this kind of Dinkin's ellipsoid inclusion in our definition of inexact global oracle.

A global inexact oracle will be used to analyze global convergence of our algorithms developed in the next sections. In order to investigate local convergence of Newton-type methods we also require a local inexact second-order oracle in addition to this global inexact one.

Definition 5.2 Local inexact second-order oracle. For a twice differentiable convex function f and a subset $\mathcal{X} \subset \mathrm{dom}(f)$, a triple $(\tilde{f}, g, H)$ is called a $(\delta_0, \delta_1, \delta_2, \delta_3)$-local inexact second-order oracle of f on $\mathcal{X}$ if (5.2) holds and additionally the following approximations for the gradient and for the Hessian maps hold:

$$\begin{cases} \|g(\mathbf{x}) - \nabla f(\mathbf{x})\|_{\mathbf{x}}^* \leq \delta_2, \\ (1-\delta_3)^2 \nabla^2 f(\mathbf{x}) \preceq H(\mathbf{x}) \preceq (1+\delta_3)^2 \nabla^2 f(\mathbf{x}), \end{cases} \tag{5.3}$$

for all $\mathbf{x} \in \mathcal{X}$, where $\delta := (\delta_0, \delta_1, \delta_2, \delta_3) \geq 0$ and $0 \leq \delta_0, \delta_3 < 1$.

In this definition we allow $\delta_2 := \delta_2(\mathbf{x})$ depending on $\mathbf{x} \in \mathcal{X}$. Note that we only require these two conditions in (5.3) in a given subset $\mathcal{X}$ of $\mathrm{dom}(f)$, therefore this inexact oracle is local. Again, we observe that any self-concordant function admits a $(0,0,0,0)$-local oracle.

Remark 5.2.1. As we will show in Lemma 5.2.1 below, the condition (5.2) is also sufficient to deduce that $\|\|g(\mathbf{x}) - \nabla f(\mathbf{x})\|\|_{\mathbf{x}}^* \leq \delta_2$. However, δ_2 will be a function of δ_0 and δ_1, and $\delta_2 = \delta_2(\delta_0, \delta_1) \to 0$ as $\delta_0, \delta_1 \to 0$. Therefore, the first condition (5.3) can be guaranteed from the global inexact oracle in Definition 5.1. In order to make our method more flexible, we use the first condition of (5.3) to define local inexact oracle instead of deriving it from a global inexact oracle as in Lemma 5.2.1.

5.2.2 Properties of global inexact oracle

Convex functions, including self-concordant functions, have many important properties on the function values, gradient and Hessian mappings [74, 75]. These properties are necessary to develop Newton-type methods and interior-point methods. In this subsection, we provide some key properties required for the analysis of our algorithms as well.

The following lemma provides some key properties of our global inexact oracle of f whose proof is given in Appendix A.3.1. Note that these properties hold for general convex functions endowed with such global inexact oracle.

Lemma 5.2.1. Let $(\tilde{f}, g, H)$ be a (δ_0, δ_1)-global inexact oracle of a convex function f as defined in Definition 5.1. Then:

(a) For any $\mathbf{x} \in \mathrm{dom}(f)$, we have

$$\tilde{f}(\mathbf{x}) \leq f(\mathbf{x}) \leq \tilde{f}(\mathbf{x}) + \delta_1. \tag{5.4}$$

(b) The inexact gradient $g(\bar{\mathbf{x}})$ certifies a δ_1-approximate minimizer $\bar{\mathbf{x}} \in \mathrm{dom}(f)$ of f with $f^\star = \inf_{\mathbf{x}} f(\mathbf{x})$. That is, if $\langle g(\bar{\mathbf{x}}), y - \bar{\mathbf{x}} \rangle \geq 0$ for all $y \in \mathrm{dom}(f)$, then

$$f^\star \leq f(\bar{\mathbf{x}}) \leq f^\star + \delta_1.$$

(c) For any $\mathbf{x} \in \mathrm{dom}(f)$, the difference between $g(\mathbf{x})$ and the true (sub)gradient of a convex function f is bounded as

$$\|\|\nabla f(\mathbf{x}) - g(\mathbf{x})\|\|_{\mathbf{x}}^* \leq \delta_2(\delta_0, \delta_1), \tag{5.5}$$

where $\delta_2(\delta_0, \delta_1)$ is the unique nonnegative solution of the equation in δ_2: $\omega\left(\frac{\delta_2}{1+\delta_0}\right) = \delta_1$ (always exists). Moreover, $\delta_2(\delta_0, \delta_1) \to 0$ as $\delta_0, \delta_1 \to 0$.

(d) For any $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$, we have

$$\omega\left(\frac{\|\|g(\mathbf{x})-\nabla f(\mathbf{y})\|\|_{\mathbf{x}}^*}{1+\delta_0}\right) \leq \|\|g(\mathbf{x}) - \nabla f(\mathbf{y})\|\|_{\mathbf{x}}^* \|\|\mathbf{y} - \mathbf{x}\|\|_{\mathbf{x}} + \delta_1, \tag{5.6}$$

5.2.3 Properties of local inexact oracle

We prove some properties of local inexact oracle in the following lemma, whose proof is given in Appendix A.3.2.

Lemma 5.2.2. Let $(\tilde{f}, g, H)$ be a local inexact oracle of a twice differentiable convex function f on $\mathcal{X} \subset \mathrm{dom}(f)$ defined in Definition 5.2. Then, for any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ and $\mathbf{x} \in \mathcal{X}$, we have

$$\begin{aligned} (1-\delta_3)\|\mathbf{u}\|_{\mathbf{x}} &\leq \|\|\mathbf{u}\|\|_{\mathbf{x}} &&\leq (1+\delta_3)\|\mathbf{u}\|_{\mathbf{x}}, \\ \tfrac{1}{1+\delta_3}\|\mathbf{v}\|_{\mathbf{x}}^* &\leq \|\|\mathbf{v}\|\|_{\mathbf{x}}^* &&\leq \tfrac{1}{1-\delta_3}\|\mathbf{v}\|_{\mathbf{x}}^*. \end{aligned} \tag{5.7}$$

If, in addition, f is self-concordant, then for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, we also have:

$$\begin{aligned} \tfrac{(1-\delta_3-\|\|\mathbf{y}-\mathbf{x}\|\|_{\mathbf{x}})^2}{1-\delta_3^2} H(\mathbf{x}) &\preceq H(\mathbf{y}) \preceq \tfrac{1-\delta_3^2}{(1-\delta_3-\|\|\mathbf{y}-\mathbf{x}\|\|_{\mathbf{x}})^2} H(\mathbf{x}) \\ \|\|(\nabla^2 f(\mathbf{x}) - H(\mathbf{x}))\mathbf{v}\|\|_{\mathbf{y}}^* &\leq \tfrac{\delta_3}{(1-\delta_3)(1-\delta_3-\|\|\mathbf{y}-\mathbf{x}\|\|_{\mathbf{x}})} \|\|\mathbf{v}\|\|_{\mathbf{x}}, \end{aligned} \tag{5.8}$$

provided that $\|\|\mathbf{y} - \mathbf{x}\|\|_{\mathbf{x}} < 1 - \delta_3$.

5.3 Examples of inexact oracles

The notion of inexact oracles naturally appears in the context of Fenchel conjugate, barrier smoothing, inexact computations, and many other situations. Below are some examples to show that our definition of inexact oracle makes sense.

5.3.1 Example 1: The generality of new global inexact oracle

We will show in this example that the class of convex functions satisfying Definition 5.1 is larger than the class of standard self-concordant functions [75] and Lipschitz gradient convex functions.

(a) Lipschitz gradient convex functions Let f be a convex function with L_f-Lipschitz gradient on $\mathrm{dom}(f) = \mathbb{R}^p$. Then, $(f, \nabla f, \frac{L_f}{4}\mathbb{I})$ is a (δ_0, δ_1)-global inexact oracle of f in the sense of Definition 5.1 with $\delta_0 = 1$, and $\delta_1 := 0$.

Indeed, we have $0 \leq f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \frac{L_f}{2}\|\mathbf{y} - \mathbf{x}\|^2$ for any $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$. The left-hand side inequality of (5.2) automatically holds since $\delta_0 = 1$.

Now, note that $\frac{\tau^2}{2} \leq \omega_*(\tau)$ for all $\tau \in [0, 1)$. Hence, using $H(\mathbf{x}) = \frac{L_f}{4}\mathbb{I}$, we can show that

$$\frac{L_f}{2}\|\mathbf{y} - \mathbf{x}\|^2 \leq \frac{4\|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}}^2}{2} \leq \omega_*(2\|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}}),$$

provided that $\|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}} < 0.5$. Therefore, we obtain $f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \omega_*(2\|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}})$, which means that the right-hand side of (5.2) holds. The second condition of Definition 5.1 automatically holds since $\mathrm{dom}(f) = \mathbb{R}^p$. This shows that our framework covers the inexact first-order oracle for smooth convex optimization introduced in [28].

(b) The sum of self-concordant and convex functions Let us consider a functions f composed of a self-concordant function f_1 and a convex function (possibly non-smooth) f_2:

$$f(\mathbf{x}) := f_1(\mathbf{x}) + f_2(\mathbf{x}). \tag{5.9}$$

We have $\mathrm{dom}(f) = \mathrm{dom}(f_1) \cap \mathrm{dom}(f_2)$. We assume that for any $g_2(\mathbf{x}) \in \partial f_2(\mathbf{x})$ there exists finite constant $\delta_1 > 0$ such that

$$f_2(\mathbf{y}) - f_2(\mathbf{x}) - \langle g_2(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \delta_1, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathrm{dom}(f),\ \|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}} < 1. \tag{5.10}$$

Then, we can construct a global inexact oracle for f in (5.9) by considering the triple

$$\tilde{f}(\mathbf{x}) := f_1(\mathbf{x}) + f_2(\mathbf{x}),\ g(\mathbf{x}) := \nabla f_1(\mathbf{x}) + g_2(\mathbf{x}) \text{ for any } g_2(\mathbf{x}) \in \partial f_2(\mathbf{x}), \text{ and } H(\mathbf{x}) := \nabla^2 f_1(\mathbf{x}),$$

and consequently $(\tilde{f}, g, H)$ is a $(0, \delta_1)$-global inexact oracle of f in (5.9) by Definition 5.1.

Indeed, since f_1 is self-concordant, we have

$$\omega\left(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right) \leq f_1(\mathbf{y}) - f_1(\mathbf{x}) - \langle \nabla f_1(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \leq \omega_*\left(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right), \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathrm{dom}(f_1),$$

for all $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$, where the right-hand side inequality holds for any $\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < 1$ and $g_2(\mathbf{x}) \in \partial f_2(\mathbf{x})$. Moreover, by convexity of f_2 and (5.10) we also have

$$0 \leq f_2(\mathbf{y}) - f_2(\mathbf{x}) - \langle g_2(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \leq \delta, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathrm{dom}(f).$$

Summing up these two in equalities, we can easily show that the triple $(\tilde{f}, g, H)$ defined above satisfies (5.2) for $(0, \delta)$-inexact global oracle.

As a special case, let us consider the following function:

$$f(\mathbf{x}) = f_1(\mathbf{x}) + \beta f_2(\mathbf{x}), \tag{5.11}$$

where f_1 is a self-concordant barrier, f_2 is an L_2-Lipschitz continuous and convex (possibly nonsmooth) function, and $\beta > 0$ is a given parameter.

Assume that the domain of f, i.e. $\mathrm{dom}(f) = \mathrm{dom}(f_1) \cap \mathrm{dom}(f_2)$ is bounded. Hence, the diameter of $\mathrm{dom}(f)$, $\mathcal{D} = \max_{\mathbf{x},\mathbf{y}\in\mathrm{dom}(f)} \|\mathbf{x}-\mathbf{y}\|$ is finite. In particular, if $\mathrm{dom}(f_1)$ or $\mathrm{dom}(f_2)$ is bounded, then $\mathrm{dom}(f)$ is bounded. Moreover, since f_2 is L_2-Lipschitz continuous, i.e., there exists $L_2 > 0$ such that $|f_2(\mathbf{x}) - f_2(\mathbf{y})| \leq L_2\|\mathbf{x}-\mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f_2)$, we have $\max_{\mathbf{x}\in\mathrm{dom}(f_2)} \|\partial f_2(\mathbf{x})\| \leq L_2$. Using these two facts, we can show that

$$0 \leq f_2(\mathbf{y}) - f_2(\mathbf{x}) - \langle g_2(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \leq L_2\|\mathbf{x}-\mathbf{y}\| + \|g_2(\mathbf{x})\|\|\mathbf{x}-\mathbf{y}\| \leq 2L_2\mathcal{D}, \quad \forall\, x, y \in \mathrm{dom}(f).$$

Therefore, we can construct a global inexact oracle for f in (5.11) with $\delta_1 = 2\beta L_2\mathcal{D}$.

(c) An example with unbounded domain The boundedness of $\mathrm{dom}(f)$ in the previous example is not necessary. For example, let us choose

$$f(x) := f_1(x) + f_2(x), \text{ where } f_1(x) := -\ln(x) \text{ and } f_2(x) := \max\{\delta_1, \delta_1 x\} \text{ for any } \delta_1 > 0.$$

It is clear that $\mathrm{dom}(f) = \{x \in \mathbb{R} \mid x > 0\}$, which is unbounded. If we take $\tilde{f}(x) := f_1(x) + f_2(x)$, $g(x) := f_1'(x) + g_2(x)$, with $g_2(x) \in \partial f_2(x)$, and $H(x) := f_1''(x)$, then it is easy to show that $(\tilde{f}, g, H)$ is a $(0, \delta_1)$-inexact global oracle of f.

Indeed, processing as before, the left-hand side inequality of (5.2) holds for $\delta_0 = 0$. The right-hand side inequality of (5.2) has to hold for $\|\|y - x\|\|_x < \frac{1}{1+\delta_0}$, which induces a bound on y of the form $(y-x)^2/x^2 \leq 1/(1+\delta_0)$, that is for $\delta_0 = 0$ we have $y \leq 2x$. Then, we get

$$f_2(y) - f_2(x) - \langle g_2(x), y - x\rangle \leq \delta_1, \quad \forall\, x, y \in \mathrm{dom}(f), \quad \|\|y - x\|\|_x < 1,$$

which shows that the triple $(\tilde{f}, g, H)$ is a $(0, \delta_1)$-global inexact oracle of the nonsmooth convex function f with unbounded domain.

5.3.2 Example 2: Inexact computation

It is natural to approximate the function value $f(\mathbf{x})$ at $\mathbf{x}$ by $\tilde{f}(\mathbf{x})$ such that $|f(\mathbf{x}) - \hat{f}(\mathbf{x})| \leq \varepsilon$ for some $\varepsilon \geq 0$. In this case, we can define a new inexact oracle as follows. Assume that the triple $(\tilde{f}, g, H)$ satisfies the following inequalities:

$$\begin{cases} |\hat{f}(\mathbf{x}) - f(\mathbf{x})| & \leq \varepsilon, \\ \|\|g(\mathbf{x}) - \nabla f(\mathbf{x})\|\|_{\mathbf{x}}^* & \leq \delta_2, \\ (1-\delta_3)^2 \nabla^2 f(\mathbf{x}) & \preceq H(\mathbf{x}) \preceq (1+\delta_3)^2 \nabla^2 f(\mathbf{x}), \end{cases} \qquad \forall\, \mathbf{x} \in \mathrm{dom}(f). \tag{5.12}$$

where $\varepsilon \geq 0$, $\delta_2 \geq 0$, and $\delta_3 \in [0, 1)$. In addition, H satisfies the condition that for any $\mathbf{x} \in \mathrm{dom}(f)$, if $\|\|\mathbf{y} - \mathbf{x}\|\|_{\mathbf{x}} < \frac{1}{1+2\delta_2+\delta_3}$ for $\mathbf{y} \in \mathbb{R}^p$, then $\mathbf{y} \in \mathrm{dom}(f)$.

Clearly, (5.12) is more restrictive than the oracles defined in Definition 5.1 and Definition 5.2 as we show in Lemma 5.3.1, whose proof can be found in Appendix A.3.3.

Lemma 5.3.1. Let $(\hat{f}, g, H)$ satisfy the condition (5.12). Given $2\delta_2 + \delta_3 < 1$, if we define $\tilde{f}(\mathbf{x}) = \hat{f}(\mathbf{x}) - \varepsilon + \tilde{\omega}(u(\delta_2, \delta_3), v(\delta_2, \delta_3))$, then $(\tilde{f}, g, H)$ is a (δ_0, δ_1)-inexact global oracle of f.

More precisely, we have the following bounds

$$\begin{aligned} f(\mathbf{y}) &\geq \tilde{f}(\mathbf{x}) + \langle g(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \omega\left((1-\delta_0)\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right) \\ f(\mathbf{y}) &\leq \tilde{f}(\mathbf{x}) + \langle g(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \omega_*\left((1+\delta_0)\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right) + \delta_1, \end{aligned} \tag{5.13}$$

where $\delta_0 := 2\delta_2 + \delta_3$, and $\delta_1 := 2\varepsilon - \tilde{\omega}\big(u(\delta_2,\delta_3), v(\delta_2,\delta_3)\big) + \tilde{\omega}\big(\frac{\beta-1}{3}, \frac{4}{3\beta}\big)$ with

$$\begin{aligned} u(\delta_2,\delta_3) &= \tfrac{\delta_2}{(1-\delta_3)^2}\left(2 - \delta_2 - 2\delta_3 - \sqrt{2(1-\delta_2-\delta_3)^2 - \delta_2^2}\right), \\ v(\delta_2,\delta_3) &= \tfrac{\delta_2}{2(1-\delta_3)} - \tfrac{1}{2(1-2\delta_2-\delta_3)}\sqrt{2(1-\delta_2-\delta_3)^2 - \delta_2^2}, \end{aligned}$$

and $\beta \in \left(1, 1 + \frac{2\delta_2}{1+\delta_3}\right)$ being the solution of a quadratic equation (always exists):

$$3(1+\delta_3)\beta^2 + (1 + 3\delta_2 + \delta_3)\beta - 4(1 + 3\delta_2 + \delta_3) = 0. \tag{5.14}$$

5.3.3 Example 3: Fenchel conjugates

Any convex function f can be written as $f(\mathbf{x}) = \sup_{\mathbf{y}\in\mathrm{dom}(f^*)}\{\mathbf{x}^\top\mathbf{y} - f^*(\mathbf{y})\}$, where f^* is the Fenchel conjugate of f. Borrowing this interpretation, we consider the following general convex function

$$f(\mathbf{x}) := \max_{\mathbf{u}\in\mathrm{dom}(\varphi)} \left\{ \left\langle \mathbf{u}, \mathbf{A}^\top\mathbf{x} \right\rangle - \varphi(\mathbf{u}) \right\}, \tag{5.15}$$

where φ is a standard self-concordant function, and $\mathbf{A}$ is a given bounded linear operator. In order to evaluate f and its derivatives, we need to solve the following convex program:

$$u^*(\mathbf{x}) := \mathrm{argmin}_{\mathbf{u}\in\mathrm{dom}(\varphi)}\left\{\varphi(\mathbf{u}) - \left\langle \mathbf{u}, \mathbf{A}^\top\mathbf{x} \right\rangle\right\}, \text{ or equivalent to } \nabla\varphi(u^*(\mathbf{x})) - \mathbf{A}^\top\mathbf{x} = 0. \tag{5.16}$$

Clearly, $u^*(\mathbf{x}) = \nabla\varphi^*(\mathbf{A}^\top\mathbf{x})$. As shown in [75], f defined by (5.15) is convex, twice differentiable, and standard self-concordant on

$$\mathrm{dom}(f) := \{\mathbf{x}\in\mathbb{R}^n \mid \varphi(\mathbf{u}) - \left\langle \mathbf{u}, \mathbf{A}^\top\mathbf{x} \right\rangle \text{ is bounded from below on } \mathrm{dom}(\varphi)\}.$$

The exact gradient and Hessian maps of f are respectively given by

$$\nabla f(\mathbf{x}) = \mathbf{A}u^*(\mathbf{x}) \;\text{ and }\; \nabla^2 f(\mathbf{x}) = \mathbf{A}[\nabla^2\varphi(u^*(\mathbf{x}))]^{-1}\mathbf{A}^\top.$$

However, in many settings, we can only approximate $u^*(\mathbf{x})$ by $\tilde{u}^*(\mathbf{x})$ up to a given accuracy δ in the following sense, which leads to inexact estimations of ∇f and $\nabla^2 f$.

Definition 5.3. Given $\mathbf{x} \in \mathrm{dom}(f)$ and $\delta \geq 0$, we say that $\tilde{u}^*(\mathbf{x}) \in \mathrm{dom}(\varphi)$ is a δ-solution of (5.16) if $\delta(\mathbf{x}) := \|\tilde{u}^*(\mathbf{x}) - u^*(\mathbf{x})\|_{\tilde{u}^*(\mathbf{x})} \leq \delta$, where the local norm is defined w.r.t. $\nabla^2\varphi(\tilde{u}^*(\mathbf{x}))$.

For $\tilde{u}^*(\cdot)$ given in Definition 5.3, we define

$$\tilde{f}(\mathbf{x}) := \langle \tilde{u}^*(\mathbf{x}), \mathbf{A}^\top\mathbf{x}\rangle - \varphi(\tilde{u}^*(\mathbf{x})), \quad g(\mathbf{x}) := \mathbf{A}\tilde{u}^*(\mathbf{x}), \text{ and } H(\mathbf{x}) := \mathbf{A}[\nabla^2\varphi(\tilde{u}^*(\mathbf{x}))]^{-1}\mathbf{A}^\top. \tag{5.17}$$

We show in the following lemma that this triple satisfies our conditions for inexact oracles. In addition, since $u^*(\mathbf{x})$ is unknown, it is impractical to check $\delta(\mathbf{x}) \leq \delta$ directly. We show how to guarantee this condition by approximately checking the optimality condition of (5.16) in the following lemma, whose proof is given in Appendix A.3.4.

Lemma 5.3.2. Let $\tilde{u}^*(\cdot)$ be a δ-approximate solution of $u^*(\cdot)$ in Definition 5.3 and $(\tilde{f}, g, H)$ be given by (5.17). If $\delta \in [0, 0.292]$, then $\hat{f}(\mathbf{x}) := \tilde{f}(\mathbf{x}) - \omega_*\big(\frac{\delta}{1-\delta}\big) + \tilde{\omega}(u(\delta,\delta_3), v(\delta,\delta_3))$ is also a (δ_0, δ_1)-global inexact oracle of f defined in Definition 5.1, where δ_0 and δ_1 are defined similarly as in Lemma 5.3.1. Moreover, we have the following estimates:

$$\|g(\mathbf{x}) - \nabla f(\mathbf{x})\|_{\mathbf{x}}^* \leq \delta, \quad \text{and} \quad (1-\delta_3)^2\nabla^2 f(\mathbf{x}) \preceq H(\mathbf{x}) \preceq (1+\delta_3)^2\nabla^2 f(\mathbf{x}), \tag{5.18}$$

with $\delta_3 := \frac{\delta}{1-\delta}$.

If $\|\nabla\varphi(\tilde{u}^*(\mathbf{x})) - \mathbf{A}^\top\mathbf{x}\|_{\tilde{u}^*(\mathbf{x})}^* \leq \frac{\delta}{1-\delta}$ for $\delta \in (0,1)$, then $\delta(\mathbf{x}) := \|\tilde{u}^*(\mathbf{x}) - u^*(\mathbf{x})\|_{\tilde{u}^*(\mathbf{x})} \leq \delta$.

As an example of (5.15), we consider the following constrained convex optimization problem:

$$\min_{u\in\mathbb{R}^n}\Big\{\phi(\mathbf{u}) \;\text{ s.t. }\; \mathbf{A}\mathbf{u} = \mathbf{b},\; \mathbf{u} \in \mathcal{U}\Big\},$$

where ϕ is a self-concordant function, $\mathbf{A} \in \mathbb{R}^{n\times p}$, $\mathbf{b} \in \mathbb{R}^n$, and $\mathcal{U}$ is a nonempty, closed and convex set in $\mathbb{R}^n$ that admits a self-concordant barrier (see [74, 75]). The dual function defined as

$$f(\mathbf{x}) := \max_{\mathbf{u}\in\mathbb{R}^n}\{\langle \mathbf{x}, \mathbf{A}\mathbf{u} - \mathbf{b}\rangle - \phi(\mathbf{u}) \mid \mathbf{u} \in \mathcal{U}\}$$

is convex and differentiable, but does not have Lipschitz gradient and is not self-concordant in general. Hence, we often smooth it using a self-concordant barrier function $b_{\mathcal{U}}$ of $\mathcal{U}$ to obtain

$$f_\gamma(\mathbf{x}) := \max_{\mathbf{u}} \Big\{ \langle \mathbf{x}, \mathbf{A}\mathbf{u} - \mathbf{b}\rangle - \phi(\mathbf{u}) - \gamma b_{\mathcal{U}}(\mathbf{u}) \Big\}, \tag{5.19}$$

where $\gamma > 0$ is a smoothness parameter. When γ is sufficiently small, $f_\gamma(\mathbf{x})$ can be consider as an approximation of the dual function $f(\mathbf{x})$ at $\mathbf{x}$. Note that in this case $\varphi = \phi + \gamma b_{\mathcal{U}}$. Similar to (5.16), very often, we cannot solve the maximization problem (5.19) exactly to evaluate f and its derivatives. We only obtain an approximate solution $\tilde{u}^*_\gamma(\mathbf{x})$ of its true solution $u^*_\gamma(\mathbf{x})$. In this case, the oracle we obtained via $\tilde{u}^*_\gamma(\cdot)$ generates an inexact oracle for the dual function $f(\cdot)$.

5.4 Inexact proximal-Newton methods using inexact oracles

We utilize our inexact oracles to develop an inexact proximal Newton algorithm (iPNA) for solving (5.1). Our algorithm allows one to use both inexact oracles and inexact computation for the proximal Newton direction. Therefore, it is different from some recent works on this topic such as [40, 66, 109]. [66, 109] only focus on inexact computation of Newton-type directions, while [40] approximates Hessian mappings using quasi-Newton schemes. Our approach combine both aspects but for a more general setting.

5.4.1 iPNA with global inexact oracle: Global convergence

We first describe our inexact proximal-Newton algorithm ((iPNA)) to solve (5.1) under the general setting.

The inexact proximal-Newton scheme Given a global inexact oracle $(\tilde{f}, g, H)$ of f, we first build a quadratic surrogate of f at $\mathbf{x}^k \in \mathrm{dom}(F)$ as

$$\mathcal{Q}(\mathbf{x}; \mathbf{x}^k) := \tilde{f}(\mathbf{x}^k) + \langle g(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle + \tfrac{1}{2}\langle H(\mathbf{x}^k)(\mathbf{x} - \mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle.$$

(iPNA) for solving (5.1) consists of two steps:

$$\left\{ \begin{array}{ll} \mathbf{z}^k & :\approx \mathrm{argmin}_{\mathbf{x}\in\mathbb{R}^p} \big\{ \hat{F}_k(\mathbf{x}) := \mathcal{Q}(\mathbf{x}; \mathbf{x}^k) + R(\mathbf{x}) \big\} \\ \mathbf{x}^{k+1} & := (1-\alpha_k)\mathbf{x}^k + \alpha_k \mathbf{z}^k = \mathbf{x}^k + \alpha_k \mathbf{d}^k \quad \text{with } \mathbf{d}^k := \mathbf{z}^k - \mathbf{x}^k, \end{array} \right. \tag{iPNA}$$

where $\mathbf{d}^k$ is called the inexact-proximal Newton direction, $\alpha_k \in (0, 1]$ is a given stepsize, and the approximation $:\approx$ means that $\mathbf{z}^k$ is computed until satisfying following stoping criterion

$$\|\nu^k\|^*_{\mathbf{x}^k} \leq \delta_4^k \ \|\mathbf{z}^k - \mathbf{x}^k\|_{\mathbf{x}^k}, \text{ where } \nu^k \in g(\mathbf{x}^k) + H(\mathbf{x}^k)(\mathbf{z}^k - \mathbf{x}^k) + \partial R(\mathbf{z}^k). \tag{5.20}$$

Note that one can solve the subproblem in iPNA by any first order scheme, such as FISTA [4], and check criterion (5.20) as described in Appendix A.3.6. Clearly, if $\delta_4^k = 0$, then $\mathbf{z}^k = \bar{\mathbf{z}}^k := \mathrm{argmin}_{\mathbf{x}\in\mathbb{R}^p} \big\{ \hat{F}_k(\mathbf{x}) := \mathcal{Q}(\mathbf{x}; \mathbf{x}^k) + R(\mathbf{x}) \big\}$, the exact solution of the subproblem in iPNA.

Global convergence We now state one of our main results, the global convergence of our inexact proximal-Newton algorithm.

Theorem 5.4.1. Assume that $(\tilde{f}, g, H)$ is a (δ_0^k, δ_1^k)-inexact global oracle of f as in Definition 5.1. Let $\{\mathbf{x}^k\}$ be the sequence computed by iPNA starting from $\mathbf{x}^0$, where α_k is computed as

$$\alpha_k := \frac{1 - \delta_4^k}{(1 + \delta_0^k)(1 + \delta_0^k + (1 - \delta_4^k)\lambda_k)}, \quad \text{with} \quad \lambda_k := \|\mathbf{d}^k\|_{\mathbf{x}^k}. \tag{5.21}$$

Then, the following descent property holds:

$$F(\mathbf{x}^{k+1}) \leq F(\mathbf{x}^k) - \omega\left(\frac{(1 - \delta_4^k)\lambda_k}{1 + \delta_0^k}\right) + \delta_1^k. \tag{5.22}$$

Assume, in addition, that δ_1^k and δ_4^k are chosen such that

$$\sum_{k=0}^{\infty} \delta_1^k < +\infty, \quad \text{and} \quad 0 \leq \delta_4^k \leq \bar{\delta}_4 < 1. \tag{5.23}$$

Then, the inexact Newton decrement $\{\lambda_k\}$ converges to zero as $k \to \infty$. Consequently, the sequence $\{\mathbf{z}^k\}$ also satisfies

$$\lim_{k\to\infty} \inf_{r(\mathbf{z}^k)\in\partial R(\mathbf{z}^k)} \|\!|\nabla f(\mathbf{z}^k) + r(\mathbf{z}^k)|\!\|_{\mathbf{x}^k}^* \equiv \lim_{k\to\infty} \inf_{\nabla F(\mathbf{z}^k)\in\partial F(\mathbf{z}^k)} \|\!|\nabla F(\mathbf{z}^k)|\!\|_{\mathbf{x}^k}^* = 0,$$

which guarantees the optimality condition of (5.1) in the weighted norm $\|\!|\cdot|\!\|_{\mathbf{x}^k}$. In particular, for any given $\varepsilon > 0$, if there exists $L \in [0, +\infty)$ such that $H(\mathbf{x}^k) \preceq L\mathbb{I}$ for all $\mathbf{x}^k \in \mathrm{dom}(F)$ with $\lambda_k \leq \varepsilon$, then we have $\lim_{k\to\infty} \inf_{\nabla F(\mathbf{z}^k)\in\partial F(\mathbf{z}^k)} \|\nabla F(\mathbf{z}^k)\|_2 = 0$.

Proof. From (5.20), we have $\nu^k + H(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{z}^k) - g(\mathbf{x}^k) \in \partial R(\mathbf{z}^k)$. Using this expression and convexity of R, with $r(\mathbf{x}^k) \in \partial R(\mathbf{x}^k)$, we can derive for any $\mathbf{x} \in \mathrm{dom}(F)$ that:

$$\begin{aligned} R(\mathbf{z}^k) &\leq R(\mathbf{x}) + \left\langle r(\mathbf{x}^k), \mathbf{z}^k - \mathbf{x} \right\rangle = R(\mathbf{x}) + \left\langle g(\mathbf{x}^k) + H(\mathbf{x}^k)(\mathbf{z}^k - \mathbf{x}^k) - \nu^k, \mathbf{x} - \mathbf{z}^k \right\rangle \\ &= R(\mathbf{x}) + \left\langle g(\mathbf{x}^k), \mathbf{x} - \mathbf{z}^k \right\rangle + \left\langle H(\mathbf{x}^k)(\mathbf{z}^k - \mathbf{x}^k), \mathbf{x} - \mathbf{z}^k \right\rangle + \left\langle \nu^k, \mathbf{z}^k - \mathbf{x} \right\rangle. \end{aligned}$$

Since $\mathbf{x}^{k+1} := (1 - \alpha_k)\mathbf{x}^k + \alpha_k \mathbf{z}^k$, we can further derive from the last inequality that

$$\begin{aligned} R(\mathbf{x}^{k+1}) &\leq (1 - \alpha_k) R(\mathbf{x}^k) + \alpha_k R(\mathbf{z}^k) \\ &\leq (1-\alpha_k) R(\mathbf{x}^k) + \alpha_k R(\mathbf{x}) + \alpha_k \left[\left\langle g(\mathbf{x}^k), x - \mathbf{z}^k \right\rangle + \alpha_k \left\langle H(\mathbf{x}^k)\mathbf{d}^k, x - \mathbf{z}^k \right\rangle + \left\langle \nu^k, \mathbf{z}^k - \mathbf{x} \right\rangle \right]. \end{aligned}$$

Now, using (5.2), we have

$$\begin{aligned} f(\mathbf{x}^{k+1}) &\overset{(5.2)}{\leq} \tilde{f}(\mathbf{x}^k) + \left\langle g(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \right\rangle + \omega_* \left((1 + \delta_0^k) \|\!| \mathbf{x}^{k+1} - \mathbf{x}^k |\!\|_{\mathbf{x}^k} \right) + \delta_1^k \\ &\leq (1 - \alpha_k) f(\mathbf{x}^k) + \alpha_k \tilde{f}(\mathbf{x}^k) + \alpha_k \left\langle g(\mathbf{x}^k), \mathbf{d}^k \right\rangle + \omega_* \left((1 + \delta_0^k) \alpha_k \|\!| \mathbf{d}^k |\!\|_{\mathbf{x}^k} \right) \\ &\quad + (1 - \alpha_k)\delta_1^k. \end{aligned}$$

Adding these two inequalities and using (5.20), we can show that

$$\begin{aligned} F(\mathbf{x}^{k+1}) &\leq (1-\alpha_k)F(\mathbf{x}^k) + \alpha_k\big[\tilde{f}(\mathbf{x}^k) + \big\langle g(\mathbf{x}^k), \mathbf{x}-\mathbf{x}^k\big\rangle + R(\mathbf{x})\big] - \alpha_k^2\lambda_k^2 \\ &\quad + \omega_*\left((1+\delta_0^k)\alpha_k\|\mathbf{d}^k\|_{\mathbf{x}^k}\right) + (1-\alpha_k)\delta_1^k + \alpha_k\delta_4^k\lambda_k^2 + \alpha_k\left\langle H(\mathbf{x}^k)\mathbf{d}^k - \nu^k, \mathbf{x}-\mathbf{x}^k\right\rangle \\ &\overset{(5.2)}{\leq} (1-\alpha_k)F(\mathbf{x}^k) + \alpha_k F(\mathbf{x}) - \alpha_k^2\lambda_k^2 + \omega_*\left((1+\delta_0^k)\alpha_k\|\mathbf{d}^k\|_{\mathbf{x}^k}\right) \\ &\quad + (1-\alpha_k)\delta_1^k + \alpha_k\delta_4^k\lambda_k^2 + \alpha_k\left\langle H(\mathbf{x}^k)\mathbf{d}^k - \nu^k, \mathbf{x}-\mathbf{x}^k\right\rangle + \delta_1^k. \end{aligned} \tag{5.24}$$

Note that the function $s_2^k(t) := \lambda_k^2(1-\delta_4^k)t - \omega_*((1+\delta_0^k)\lambda_k t)$ achieves a maximum at

$$t_k^* = \frac{1-\delta_4^k}{(1+\delta_0^k)(1+\delta_0^k+(1-\delta_4^k)\lambda_k)},$$

with the optimal value $s_2^k = \omega\left(\frac{(1-\delta_4^k)\lambda_k}{1+\delta_0^k}\right)$. Substituting s_2^k into (5.24), we get

$$F(\mathbf{x}^{k+1}) \leq (1-\alpha_k)F(\mathbf{x}^k) + \alpha_k F(\mathbf{x}^k) - \omega\left(\frac{(1-\delta_4^k)\lambda_k}{1+\delta_0^k}\right) + \delta_1^k + \alpha_k\langle H(\mathbf{x}^k)\mathbf{d}^k - \nu^k, \mathbf{x}-\mathbf{x}^k\rangle \tag{5.25}$$

for all $\mathbf{x} \in \mathrm{dom}(F)$. Substituting now $\mathbf{x} = \mathbf{x}^k$ into this inequality, we obtain (5.22). Since $F(\mathbf{x}^k) \geq F^\star > -\infty$, by induction, we obtain from (5.22) that

$$\sum_{k=0}^{\infty} s_2^k \leq F(\mathbf{x}^0) - F^* + \sum_{k=0}^{\infty}\delta_1^k < +\infty.$$

Hence, we obtain $\sum_{k=0}^{\infty} s_2^k < +\infty$, which yields $\lim_{k\to\infty}\omega\left(\frac{(1-\delta_4^k)\lambda_k}{1+\delta_0^k}\right) = 0$. By the choice of δ_0^k and δ_4^k, and the definition of ω, we have $\lim_{k\to\infty}\lambda_k = 0$.

Further, we can write the optimality condition of (5.20) as $\nu^k = g(\mathbf{x}^k) + H(\mathbf{x}^k)(\mathbf{z}^k - \mathbf{x}^k) + r(\mathbf{z}^k)$ where $r(\mathbf{z}^k) \in \partial R(\mathbf{z}^k)$. Since $\lambda_k := \|\mathbf{x}^k - \mathbf{z}^k\|_{\mathbf{x}_k} \to 0$ as $k \to +\infty$, for k sufficiently large, and $\delta_0 \in [0,1]$, we obtain $\mathbf{z}^k \in \mathrm{dom}(F)$ by Definition 5.1. The above optimality condition leads to

$$\nabla f(\mathbf{z}^k) + r(\mathbf{z}^k) = -H_k(\mathbf{z}^k - \mathbf{x}^k) + (\nabla f(\mathbf{z}^k) - g(\mathbf{x}^k)) + \nu^k.$$

By property of the norm and the definition of our stopping criterion, we have:

$$\begin{aligned}\|\|\nabla f(\mathbf{z}^k) + r(\mathbf{z}^k)\|\|_{\mathbf{x}^k}^* &\leq \|\|H_k(\mathbf{z}^k - \mathbf{x}^k)\|\|_{\mathbf{x}^k}^* + \|\|\nu^k\|\|_{\mathbf{x}^k}^* + \|\|\nabla f(\mathbf{z}^k) - g(\mathbf{x}^k)\|\|_{\mathbf{x}^k}^* \\ &\leq (1+\delta_4^k)\lambda_k + \|\|\nabla f(\mathbf{z}^k) - g(\mathbf{x}^k)\|\|_{\mathbf{x}^k}^*.\end{aligned} \tag{5.26}$$

From (5.6) it follows that $\omega\left(\frac{\|\|g(\mathbf{x}^k)-\nabla f(\mathbf{z}^k)\|\|_{\mathbf{x}^k}^*}{1+\delta_0^k}\right) \leq \|\|g(\mathbf{x}^k) - \nabla f(\mathbf{z}^k)\|\|_{\mathbf{x}^k}^* \lambda_k + \delta_1^k$. This implies that $\left(\frac{1}{1+\delta_0^k} - \lambda_k\right)\|\|g(\mathbf{x}^k) - \nabla f(\mathbf{z}^k)\|\|_{\mathbf{x}^k}^* - \ln\left(1 + \frac{\|\|g(\mathbf{x}^k)-\nabla f(\mathbf{z}^k)\|\|_{\mathbf{x}^k}^*}{1+\delta_0^k}\right) \leq \delta_1^k$. Since $\lim_{k\to\infty}\delta_1^k = \lim_{k\to\infty}\lambda_k = 0$ and $\delta_0^k \in [0,1]$ (Definition 5.1), the last inequality implies that $\lim_{k\to\infty}\|\|g(\mathbf{x}^k) - \nabla f(\mathbf{z}^k)\|\|_{\mathbf{x}^k}^* = 0$. Using this limit together with $\lim_{k\to\infty}\lambda_k = 0$ into (5.26), we can conclude that $\lim_{k\to\infty}\|\|\nabla f(\mathbf{z}^k) + r(\mathbf{z}^k)\|\|_{\mathbf{x}^k}^* = 0$. Consequently, we get our statement

$$\lim_{k\to\infty}\inf_{r(\mathbf{z}^k)\in\partial R(\mathbf{z}^k)}\|\|\nabla f(\mathbf{z}^k) + r(\mathbf{z}^k)\|\|_{\mathbf{x}^k}^* = 0.$$

Since $\partial F(\mathbf{z}^k) = \nabla f(\mathbf{z}^k) + \partial R(\mathbf{z}^k)$, this limit implies $\lim_{k\to\infty}\inf_{\nabla F(\mathbf{z}^k)\in\partial F(\mathbf{z}^k)}\|\|\nabla F(\mathbf{z}^k)\|\|_{\mathbf{x}^k}^* = 0$. Finally, the last statement of this theorem is an immediate consequence of the previous one since $\frac{1}{\sqrt{L}}\|\nabla f(\mathbf{z}^k) + r(\mathbf{z}^k)\|_2 \leq \|\|\nabla f(\mathbf{z}^k) + r(\mathbf{z}^k)\|\|_{\mathbf{x}^k}^*$. □

Remark 5.4.1. Since $\lim_{k\to\infty}\lambda_k = \lim_{k\to\infty}\|\|\mathbf{z}^k - \mathbf{x}^k\|\|_{\mathbf{x}^k} = 0$ in Theorem 5.4.1, we can see that if there exists $L \in [0,+\infty)$ such that $H(\mathbf{x}^k) \preceq L\mathbb{I}$ for all $\mathbf{x}^k \in \mathrm{dom}(F)$ with $\lambda_k \leq \varepsilon$, then $\lim_{k\to\infty}\mathbf{z}^k = \lim_{k\to\infty}\mathbf{x}^k = \mathbf{x}^*$ if these limits exist (at least via a subsequence). Hence, by [92, Theorem 24.4], we have $\inf_{\mathbf{r}^*\in\partial R(\mathbf{x}^*)}\|\nabla f(\mathbf{x}^*) + \mathbf{r}^*\|_2 = 0$.

Remark 5.4.2. To guarantee only the descent property (5.22), one can use a weaker stopping criterion $\langle\nu^k, \mathbf{z}^k - \mathbf{x}^k\rangle \leq \delta_4^k\lambda_k^2$ along with $\delta_4^k < 1$ instead of (5.20) to avoid the inverse computation in $\|\|\nu^k\|\|_{\mathbf{x}^k}^*$. In addition, the proof of (5.22) holds using this criterion even δ_4^k is nonpositive.

5.4.2 iPNA with local inexact oracle: Local convergence

In this subsection, we analyze local convergence of iPNA for solving (5.1) with local inexact oracle under the self-concordance of f. The following lemma is key to our analysis, whose proof is defered to Appendix A.3.5.

Lemma 5.4.2. Let $\{\mathbf{x}^k\}$ be the sequence generated by iPNA algorithm. Then:

$$\begin{aligned}\lambda_{k+1} \quad &\leq \frac{1}{1-\delta_4^{k+1}}\Big\{\delta_2^{k+1} + \frac{1}{(1-\delta_3^{k+1})\big(1-\delta_3^k-\alpha_k\lambda_k\big)}\big[(1-(\delta_3^k)^2)\delta_2^k + (1-(\delta_3^k)^2)\delta_4^k\lambda_k \\ &+ (1-\alpha_k)(3-2(\delta_3^{k+1})^2-(\delta_3^k)^2)\lambda_k + \alpha_k(2+\delta_3^k)\delta_3^k\lambda_k + \frac{\alpha_k^2\lambda_k^2}{1-\delta_3^k-\alpha_k\lambda_k}\big]\Big\},\end{aligned} \tag{5.27}$$

provided that $\alpha_k\lambda_k + \delta_3^k < 1$ and $\delta_4^k < 1$.

Based on the Lemma 5.4.2 and using either full-step or damped-step we can prove local convergence of iPNA in the following theorems.

Theorem 5.4.3. Let $\{\mathbf{x}^k\}$ be the sequence generated by iPNA using a full-step $\alpha_k := 1$ and fix a constant $\rho := 0.8$. Then:

(i) If we choose $0 \leq \delta_3^k, \delta_4^k \leq \frac{1}{100}$, and $0 \leq \delta_2^k \leq \frac{\rho^{k+1}}{50}$ for a given $k \geq 0$, then

$$\lambda_k \leq \frac{\rho^k}{10} \quad \Rightarrow \quad \lambda_{k+1} \leq \frac{\rho^{k+1}}{10}.$$

Consequently, if we choose $\mathbf{x}^0 \in \mathrm{dom}(F)$ such that $\lambda_0 \leq \frac{1}{10}$, then $\{\lambda_k\}$ converges to zero at an R-linear rate with a factor of ρ.

(ii) If we choose δ_2^k, δ_3^k and δ_4^k such that $0 \leq \delta_2^k \leq \frac{\rho^{\frac{k(k+1)}{2}}}{50}$ and $0 \leq \delta_3^k, \delta_4^k \leq \min\{\frac{1}{100}, \frac{\rho^k}{10}\}$, for some $k \geq 0$, then

$$\lambda_k \leq \frac{\rho^{\frac{(k-1)k}{2}}}{10} \quad \Rightarrow \quad \lambda_{k+1} \leq \frac{\rho^{\frac{k(k+1)}{2}}}{10}.$$

Consequently, if we choose $\mathbf{x}^0 \in \mathrm{dom}(F)$ such that $\lambda_0 \leq \frac{1}{10}$, then $\{\lambda_k\}$ converges to zero at an R-superlinear rate.

(iii) If we choose δ_2^k, δ_3^k and δ_4^k such that $0 \leq \delta_2^k \leq \frac{\rho^{2^{k+1}}}{50}$ and $0 \leq \delta_3^k, \delta_4^k \leq \min\{\frac{1}{100}, \frac{\rho^{2^k}}{50}\}$, for some $k \geq 0$, then

$$\lambda_k \leq \min\{\tfrac{1}{10}, \rho^{2^k}\} \quad \Rightarrow \quad \lambda_{k+1} \leq \min\{\tfrac{1}{10}, \tfrac{7}{5}\rho^{2^{k+1}}\}.$$

Consequently, if we choose $\mathbf{x}^0 \in \mathrm{dom}(F)$ such that $\lambda_0 \leq \frac{1}{10}$, then $\{\lambda_k\}$ converges to zero at an R-quadratic rate.

In addition, we have

$$\inf_{\nabla F(\mathbf{z}^k)\in\partial F(\mathbf{z}^k)} \|\nabla F(\mathbf{z}^k)\|^*_{\mathbf{x}^k} \leq \mathcal{O}\left(\max\{\lambda_k, \delta_1^k\}\right).$$

Hence, $\{\inf_{\nabla F(\mathbf{z}^k)\in\partial F(\mathbf{z}^k)} \|\nabla F(\mathbf{z}^k)\|^*_{\mathbf{x}^k}\}$ converges to zero at the same rate of $\{\max\{\lambda_k, \delta_1^k\}\}$.

Proof. (a) For the full-step case, we set $\alpha_k = 1$ in (5.27). If we have $\lambda_k \leq \frac{1}{10}$ for a given $k \geq 0$, then from (5.27), we can derive

$$\lambda_{k+1} \leq \frac{\delta_2^{k+1}}{1-r} + \frac{\delta_2^k}{(1-r)^2(0.9-r)} + \frac{3+r}{(1-r)^2(0.9-r)} r_k\lambda_k + \frac{1}{(1-r)^2(0.9-r)^2}\lambda_k^2 \qquad (5.28)$$

provided that $0 \leq \delta_3^k, \delta_4^k \leq \min\{r, r_k\}$. We note that the left-hand side of (5.28) is an increasing functions of λ_k, r, r_k, and δ_2^k and δ_2^{k+1}. If we impose $\lambda_k \leq \frac{1}{10}$, then by substituting the upper bounds $r_k \leq \frac{1}{100}$, $r = \frac{1}{100}$ and $\delta_2^k, \delta_2^{k+1} \leq \frac{1}{50}$ into (5.28), we can over-estimate it as

$$\lambda_{k+1} \leq \frac{3}{50} \leq \frac{1}{10}.$$

This shows that $\lambda_k \leq \frac{1}{10}$ implies $\lambda_{k+1} \leq \frac{1}{10}$ as long as $\delta_2^k \leq \frac{1}{50}$ and $r_k \leq \frac{1}{100}$, which are satisfied by all the conditions of (i), (ii) and (iii). By choosing δ_2^k, δ_3^k, and δ_4^k as in (i), (ii), and (iii), respectively, then utilizing (5.28), we can directly get the conclusion of (i), (ii), and (iii), respectively. □

Theorem 5.4.4. Let $\{\mathbf{x}^k\}$ be the sequence generated by iPNA using the damped-step (5.21) and fix a constant $\rho := 0.9$. Then:

(i) If we choose $0 \leq \delta_0^k, \delta_3^k, \delta_4^k \leq \frac{1}{100}$, and $0 \leq \delta_2^k \leq \frac{3\rho^{k+1}}{200}$ for all $k \geq 0$, then

$$\lambda_k \leq \tfrac{\rho^k}{10} \quad \Rightarrow \quad \lambda_{k+1} \leq \tfrac{\rho^{k+1}}{10}.$$

Consequently, if we choose $\mathbf{x}^0 \in \mathrm{dom}(F)$ such that $\lambda_0 \leq \frac{1}{10}$, then $\{\lambda_k\}$ converges to zero at an R-linear rate with a factor of ρ.

(ii) If we choose δ_0^k, δ_2^k, δ_3^k, and δ_4^k such that $0 \le \delta_2^k \le \frac{3\rho^{\frac{k(k+1)}{2}}}{200}$ and $0 \le \delta_0^k, \delta_3^k, \delta_4^k \le \frac{\rho^k}{100}$ for some $k \ge 0$, then

$$\lambda_k \le \frac{\rho^{\frac{(k-1)k}{2}}}{10} \quad \Rightarrow \quad \lambda_{k+1} \le \frac{\rho^{\frac{k(k+1)}{2}}}{10}.$$

Consequently, if we choose $\mathbf{x}^0 \in \mathrm{dom}(F)$ such that $\lambda_0 \le \frac{1}{10}$, then $\{\lambda_k\}$ converges to zero at an R-superlinear rate.

(iii) If we choose δ_0^k, δ_2^k, δ_3^k, and δ_4^k such that $0 \le \delta_2^k \le \frac{3\hat{\rho}^{2^{k+1}}}{40}$, and $0 \le \delta_0^k, \delta_3^k, \delta_4^k \le \frac{\hat{\rho}^{2^k}}{100}$ for some $k \ge 0$, where $\hat{\rho} := \frac{11}{25}$, then

$$\lambda_k \le \min\{\tfrac{1}{10}, \hat{\rho}^{2^k}\} \quad \Rightarrow \quad \lambda_{k+1} \le \min\{\tfrac{1}{10}, \tfrac{497}{100}\hat{\rho}^{2^{k+1}}\}.$$

Consequently, if we choose $\mathbf{x}^0 \in \mathrm{dom}(F)$ such that $\lambda_0 \le \frac{1}{10}$, then $\{\lambda_k\}$ converges to zero at an R-quadratic rate.

Moreover, $\{\inf_{\nabla F(\mathbf{z}^k) \in \partial F(\mathbf{z}^k)} \|\nabla F(\mathbf{z}^k)\|_{\mathbf{x}^k}^*\}$ converges to zero at the same rate as $\{\max\{\lambda_k, \delta_1^k\}\}$.

Proof. For the damped-step case, if $0 \le \delta_0^k, \delta_4^k \le t_k$, then α_k defined in (5.21) satisfies

$$1 - \alpha_k \le 1 - \frac{1 - t_k}{(1 + t_k)(1 + t_k + \lambda_k)} = \frac{t_k^2 + 3t_k + (1 + t_k)\lambda_k}{(1 + t_k)(1 + t_k + \lambda_k)} \le 3t_k + \lambda_k. \tag{5.29}$$

Similar to the proof given previously for the full step, if $\lambda_k \le \frac{1}{10}$ for some $k \ge 0$, then from (5.27) and (5.29) we can show that

$$\lambda_{k+1} \le \frac{\delta_2^{k+1}}{1 - t} + \frac{\delta_2^k}{(1 - t)^2(0.9 - t)} + \frac{12 + t}{(1 - t)^2(0.9 - t)} t_k \lambda_k + \frac{3.7 - 3t}{(1 - t)^2(0.9 - t)^2} \lambda_k^2, \tag{5.30}$$

given that $0 \le \delta_3^k, \delta_4^k \le \min\{t, t_k\}$. By taking $t := \frac{1}{100}$ in the above estimate, then after a few elementary calculations, one can shos that $\lambda_k \le \frac{1}{10}$ implies $\lambda_{k+1} \le 0.094 \le \frac{1}{10}$ as long as $\delta_2^k \le 0.015$ and $t_k \le \frac{1}{100}$. These estimates satisfy all the conditions given in (i), (ii), and (iii) of (b). Finally, by choosing δ_0^k, δ_2^k, δ_3^k, and δ_4^k as given in (i), (ii), and (iii), respectively, from (5.30), we can directly get the conclusion of (i), (ii), and (iii), respectively. □

Remark 5.4.3. The last statement of Theorem 5.4.3 and 5.4.4 shows the convergence of subgradient sequence $\{\inf_{\nabla F(\mathbf{z}^k)\in\partial F(\mathbf{z}^k)} \|\nabla F(\mathbf{z}^k)\|^*_{\mathbf{x}^k}\}$ of F. If we choose $\{\delta_1^k\}$ with the same rate as $\{\lambda_k\}$, then $\{\inf_{\nabla F(\mathbf{z}^k)\in\partial F(\mathbf{z}^k)} \|\nabla F(\mathbf{z}^k)\|^*_{\mathbf{x}^k}\}$ converges to zero with the same rate of $\{\lambda_k\}$.

Remark 5.4.4. Due to the complexity of (5.27), we only provide one explicit range of δ_i^k and λ_k by numerically computing their upper bounds. However, we can choose different values than the ones we provide in Theorems 5.4.3 and 5.4.4.

5.4.3 Relationship to other inexact methods

We show that our iPNA covers both inexact Newton methods in [63, 109] and quasi-Newton method in [40].

(a) Inexact proximal-Newton methods In [63], the authors discussed a proximal Newton method where the inexactness lies on the subproblem of computing proximal-Newton direction. This method can be viewed as a special case of our method by choosing $\delta_0^k = \delta_1^k = \delta_2^k = \delta_3^k = 0$ (i.e., no inexact oracle was considered in [63]). In this case, the subproblem (5.20) reduces to the following one by using $\delta_4^k = 1 - \theta_k$:

$$\nu^k \in g^k + h_k(\bar{\mathbf{z}}^k - \mathbf{x}^k) + \partial R(\bar{\mathbf{z}}^k), \tag{5.31}$$

where $\|\nu^k\|^*_{\mathbf{x}^k} \leq (1-\theta_k)\|\bar{\mathbf{z}}^k - \mathbf{x}^k\|_{\mathbf{x}^k}$. For the damped-step proximal Newton method, the corresponding step-size reduces to $\alpha_k = \frac{1-\delta_4^k}{1+(1-\delta_4^k)\lambda_k} = \frac{\theta_k}{1+\theta_k\lambda_k}$, which is the same as the step-size defined in [63]. For global convergence, [63, Theorem 3] is a special case of Theorem 5.4.1 with exact Hessian, gradient, and function values. Furthermore, if we let $\alpha_k = 1$ in Lemma 5.4.2, then we get the same local convergence result as shown in [63, Theorem 2].

(b) Quasi-Newton methods In [40] a quasi-Newton method for self-concordant minimization is proposed based on a curvature-adaptive step-sizes that involve both the inexact and the real Hessian at each loop. However, with our inexact oracle algorithmic setting we can reproduce the same descent and convergence results as in [40]. In particular, we can recover the descent property from [40, Section 4]. In order to avoid the notation ambiguity, we express all quantities appearing in [40] with an additional $\cdot^G$ (e.g. α_k^G means the α_k in [40]), and let $\mathbf{B}_k^{\text{inv}}$ be the

inverse inexact Hessian in [40]. Since f is self-concordant, by using $(\tilde{f}, g, H) = (f, \nabla f, \nabla^2 f)$, we obtain a $(0, 0)$-global inexact oracle as in Definition 5.1. Since [40] only deals with the non-composite case, $R(\mathbf{x}) \equiv 0$ in this case. Therefore, our inexact proximal-Newton scheme (iPNA) is reduced to the inexact Newton scheme:

$$\left\{\begin{array}{ll} \mathbf{z}^k & :\approx \mathbf{x}^k - (\nabla^2 f(\mathbf{x}^k))^{-1}\nabla f(\mathbf{x}^k) \\ \mathbf{x}^{k+1} & := (1-\alpha_k)\mathbf{x}^k + \alpha_k \mathbf{z}^k = \mathbf{x}^k + \alpha_k \mathbf{d}^k, \quad \text{where} \quad \mathbf{d}^k := \mathbf{z}^k - \mathbf{x}^k, \end{array}\right. \tag{iNA}$$

with $\nu^k = \nabla f(\mathbf{x}^k) + \nabla^2 f(\mathbf{x}^k)\mathbf{d}^k$ and $\lambda_k := \|\mathbf{d}^k\|_{\mathbf{x}^k}$. Now, by setting $\mathbf{z}^k = \mathbf{x}^k - \mathbf{B}_k^{\text{inv}}\nabla f(\mathbf{x}^k)$, then by (iNA), $\mathbf{d}^k = -\mathbf{B}_k^{\text{inv}}\nabla f(\mathbf{x}^k)$ is exactly the descent direction $\mathbf{d}_k^G$ in [40]. Moreover, if we set:

$$\delta_4^k := 1 - \alpha_k^G = 1 - \frac{\left\langle \nabla f(\mathbf{x}^k), -\mathbf{d}^k \right\rangle}{\|\mathbf{d}^k\|_{\mathbf{x}^k}^2} = 1 + \frac{\left\langle \nabla f(\mathbf{x}^k), \mathbf{d}^k \right\rangle}{\|\mathbf{d}^k\|_{\mathbf{x}^k}^2} = 1 + \frac{\left\langle \nabla f(\mathbf{x}^k), \mathbf{d}^k \right\rangle}{\lambda_k^2},$$

then from Theorem 5.4.1 we get that

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \omega((1-\delta_4^k)\lambda_k). \tag{5.32}$$

In particular, the proof in Theorem 5.4.1 is reduced to:

$$\begin{aligned} f(\mathbf{x}^{k+1}) &\leq f(\mathbf{x}^k) + \left\langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \right\rangle + \omega_*(\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_{\mathbf{x}^k}) \\ &= f(\mathbf{x}^k) + \alpha_k \left\langle \nabla f(\mathbf{x}^k), \mathbf{d}^k \right\rangle + \omega_*(\alpha_k \lambda_k). \end{aligned} \tag{5.33}$$

Minimizing the right-hand side over the step-size α_k, we obtain the optimal α_k as follows:

$$\alpha_k = \frac{-\left\langle \nabla f(\mathbf{x}^k), \mathbf{d}^k \right\rangle}{\lambda_k(\lambda_k - \langle \nabla f(\mathbf{x}^k), \mathbf{d}^k \rangle)} = \frac{-\left\langle \nabla f(\mathbf{x}^k), \mathbf{d}^k \right\rangle / \lambda_k^2}{\lambda_k(\lambda_k - \langle \nabla f(\mathbf{x}^k), \mathbf{d}^k \rangle)/\lambda_k^2} = \frac{1-\delta_4^k}{1 + (1-\delta_4^k)\lambda_k}.$$

Substituting this α_k into (5.33) we obtain (5.32). Rearranging our step-size we get:

$$\alpha_k = \frac{1-\delta_4^k}{1 + (1-\delta_4^k)\lambda_k} = \frac{\alpha_k^G}{1 + \alpha_k^G \delta_k^G} = t_k^G,$$

which is exactly the step-size used in [40]. For the descent property, the conclusion in [40, Lemma 4.1] is $f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \omega(\eta_k^G)$. Comparing this with our descent (5.32), we have:

$$(1-\delta_4^k)\lambda_k = \frac{\langle \nabla f(\mathbf{x}^k), -\mathbf{d}^k \rangle}{\|\mathbf{d}^k\|_{\mathbf{x}^k}^2} \cdot \|\mathbf{d}^k\|_{\mathbf{x}^k} = \frac{\langle \nabla f(\mathbf{x}^k), \mathbf{B}_k^{-1} \nabla f(\mathbf{x}^k) \rangle}{\lambda_k} = \frac{\rho_k^G}{\delta_k^G} = \eta_k^G.$$

Hence we have recovered the main result of [40, Section 4] by using our oracle setting as defined above and Theorem 5.4.1. Furthermore, [40, Section 5] analysis the convergence behavior of the quasi-Newton method for a $\mathbf{B}_k^{\text{inv}}$ that satisfies the condition $\lambda \mathbb{I} \preceq \mathbf{B}_k^{\text{inv}} \preceq \Lambda \mathbb{I}$ for either $\lambda = \Lambda = 1$ (Gradient Descent) or λ and Λ chosen as in [40, Theorem 5.5] (L-BFGS). Moreover, [40, Section 6] derives similar results for $\mathbf{B}_k^{\text{inv}}$ based on BFGS updates. Since [40, Sections 5 and 6] are just two particular choices for $\mathbf{B}_k^{\text{inv}}$ based on the scheme of [40, Section 4], from previous discussion it follows immediately that we can recover all the local and global convergence results in [40] under the Lipschitz gradient and strong convexity assumptions considered in that paper.

5.5 Application to primal-dual methods

We have shown in Subsection 5.3.3 that inexact oracles of a convex function can be controlled by approximately evaluating its Fenchel conjugate. In this section, we show how to apply this theory to design a primal-dual method for solving composite minimization of a self-concordant objective and a nonsmooth convex regularizer.

We consider the following composite convex problem:

$$G^\star := \min_{\mathbf{y} \in \mathbb{R}^n} \Big\{ G(\mathbf{y}) := \varphi(\mathbf{A}^\top \mathbf{y}) + \psi(\mathbf{y}) \Big\}, \tag{5.34}$$

where $\varphi : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is proper, closed, and convex, and $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a smooth convex function. We assume that ψ is self-concordant, φ is proximally tractable, and $\mathbf{A} \in \mathbb{R}^{n \times p}$ is not diagonal. Problem (5.34) covers many applications in the literature such as image denoising and restoration [8, 16], sparse inverse covariance estimation [39], distance weighted discrimination [67], robust PCA [81], and fused lasso problems [99].

Since φ is nonsmooth, and $\mathbf{A}$ is not diagonal, the proximal operator of $\varphi(\mathbf{A}^\top(\cdot))$ is not tractable. We instead consider the dual problem of (5.34). Using Fenchel conjugate, the dual

problem of (5.34) can be written as

$$F^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \Big\{ F(\mathbf{x}) := f(\mathbf{x}) + R(\mathbf{x}) \equiv \psi^*(\mathbf{A}\mathbf{x}) + \varphi^*(-\mathbf{x}) \Big\}, \tag{5.35}$$

which is exactly of the form (5.1), where $f(\mathbf{x}) := \psi^*(\mathbf{A}\mathbf{x})$ and $R(\mathbf{x}) := \varphi^*(-\mathbf{x})$. Under our assumptions, strong duality holds, i.e. (5.35) is also feasible and $G^\star + F^\star = 0$. The optimality condition of (5.34) and (5.35) becomes

$$\mathbf{A}\mathbf{x}^\star = \nabla\psi(\mathbf{y}^\star) \;\text{ and }\; -\mathbf{x}^\star \in \partial\varphi(\mathbf{A}^\top\mathbf{y}^\star) \;\Leftrightarrow\; 0 \in -\mathbf{A}^\top\mathbf{y}^\star + \partial\varphi^*(-\mathbf{x}^\star). \tag{5.36}$$

Let $y^*(\mathbf{x}) \in \operatorname{argmax}_{\mathbf{y}\in\operatorname{dom}\psi}\{\left\langle \mathbf{x}, \mathbf{A}^\top\mathbf{y} \right\rangle - \psi(\mathbf{y})\}$. Since the optimal set of (5.34) is nonempty and φ is self-concordant, $y^*(\mathbf{x})$ exists and is unique. Moreover, we can show that the exact gradient and Hessian mappings of f are $\nabla f(\mathbf{x}) = \mathbf{A}^\top y^*(\mathbf{x})$ and $\nabla^2 f(\mathbf{x}) = \mathbf{A}^\top\nabla^2\psi(y^*(\mathbf{x}))^{-1}\mathbf{A}$, respectively. However, in practice, we can only evaluate an inexact oracle of f as

$$g(\mathbf{x}) := \mathbf{A}^\top\tilde{y}^*(\mathbf{x}), \quad \text{and} \quad H(\mathbf{x}) := \mathbf{A}^\top\nabla^2\psi(\tilde{y}^*(\mathbf{x}))^{-1}\mathbf{A}, \tag{5.37}$$

that approximate $\nabla f(\mathbf{x})$ and $\nabla^2 f(\mathbf{x})$, respectively, where $\tilde{y}^*(\mathbf{x})$ is an approximate solution of $y^*(\mathbf{x})$ such that $\|\mathbf{A}\mathbf{x} - \nabla\psi(\tilde{y}^*(\mathbf{x}))\|_{\tilde{y}^*(\mathbf{x})} \le \frac{\delta}{1-\delta}$ as suggested by Lemma 5.3.2.

Now, we can develop an inexact primal-dual method to solve (5.34) as follows. Starting from an initial point $\mathbf{x}^0 \in \operatorname{dom}(f)$, at each iteration $k \ge 0$, perform the following steps:

1. Approximately compute $\tilde{y}^*(\mathbf{x}^k)$ from $\|\mathbf{A}\mathbf{x}^k - \nabla\psi(\tilde{y}^*(\mathbf{x}^k))\|^*_{\tilde{y}^*(\mathbf{x}^k)} \le \frac{\delta_k}{1-\delta_k}$, where δ_k is chosen according to Lemma 5.3.2 and Theorem 5.4.1.

2. Form an inexact oracle $g(\mathbf{x}^k) := \mathbf{A}^\top\tilde{y}^*(\mathbf{x}^k)$ and $H(\mathbf{x}^k) := \mathbf{A}^\top\nabla^2\psi(\tilde{y}^*(\mathbf{x}^k))^{-1}\mathbf{A}$ of f at $\mathbf{x}^k$.

3. Approximately solve $\mathbf{z}^k \approx \bar{\mathbf{z}}^k := \arg\min\big\{\tilde{Q}(\mathbf{x};\mathbf{x}^k) + R(\mathbf{x})\big\}$ as in (iPNA).

4. Compute a step-size α_k as in (5.21).

5. Update $\mathbf{x}^{k+1} := (1-\alpha_k)\mathbf{x}^k + \alpha_k\mathbf{z}^k$.

Finally, we recover an approximate solution $\mathbf{y}^k := \tilde{y}^*(\mathbf{x}^k)$ of $\mathbf{y}^\star$ for (5.34).

The following lemma shows that $\tilde{y}^*(\mathbf{x}^k)$ is indeed an approximate solution of (5.34).

Lemma 5.5.1. Let $\{(\mathbf{z}^k, \mathbf{y}^k)\}$ be the sequence generated by our primal-dual scheme above. Then

$$\|A\mathbf{z}^k - \nabla\psi(\mathbf{y}^k)\|^*_{\mathbf{y}^k} \leq \frac{\delta_k}{1-\delta_k} + \lambda_k \;\text{ and }\; \mathbf{r}^k \in \mathbf{A}^\top\mathbf{y}^k - \partial\varphi^*(-\mathbf{z}^k) \text{ with } \|\mathbf{r}^k\|^*_{\mathbf{x}^k} \leq (1+\delta_4^k)\lambda_k. \quad (5.38)$$

Consequently, if we compute λ_k and choose δ_k such that $\delta_k + \lambda_k \leq \frac{\varepsilon}{1+\varepsilon}$ and $\lambda_k \leq \frac{\varepsilon}{2}$, then $(\mathbf{z}^k, \mathbf{y}^k)$ is an ε-solution of the primal problem (5.34) and its dual (5.35), i.e., $\|\mathbf{A}\mathbf{z}^k - \nabla\psi(\mathbf{y}^k)\|^*_{\mathbf{y}^k} \leq \varepsilon$ and $\|\mathbf{r}^k\|^*_{\mathbf{x}^k} \leq \varepsilon$ such that $\mathbf{r}^k \in \mathbf{A}^\top\mathbf{y}^k - \partial\varphi^*(-\mathbf{z}^k)$.

Proof. Since we define $\mathbf{y}^k := \tilde{y}^*(\mathbf{x}^k)$, from (iPNA) and (5.37), we have

$$\nu^k \in \mathbf{A}^\top\mathbf{y}^k + \mathbf{A}^\top\nabla^2\psi(\mathbf{y}^k)^{-1}\mathbf{A}(\mathbf{z}^k - \mathbf{x}^k) - \partial\varphi^*(-\mathbf{z}^k).$$

Let us define $\mathbf{r}^k := \nu^k - \mathbf{A}^\top\nabla^2\psi(\mathbf{y}^k)^{-1}\mathbf{A}(\mathbf{z}^k - \mathbf{x}^k)$. Then, the last condition leads to $\mathbf{r}^k \in \mathbf{A}^\top\mathbf{y}^k - \partial\varphi^*(-\mathbf{z}^k)$. Hence, we can estimate $\|\mathbf{r}^k\|_{\mathbf{x}^k}$ as follows:

$$\|\mathbf{r}^k\|^*_{\mathbf{x}^k} \leq \|\nu^k\|^*_{\mathbf{x}^k} + \|\mathbf{A}^\top\nabla^2\psi(y^k)^{-1}\mathbf{A}(\mathbf{z}^k - \mathbf{x}^k)\|^*_{\mathbf{x}^k} \leq \delta_4^k\lambda_k + \lambda_k = (1+\delta_4^k)\lambda_k.$$

Therefore, we get the second part of (5.38).

Note that $\|\mathbf{A}\mathbf{x}^k - \nabla\psi(\mathbf{y}^k)\|^*_{\mathbf{y}^k} \leq \frac{\delta_k}{1-\delta_k}$. Hence, we can show that $\|\mathbf{A}\mathbf{z}^k - \nabla\psi(\mathbf{y}^k)\|^*_{\mathbf{y}^k} \leq \frac{\delta_k}{1-\delta_k} + \|\mathbf{A}(\mathbf{z}^k - \mathbf{x}^k)\|^*_{\mathbf{y}^k} = \frac{\delta_k}{1-\delta_k} + \|\mathbf{z}^k - \mathbf{x}^k\|_{\mathbf{x}^k} = \frac{\delta_k}{1-\delta_k} + \lambda_k$, which proves the first part of (5.38). The rest of this lemma is a direct consequence of (5.38). □

Note that both $\|A\mathbf{z}^k - \nabla\psi(\mathbf{y}^k)\|^*_{\mathbf{y}^k}$ and $\|\mathbf{r}^k\|^*_{\mathbf{x}^k}$ are controlled by λ_k. By Theorem 5.4.1, we have $\lim_{k\to\infty}\lambda_k = 0$. Consequently, $\lim_{k\to\infty}\|\mathbf{A}\mathbf{z}^k - \nabla\psi(\mathbf{y}^k)\|^*_{\mathbf{y}^k} = \lim_{k\to\infty}\|\mathbf{r}^k\|^*_{\mathbf{x}^k} = 0$. Hence, we can say that $(\mathbf{z}^k, \mathbf{y}^k)$ converges to the solution of (5.34)-(5.35). By Theorem 5.4.3 and 5.4.4, we can also prove locally linear/superlinear/quadratic convergence rates of the two residual sequences $\{\|\mathbf{A}\mathbf{z}^k - \nabla\psi(\mathbf{y}^k)\|^*_{\mathbf{y}^k}\}$ and $\{\|\mathbf{r}^k\|^*_{\mathbf{x}^k}\}$.

5.6 Preliminary numerical experiments

We provide two numerical examples to verify several aspects of our theoretical results and compare our algorithms with some existing methods. These algorithms are implemented in Matlab 2018a running on a Lenovo Thinkpad 2.60GHz Intel Core i7 Laptop with 8Gb memory.

5.6.1 Composite Log-barrier+ℓ_p-norm models

This example aims at studying several theoretical aspects of our theory developed in the previous sections. For this purpose, we consider the following composite log-barrier+ℓ_p-norm model as a special case of (5.34):

$$G^\star := \min_{\mathbf{y}\in\mathbb{R}^p} \Big\{ G(\mathbf{y}) := \varphi(\mathbf{A}^\top\mathbf{y}) + \psi(\mathbf{y}) \Big\}, \tag{5.39}$$

where $\varphi : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ is a proper, closed, and convex function, $\psi(\mathbf{y}) := -\sum_{i=1}^m w_i \ln(d_i - \mathbf{c}_i^\top\mathbf{y})$, which can be viewed as a barrier function of a polyhedron $\mathcal{P} := \big\{\mathbf{y}\in\mathbb{R}^p \mid \mathbf{C}^\top\mathbf{y} \le \mathbf{d}\big\}$, $\mathbf{A}\in\mathbb{R}^{p\times n}$, and $\mathbf{w}\in\mathbb{R}^m_+$ is a weight vector. In our experiments, we focus on the case φ is a finite sum of ℓ_p-norms.

Problem (5.39) has concrete applications including solving systems of linear equations and inequalities [43], Poisson image processing [47, 61], and robust optimization [7].

Unlike several existing models, the linear operator $\mathbf{A}$ in (5.39) is composited into a nonsmooth term φ, which makes first-order methods to be intractable. Instead of solving the primal problem (5.39) directly, we consider its dual formulation as in Section 5.5:

$$F^\star := \min_{\mathbf{x}} \Big\{ F(\mathbf{x}) := \varphi^*(-\mathbf{x}) + \psi^*(\mathbf{A}\mathbf{x}) \Big\}, \tag{5.40}$$

where φ^* and ψ^* are the Fenchel conjugates of φ and ψ, respectively. Clearly, since ψ is smooth, one can evaluate its conjugate ψ^* as well as the derivatives of ψ^* by solving

$$\psi^*(\mathbf{A}\mathbf{x}) := \max_{\mathbf{u}\in\mathbb{R}^n} \Big\{ h(\mathbf{u}) := \langle \mathbf{A}\mathbf{x}, \mathbf{u}\rangle + \sum_{i=1}^m w_i \ln(d_i - \mathbf{c}_i^\top\mathbf{u}) \Big\}. \tag{5.41}$$

Let us denote by $u^*(\mathbf{x})$ the solution of this problem. Since the underlying function is self-concordant, one can apply Newton method to compute $u^*(\mathbf{x})$ [74]. However, we can only approximately compute $u^*(\mathbf{x})$, which leads to inexact oracle of ψ^*. Hence, our theory, in particular, the results developed in Section 5.5 can be applied to solve (5.41) inexactly.

5.6.1.1 The effect of inexactness to the convergence of iPNA

First, we show how the accuracy of inexact oracles affects the overall convergence of iPNA when solving (5.40). As indicated by Theorems 5.4.1, 5.4.3, and 5.4.4, iPNA can achieve different local convergence rates, or can be diverged. In this experiment, we plan to analyze the convergence or divergence of iPNA under different accuracy levels of inexact oracles.

In this experiment, we generate data according to Subsection 5.6.1.2 below but using $\mathbf{A} := \texttt{rand}(p, 0.1p)$, where $p = 500$. For configuration of the experiment, we set the maximum number of iterations at 100 as a safeguard, but also terminate the algorithm if $\lambda_k \leq 10^{-9}$ and the relative objective value satisfies $F(\mathbf{x}^k) - F^\star \leq \varepsilon \max\{1, |F^\star|\}$, where $\varepsilon = 10^{-11}$ for the linear convergence rate, and $\varepsilon = 10^{-12}$ for the quadratic convergence rate, respectively. The optimal value $F^\star$ is computed by running SDPT3 up to high accuracy. The global convergence of iPNA is reflected in Figure 5.1, where the sum of errors $\sum_{k=0}^{k_{\max}} \delta_1^k$ presented in (5.23) of Theorem 5.4.1 is given on the left-most plot, the proximal Newton decrement λ_k is in the middle plot, and the relative objective residual is on the right-most plot.

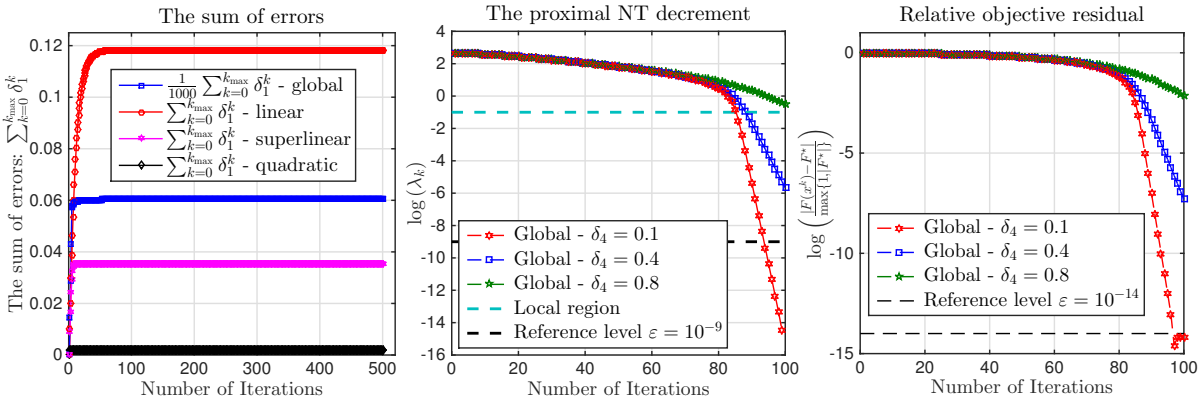

Figure 5.1: Global convergence behavior of iPNA in Theorem 5.4.1.

The left-most plot shows the sum of errors δ_1 arisen from δ, the accuracy of the conjugate function ψ^* as shown in Definition 5.3. More precisely, if δ is chosen according to Lemma 5.3.2

to achieve linear, superlinear and quadratic convergence as in Theorem 5.4.3, then, the sum of errors $\sum_{k=0}^{k_{\max}} \delta_1^k$ rendering from Theorem 5.4.1 is given in the left-most plot of Figure 5.1. The blue line is just the sum of errors when iPNA is convergent as required in Theorem 5.4.1.

The middle plot reveals the inexact proximal Newton decrement λ_k computed from different accuracy levels of the subproblem in (5.20). Clearly, the more accurate in (5.20) is given, the faster convergence in λ_k is achieved. The right-most plot provides the convergence of the relative objective residuals under different accuracy level δ_4 of the subproblem.

Our next step is to verify the local convergence represented in Theorem 5.4.3, and how inexact oracles affect the local convergence of iPNA. By choosing different values of δ we obtain different levels of inexact oracles in ψ^*. Figure 5.2, Figure 5.3, and Figure 5.4 show an R-linear, R-superlinear, and R-quadratic convergence rate of iPNA, respectively. Here, the reference level ε representing the desired accuracy of the solution is given in the legend of these figures.

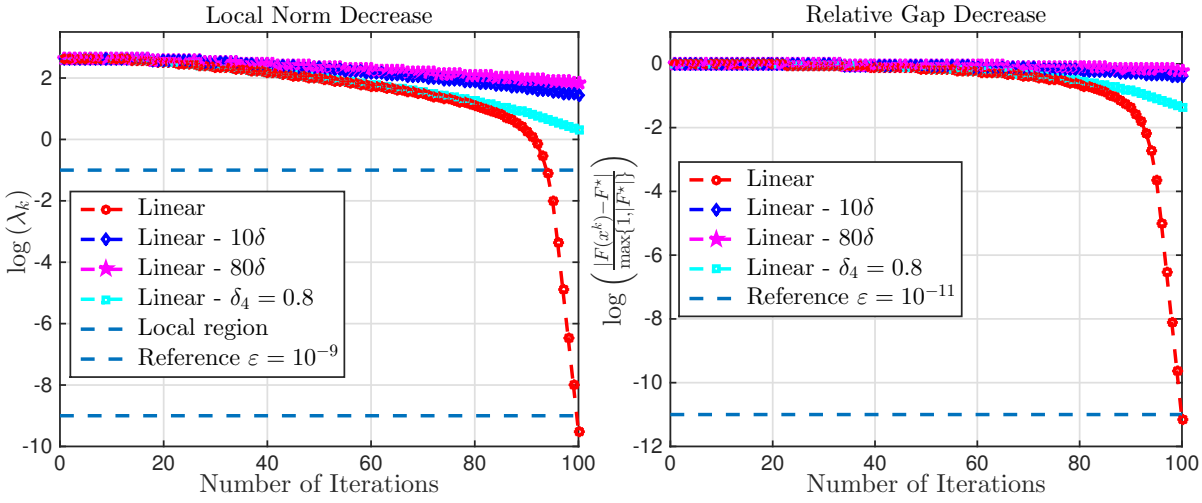

Figure 5.2: The local linear convergence of iPNA under the effect of inexact oracles.

As we can see from Figure 5.2, if we choose the parameters as in Theorem 5.4.3, 5.4.4 (i) to reflect a local linear convergence rate, we observe a sublinear convergence in a few dozen of iterations due to slow global convergence rate, but we can see a fast local convergence at the last iterations. Notice that this convergence rate is even better than linear in terms of λ_k or the relative objective residuals, since we only use the quantity δ of conjugate subproblem to measure derivatives accuracy via Lemma 5.3.2. δ is controlled by the most accurate tolerance among δ_0, δ_2 and δ_3 in Theorem 5.4.3, 5.4.4, which gives the convergence rate better than linear.

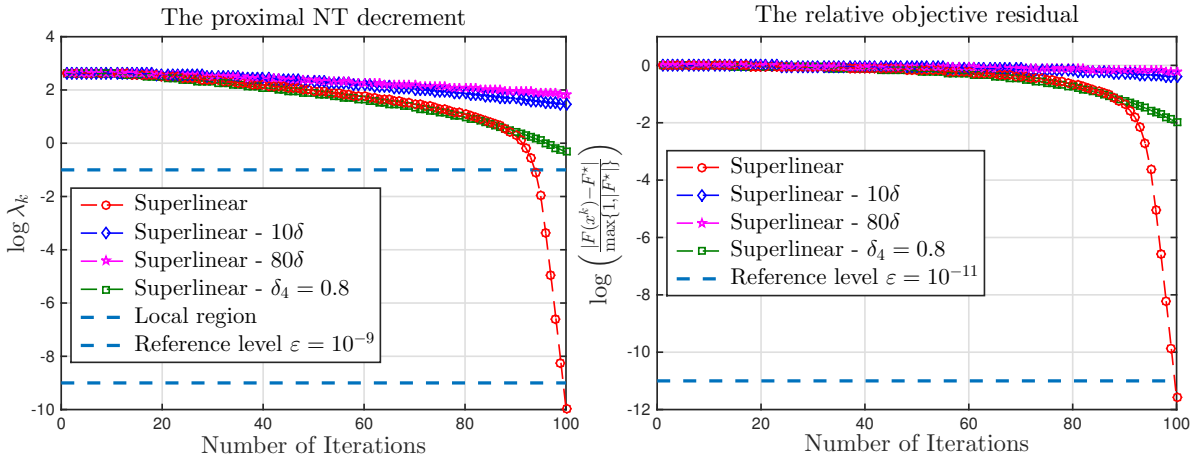

Figure 5.3: The local superlinear convergence of iPNA under the effect of inexact oracles.

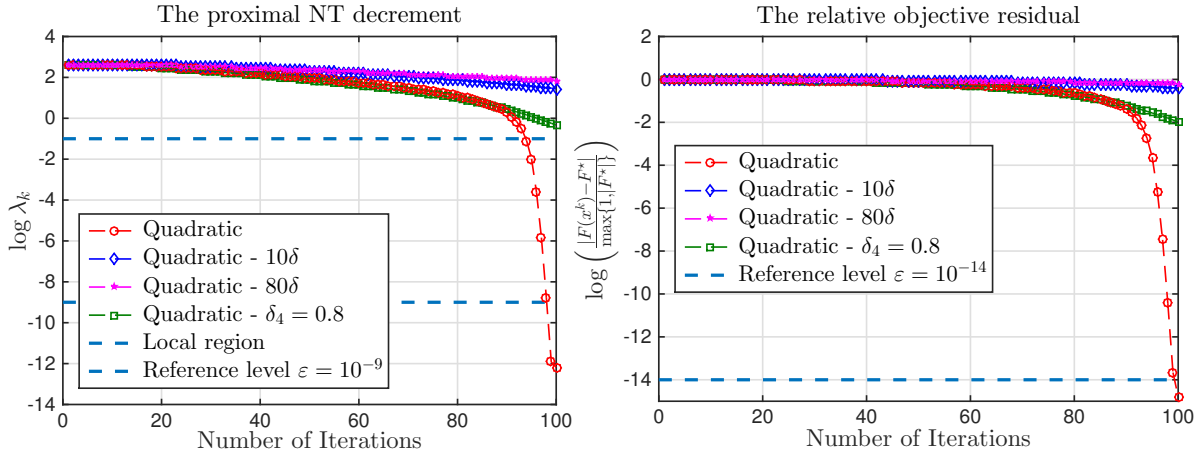

Figure 5.4: The local quadratic convergence of iPNA under the effect of inexact oracles.

If we multiply the accuracy δ by 10, and 80, respectively, we can see from this figure that the linear convergence is destroyed, and the method tends to diverge. If we choose the inexact level δ_4 of the subproblem in (5.20) to 0.8, we also get a significantly slow linear convergence rate.

The superlinear and quadratic convergence rates are reflected in Figure 5.3 and Figure 5.4, respectively. Both figures look very similar, but the quadratic convergence case achieves much higher accuracy up to 10^{-12} after around 100 iterations. If we increase the inexactness of the inexact oracle by multiplying δ by 10 and 80, respectively, iPNA shows its slow convergence or even diverges. If we increase the inexactness δ_4 of the subproblem in (5.20) to 0.8, we again obtain a much slower convergence rate.

5.6.1.2 Application to a network allocation problem

The composite model (5.39) can be applied to solve the following allocation problem. Assume that we have K cities described by polytopes as their possible area $\mathcal{P}_{[i]} := \{\mathbf{y} \in \mathbb{R}^p \mid \mathbf{C}^{[i]}\mathbf{y} \leq d^{[i]}\}$ for $i = 1, \cdots, K$. These cities are connected by a delivery network describing the routes between each pair of cities. Our goal is to locate a delivery center $y^{[i]} \in \mathcal{P}_{[i]}$ such that the total distances (or the total delivery costs) between these cities is minimized.

In order to guarantee $y^{[i]} \in \mathcal{P}_{[i]}$, we use a log-barrier function to handle this constraint. Therefore, one way to model this problem is to fit it into (5.39), where

$$\varphi(\mathbf{Ay}) := \mu \sum_{(i,j)\in\mathcal{E}} c_{ij}\|y^{[i]} - y^{[j]}\|_2 = \mu \sum_{(i,j)\in\mathcal{E}} c_{ij}\sqrt{(y_1^{[i]} - y_1^{[j]})^2 + (y_2^{[i]} - y_2^{[j]})^2},$$

where $c_{ij} \geq 0$ is the cost that is proportion to the distance between the i-th and j-th city, and $\mathcal{E}$ is the set of edges of the graph described this network, $\mu > 0$ is a penalty parameter in the barrier formulation (5.39), and $\mathbf{A}$ is a matrix describing the difference operator.

We first illustrate this model through a toy example, which creates a shape of UNC and STOR. Figure 5.5 demonstrates two unicursal routes of those word abbreviations.

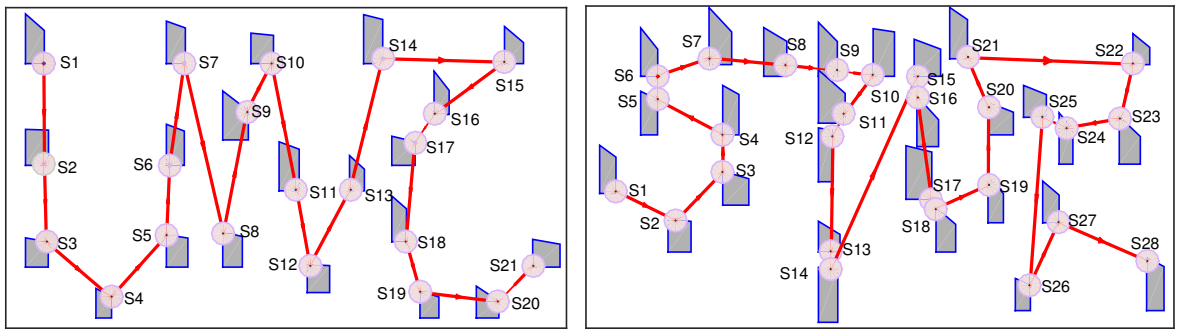

Figure 5.5: Optimal site allocation for routes UNC and STOR.

Figure 5.6 illustrates a real transportation network of US in 2015[2], and its actual optimal allocation solution. As we can see in the demo figure, this network model can be widely applied to airport and subway allocation, bipartite graph allocation, and many other fields with site allocation problems.

[2]http://esciencecommons.blogspot.com/2015/06/how-flu-viruses-use-transportation.html

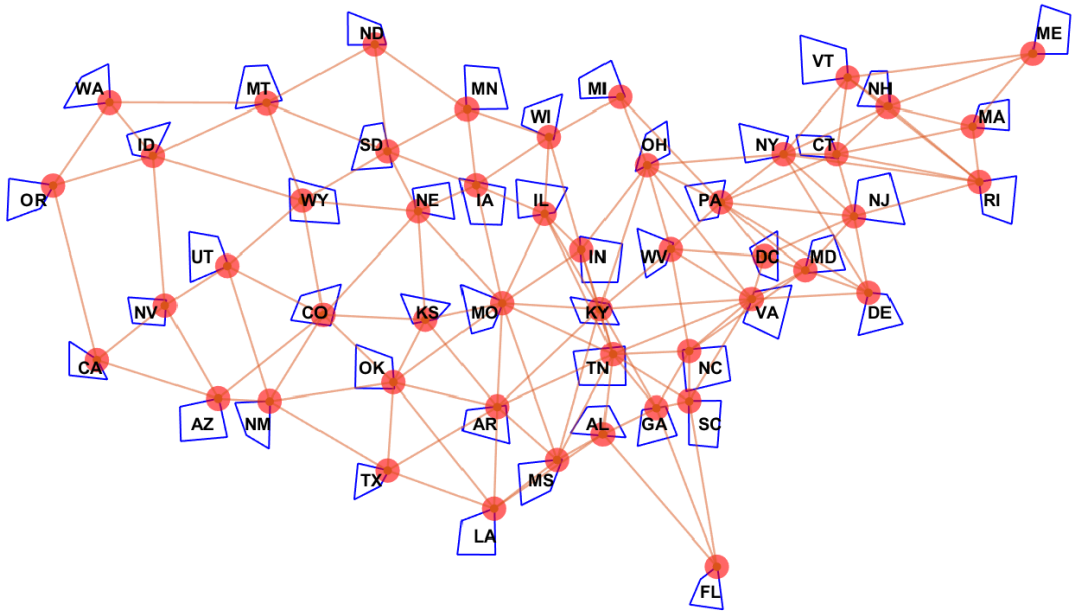

Figure 5.6: Optimal site allocation for US Network

Next, we test our methods on a collection of problems generated synthetically. We simulate the data by generating 17 problems with sparse network ($\rho = 0.04$) and 13 problems with dense network ($\rho = 0.15$). For problem of size $2p$, we generate an l-by-n rectangle area with $l = 10$ and $n = p/5$ in our case, with each area a 10×10 square. We randomly select p positions from the $2p$ square. For each chosen position i, with the central point being the origin, we again randomly generates one point as a vertex in each quadrant of the square, and then link them together as the feasible region of site i, where $i = 1, 2, \ldots, p$, and the matrix and vector $\mathbf{C}$ and $\mathbf{d}$ are generated from all feasible regions. We also generate a random adjacency matrix of size $p \times p$ with density $\rho = 0.04$ as the network, which corresponds to the linear operator $\mathbf{A}$ in the model setting. In practice, we choose $\mu = 10$, which is large enough to guarantee that the optimal points are near the boundary of feasible regions. (In fact, we exactly use $\mu = 10$ in Figure 5.5 and 5.6 below.) We choose all c_{ij}'s to be 1 in our case. Of course one can also use different c_{ij}'s to reflect the situation of different cities, or change the density or the shape of the network to reflect different real situations.

We solve this problem using inexact Newton method as before. Since the problem shares a sparse structure of matrix $\mathbf{A}$, we set the tolerances of the main loop to be $\texttt{tol}_{gap} := 10^{-10}$, and $\texttt{tol}_{sol} := 10^{-8}$. which measures the relative primal-dual gap defined by $r_{gap} := \frac{|F^* + G^*|}{1 + |F^*| + |G^*|}$,

and the maximum relative solution difference of primal and dual solutions defined by

$$r_{sol} := \max\left\{\frac{\|\mathbf{x}^{k+1}-\mathbf{x}^k\|_2}{\max\{1,\|\mathbf{x}^k\|_2\}}, \frac{\|\mathbf{y}^{k+1}-\mathbf{y}^k\|_2}{\max\{1,\|\mathbf{y}^k\|_2\}}\right\}$$

separately. We terminate our algorithm when the solution pair meets both criteria: (1) $r_{gap} < \texttt{tol}_{gap}$; and (2) $r_{sol} < \texttt{tol}_{sol}$.

5.6.1.3 Comparison to other methods

In this test, we show the advantages of our iPNA to existing state-of-the-arts such as SDPT3: a well-established interior-point solver to solve (5.39) [100], ADMM: the alternating direction method of multipliers [12], and CP: Chambolle-Pock's primal-dual first-order algorithm [16]. We note that since ψ in (5.39) does not have Lipschitz gradient, existing first-order methods such as proximal gradient-type, Frank-Wolfe, and coordinate descent methods are not applicable due to the lack of theoretical guarantees. For those three methods, we terminate all methods when both tolerance $\texttt{tol}_{gap}$ and $\texttt{tol}_{sol}$ are met. For SDPT3, since the formulation is different from others, we use their default measurement of the relative gap and solution feasibility. For the first order methods ADMM and CP, since it takes a long time for both to reach a high solution difference tolerance, we lower $\texttt{tol}_{sol}$ to 10^{-6}, instead of 10^{-8} in our algorithm. We also run CP for $10,000$ iterations to get a solution with a very high accuracy as a ground truth, and compare the relative primal solution error of all algorithms comparing with the ground truth, and the quantity is denoted by `qsol`, which measures the solution correctness and quality of each algorithm. Since there is no convergence rate guarantee at the first phase of ISNA algorithm, we use "n/t" to represent the number of iterations starts from $\mathbf{x}^k$ jumping into the local quadratic convergence range (measured by $\lambda_k < 0.1$, where we start to apply Theorem 5.4.3, 5.4.4), over the total number of ISNA iterations. If fact, n is the true number of iterations of the second-order method. The result is listed in Table 5.1.

The performance profile was studied in [32], which can be considered as a standard way to compare different optimization algorithms. A performance profile is built based on a set $\mathcal{S}$ of n_s algorithms (solvers) and a collection $\mathcal{P}$ of n_p problems. We build a profile based on computational time. We denote by $T_{ij} :=$ *computational time required to solve problem* i *by solver* j.

Table 5.1: The performance of two solvers for $l_{1,2}$-log barrier of 30 problems.

Problem	IPNA			SDPT3			ADMM			Chambolle-Pock		
Name	n/t	t[s]	qsol	iter	t[s]	qsol	iter	t[s]	qsol	iter	t[s]	qsol
sparse network												
p004120	16/72	2.8	9.0e-06	22	2.4	3.0e-05	207	0.4	9.3e-07	644	2.4	6.9e-05
p004160	16/79	3.1	4.4e-05	28	4.9	5.8e-06	253	0.7	1.2e-06	681	3.5	3.7e-04
p004200	16/91	6.0	4.3e-05	31	8.0	9.9e-06	329	1.3	7.6e-07	701	5.5	3.1e-04
p004240	17/98	6.9	1.1e-05	29	10.6	5.9e-06	336	5.2	1.4e-06	789	9.1	9.6e-05
p004280	16/105	8.7	8.2e-05	34	18.6	5.1e-06	397	13.9	3.3e-06	776	12.0	2.1e-04
p004320	18/114	9.0	1.4e-05	34	21.5	6.5e-06	375	17.6	1.7e-06	733	14.4	7.4e-05
p004360	16/118	10.0	4.1e-05	36	32.4	3.8e-06	308	20.6	1.6e-06	813	21.9	1.5e-04
p004400	18/131	20.2	2.1e-05	41	50.9	2.9e-06	677	64.9	4.2e-06	866	30.0	6.3e-05
p004440	18/132	18.7	7.7e-05	35	60.4	5.2e-06	524	59.7	2.6e-06	843	39.5	1.4e-04
p004480	20/146	26.1	1.5e-05	42	103.8	1.7e-06	584	84.8	5.9e-07	790	60.7	9.5e-05
p004520	17/146	29.3	3.1e-05	34	99.1	3.2e-06	577	102.7	2.0e-06	848	96.4	1.6e-04
p004560	17/150	29.2	2.6e-05	31	98.5	4.2e-06	447	89.9	6.7e-07	815	127.1	1.2e-04
p004600	20/158	42.5	3.6e-05	37	264.6	2.4e-06	564	141.3	2.0e-06	974	197.3	1.3e-04
p004640	18/172	54.0	2.8e-05	36	317.5	2.4e-06	649	184.3	1.1e-06	889	197.4	9.7e-05
p004680	19/172	61.2	3.4e-05	34	380.9	1.8e-06	688	230.6	1.0e-06	1042	267.9	9.2e-05
p004720	17/177	68.5	1.4e-05	38	539.5	2.8e-06	659	269.0	4.4e-07	844	290.1	7.0e-05
p004760	20/190	84.5	3.7e-05	40	742.7	1.5e-06	780	374.2	7.4e-06	1311	1544.9	8.6e-05
dense network												
p01580	17/75	1.7	2.7e-05	20	3.4	1.7e-05	356	0.5	1.0e-06	1107	3.4	3.1e-04
p015120	18/86	2.9	3.1e-06	22	8.0	1.2e-05	372	0.9	1.1e-07	491	2.6	2.8e-05
p015160	17/97	3.9	3.9e-06	22	15.7	5.8e-06	501	6.3	5.2e-07	640	6.4	4.0e-05
p015200	16/109	5.7	8.5e-06	28	37.1	1.0e-05	580	16.1	4.5e-07	901	12.6	8.7e-05
p015240	19/121	8.7	4.9e-06	29	59.3	6.3e-06	469	20.4	3.5e-07	613	16.3	3.9e-05
p015280	21/135	13.4	8.2e-06	32	193.6	6.5e-06	599	46.3	4.4e-07	861	25.1	7.8e-05
p015320	20/152	27.4	6.4e-06	33	333.0	4.8e-06	736	81.2	5.1e-07	1070	44.9	6.0e-05
p015360	19/161	33.8	2.6e-06	32	543.1	3.0e-06	694	107.8	3.8e-07	805	46.2	1.9e-05
p015400	20/164	34.2	1.1e-05	33	991.1	4.9e-06	1042	205.0	1.9e-06	946	78.5	7.9e-05
p015440	23/167	41.7	5.5e-06	33	1598.9	4.8e-06	755	225.1	8.0e-07	997	118.6	5.9e-05
p015480	20/188	82.0	2.0e-05	36	2380.3	4.0e-06	854	300.2	9.8e-07	872	213.6	8.6e-05
p015520	24/203	103.8	1.2e-05	40	3571.9	2.5e-06	820	353.1	3.2e-07	922	539.1	7.1e-05
p015560	19/206	121.1	5.8e-06	42	5365.4	1.8e-06	823	412.2	1.1e-06	1003	776.1	4.5e-05

We compare the performance of solver j on problem i with the best performance of any algorithm on this problem; that is we compute the performance ratio $r_{ij} := \frac{T_{ij}}{\min\{T_{ik}|k\in\mathcal{S}\}}$. Now, let $\tilde{\rho}_j(\tilde{\tau}) := \frac{1}{n_p}\text{size}\{i \in \mathcal{P} \mid r_{ij} \leq \tilde{\tau}\}$ for $\tilde{\tau} \in \mathbb{R}_+$. The function $\tilde{\rho}_j : \mathbb{R} \to [0,1]$ is the probability for solver j that a performance ratio is within a factor $\tilde{\tau}$ of the best possible ratio. We use the term "performance profile" for the distribution function $\tilde{\rho}_j$ of a performance met-

ric. In the following numerical examples, we plotted the performance profiles in $\log_2$-scale, i.e. $\rho_j(\tau) := \frac{1}{n_p}\text{size}\,\{i \in \mathcal{P} \mid \log_2(r_{ij}) \leq \tau := \log_2 \tilde{\tau}\}$.

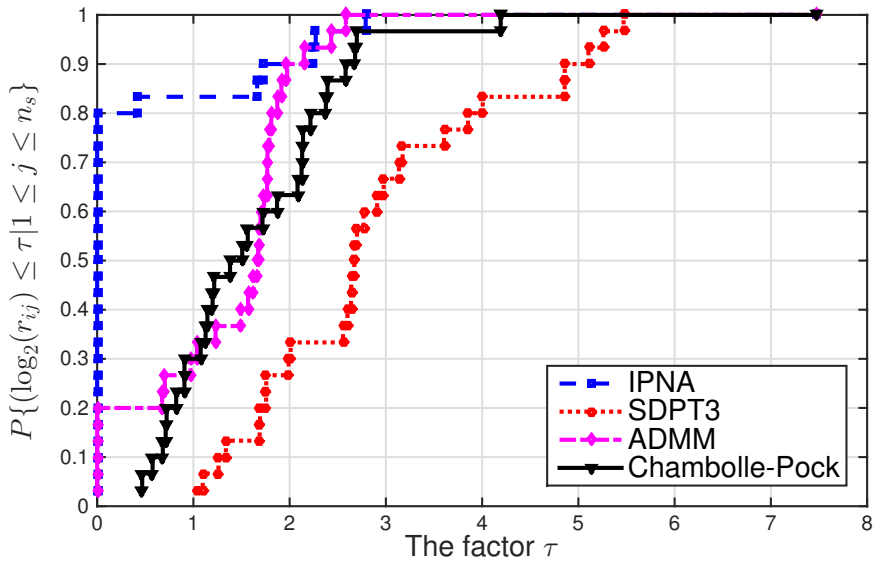

Figure 5.7: Performance Profile in time[s] of 4 methods and 30 problems

Figure 5.7 shows the performance profile of the four algorithms on a collection of the 30 problems indicated above. ISNA achieves 24/30 (80%) with the best performance, while ADMM obtains 6/30 (20%) with the best performance. In terms of computational time, both proximal inexact Newton method and first-order methods outperform SDPT3 in this experiment. We can also see from Table 5.1 that ADMM gives the best solution quality in most cases, while CP gives the worst solution quality.

5.6.2 iPNA for Graphical LASSO with inexact oracles

Proximal-Newton-type methods have been proven to be efficient for graphical LASSO [29, 39, 51, 52, 83]. In this example, we show that our theory can be useful for this problem. We illustrate this ability by considering a recent setting in [108]. Assume the data has a sparse graph structure G, then the original graphical lasso model can be written as

$$F^\star := \min_{\mathbf{X} \succ 0} \Big\{ F(\mathbf{X}) := \langle \mathbf{C}_\lambda, \mathbf{X} \rangle - \log\det(\mathbf{X}) \quad | \quad X_{ij} = 0,\ \forall\ (i,j) \notin G \Big\},$$

where $\mathbf{C}_\lambda$ is a soft-threshold operator which serves as the penalty item, that can recover the sparse graph G. The form we are interested in is the dual problem (15) of [108]. We focus on two-folds of the inexactness: (1) the inexactness of the solution of subproblem (5.20), where $R(\mathbf{x}) \equiv 0$ in this case; (2) the Hessian and the Newton decrement measurement reflected by Cholesky decomposition. Besides, instead of using line-search, we use the step-size given by (5.21), the self-concordance theory.

For (1), we solve the Newton decrement inexactly by controlling the tolerance of the preconditioned conjugate gradient (PCG) method. Here we set the tolerance of PCG to be 10^{-3}. For (2), since we are dealing with the data that shares the sparsity structure, we use the incomplete Cholesky decomposition instead of exact Cholesky decomposition. In detail, when we are solving the lower triangular matrix $\tilde{\mathbf{L}}$ such that $\mathbf{A} \approx \tilde{\mathbf{L}}\tilde{\mathbf{L}}^\top$, we fill all other off-diagonal elements to 0, if the original entry of $\mathbf{A}$ is 0. By doing this we take further advantages of sparsity structure than the original method, and bring the inexactness to the Hessian-related quantity indirectly.

For data, we use both the real-world biology dataset from [64] and the synthetic data with sample covariance matrices and the threshold parameter generated from real sparse matrix/graph collection[3] as the way did in [108]. Since the Newton-CG(NCG) method with line-search proposed in the latest paper [108] already compared and beaten QUIC in their experiments, we make use of the chordal property of the graph structure and only compare our algorithm with the proposed algorithm in their paper. Following their paper, we measure the stopping criterion of both algorithms by λ_k. We set the tolerance to be 10^{-6}. For the subproblem, we use the original stop criterion for NCG, but our criteria listed above for our inexact self-concordant Newton algorithm (ISNA). The results are listed in the following Table 5.2.

In the table, p is the dimension of the original graph/data, "`iter`" means number of iterations in the main loop, "λ_e" means the weighted norm λ_k which is used by NCG when the algorithm stops. "`soldiff`" measures the relative solution difference of two methods for primal solution, and "$\mathtt{t}_{\text{ratio}}$" represents the time ratio of NCG over ISNA.

From the table we can see that for both synthetic data derived from real sparse graph structure and real data, we performed better than the state-of-the-art algorithm NCG with

[3] https://sparse.tamu.edu/

Table 5.2: The performance of NCG and ISNA for solving the graphical lasso problem.

Problem		IPNA			NCG with ls			Comparison	
Name	p	`iter`	`time[s]`	λ_e	`iter`	`time[s]`	λ_e	`soldiff`	$\mathtt{t}_{\text{ratio}}$
Synthetic Data									
3eltdual	9000	4	11.45	3.0e-07	3	13.15	2.7e-07	3.0e-12	1.15
bcsstm38	8032	3	2.84	6.1e-07	3	4.36	5.2e-10	4.5e-12	1.54
cage8	1015	7	62.99	3.2e-07	4	116.64	3.1e-10	1.1e-09	1.85
cryg10000	10000	6	543.31	4.1e-08	4	634.06	2.8e-10	5.6e-12	1.17
FlyingRobot1	798	6	4.23	5.3e-07	4	9.98	5.2e-11	5.7e-10	2.36
G32	2000	4	2.79	7.3e-07	4	5.17	7.5e-12	4.9e-11	1.85
G50	3000	5	5.49	3.9e-09	4	7.75	6.1e-11	2.6e-13	1.41
G57	5000	5	9.10	2.1e-07	4	12.75	2.6e-10	5.5e-12	1.40
lshp2614	2614	6	108.29	1.8e-07	4	162.54	7.4e-11	1.3e-10	1.50
lshp3025	3025	6	137.24	3.8e-07	4	215.66	6.7e-11	3.2e-10	1.57
NotreDamey	2114	3	1.57	7.6e-08	3	2.19	4.1e-11	1.8e-12	1.40
orsirr2	886	6	7.17	1.7e-07	4	13.35	2.2e-10	4.8e-10	1.86
sherman3	5005	5	56.11	5.1e-07	4	99.77	1.3e-11	3.0e-10	1.78
ukerbe1	5981	3	5.23	5.8e-07	3	8.63	1.2e-10	1.2e-11	1.65
USpowerGrid	4941	3	4.66	4.9e-07	3	7.09	6.7e-09	5.2e-12	1.52
Real Data									
Arabidopsis	834	4	1.27	1.2e-07	4	1.41	5.0e-09	2.8e-12	1.11
ER	692	4	0.89	1.5e-08	4	1.25	5.8e-11	8.8e-14	1.40
hereditarybc	1869	4	21.06	2.9e-07	4	35.39	1.7e-07	7.3e-12	1.68
Leukemia	1255	3	0.60	7.6e-08	3	0.76	2.7e-09	8.6e-13	1.25
Lymph	587	4	0.24	8.5e-10	3	0.25	9.1e-07	2.4e-14	1.03

linesearch. Although for some graphs we cannot accelerate too much, we point out that NCG already taken the advantages of the chordal structure and used the linesearch, while our methods specify a step-size and the acceleration is highly related to the sparsity and the shape of the graph. Besides, we need slightly more iterations and end up with a greater λ_e than NCG, because we did not solve the subproblem to a very high accuracy, which leads to a smaller descent. However, we still met the terminating criterion and obtained the same solution (`soldiff`) of the target problem.

5.7 Conclusion

In this chapter we introduced the concept of inexact oracle, which consists of both global and local inexact oracles. Following the definition, we developed some key properties using such oracles and presented several examples. We then developed the inexact proximal Newton-type methods and showed that the obtained algorithms achieved both global convergence and local convergence from linear to quadratic rate. We also showed that our methods cover some existing inexact methods in the literature as special cases. For application, we developed the corresponding theory for primal-dual methods, and provided some representative examples to illustrate the entire inexact oracle theory.

CHAPTER 6

Conclusions and future works

6.1 Conclusions

In this thesis, we have introduced two new concepts for a class of convex functions. The first concept is a so-called "generalized self-concordance" notion, which can be considered as a generalization of the standard self-concordant concept introduced by Nesterov and Nemirovski in the early 1990s. The second one is a newly inexact oracle notion in the context of composite convex optimization. Both concepts cover a wide range class of convex functions than existing structural assumptions used in the literature.

The generalized self-concordant function class covers many important and well-known models in convex optimization, machine learning, and statistics. Relying on our new definition, we have developed several fundamental properties for this class of functions and provided a unified framework to develop new numerical methods. As byproducts, we have applied our new theory to develop a class of Newton-type methods that include different variants such as damped-step Newton, full-step Newton, quasi-Newton, and proximal Newton-type methods. Our new theory allows us to analyze both global and local convergence of the new algorithms in a rigorous manner without imposing any unverifiable assumptions as in existing methods. We have also illustrated the benefits of the proposed methods through some numerical examples using both synthetic and real datasets, and compared them with some state-of-the-art algorithms.

In the second part, we have introduced novel global and local second-order inexact oracle concepts for a wide class of convex functions. Our global inexact oracle covers both the well-known Lipschitz gradient and self-concordant convex function classes as special cases. Utilizing our new definitions, we have developed several key properties and provided representative examples for our new function class. Then, we have developed an inexact proximal Newton methods under inexact oracles. We have proved both global and local convergence of the proposed meth-

ods. We have achieved different local convergence rates ranging from R-linear, R-superlinear to R-quadratic by controlling the inexactness levels in our oracles under the self-concordant assumption. We have also customized our method to handle a primal-dual formulation. Our theoretical results have been verified through several numerical results in comparison with other state-of-the-art methods.

6.2 Future works

The theory and numerical methods developed in this thesis are expected to have broad applications in different fields. For the generalized self-concordance notion, firstly, we plan to apply our results to solve some representative applications in high-dimensional spaces, where existing methods do not have a theoretical guarantee. Our next idea is to combine this new theory and some recent advanced techniques such as stochastic, randomized, and conjugate gradient methods to scale up the problem sizes. Secondly, we wish to further accelerate existing methods to solve more problems which possess proper smoothness structures. To do this, we will combine other smoothness structures, such as the Lipschitz gradient, and/or strong convexity structures together with *generalized self-concordant* settings, explore key properties of the joint structures, and develop the corresponding numerical methods.

For our new inexact oracle concepts, firstly, we plan to handle inexactness situations for a wider class of functions than the current settings, especially in the local convergent stage. Since iterative schemes based on local oracle concepts are still limited to the self-concordant function class, one of our approach is to expand the inexact oracle settings to the *generalized self-concordant* function class developed in the first part of this thesis. By doing this, we can accelerate a broader range of composite optimization problems than existing cases. Secondly, we expect to customize our inexact oracle settings to one or more general algorithmic schemes, including but not limited to [block] coordinate descent, sketching, and random subsampling methods, to develop new inexact methods with better performance or stability. To do this, we need to expand our theory from existing deterministic schemes to stochastic ones, and study further the relations between inexact oracles and distributed and parallel computations.

We emphasize that our framework can be extended to handle constrained convex problems by combining with duality theory. We have illustrated this idea in the second part of this thesis, but we still expect to develop other schemes for solving constrained convex problems by utilizing our new concepts introduced in this thesis.

APPENDIX A

PROOFS OF TECHNICAL RESULTS

This appendix provides the full proofs of technical results that are not shown in the main text of each chapter.

A.1 Technical proofs of results in Chapter 3

A.1.1 The proof of Proposition 3.5.1: Fenchel's conjugate

Let us consider the set $\mathcal{X} := \{\mathbf{x} \in \mathbb{R}^p \mid f(\mathbf{u}) - \langle \mathbf{x}, \mathbf{u} \rangle \text{ is bounded from below on } \mathrm{dom}(f)\}$. We first show that $\mathrm{dom}(f^*) = \mathcal{X}$.

By the definition of $\mathrm{dom}(f^*)$, we have $\mathrm{dom}(f^*) = \{\mathbf{x} \in \mathbb{R}^p \mid f^*(\mathbf{x}) < +\infty\}$. Take any $\mathbf{x} \in \mathrm{dom}(f^*)$, one has $f^*(\mathbf{x}) = \max_{\mathbf{u} \in \mathrm{dom}(f)} \{\langle \mathbf{x}, \mathbf{u} \rangle - f(\mathbf{u})\} < +\infty$. Hence $f(\mathbf{u}) - \langle \mathbf{x}, \mathbf{u} \rangle \geq -f^*(\mathbf{x}) > -\infty$ for all $\mathbf{u} \in \mathrm{dom}(f)$, which implies $\mathbf{x} \in \mathcal{X}$.

Conversely, assume that $\mathbf{x} \in \mathcal{X}$. By the definition of $\mathcal{X}$, $f(\mathbf{u}) - \langle \mathbf{x}, \mathbf{u} \rangle$ is bounded from below for all $\mathbf{u} \in \mathrm{dom}(f)$. That is, there exists $M \in [0, +\infty)$, such that $f(\mathbf{u}) - \langle \mathbf{x}, \mathbf{u} \rangle \geq -M$ for all $\mathbf{u} \in \mathrm{dom}(f)$. By the definition of the conjugate, $f^*(\mathbf{x}) = \max_{\mathbf{u} \in \mathrm{dom}(f)} \{\langle \mathbf{x}, \mathbf{u} \rangle - f(\mathbf{u})\} \leq M < +\infty$. Hence, $\mathbf{x} \in \mathrm{dom}(f^*)$.

For any $\mathbf{x} \in \mathrm{dom}(f^*)$, the optimality condition of $\max_{\mathbf{u}} \{\langle \mathbf{x}, \mathbf{u} \rangle - f(\mathbf{u})\}$ is $\mathbf{x} = \nabla f(\mathbf{u})$. Let us denote by $x(\mathbf{u}) = \nabla f(\mathbf{u})$. Then, we have $f^*(x(\mathbf{u})) = \langle x(\mathbf{u}), \mathbf{u} \rangle - f(\mathbf{u})$. Taking derivative of f^* with respect to $\mathbf{x}$ on both sides, and using $x(\mathbf{u}) = \nabla f(\mathbf{u})$, we have

$$\nabla_{\mathbf{x}} f^*(x(\mathbf{u})) = \mathbf{u} + u'_{\mathbf{x}} x(\mathbf{u}) - u'_{\mathbf{x}} \nabla f(\mathbf{u}) = \mathbf{u}.$$

We further take the second-order derivative of the above equation with respect to $\mathbf{u}$ to get

$$\nabla^2 f^*(x(\mathbf{u})) x'_{\mathbf{u}}(\mathbf{u}) = \mathbb{I}.$$

Using the two relations above and the fact that $x'_{\mathbf{u}}(\mathbf{u}) = \nabla^2 f(\mathbf{u})$, we can derive

$$\left\langle \nabla f^*(x(\mathbf{u})), x'_{\mathbf{u}}(\mathbf{u})\mathbf{v} \right\rangle = \left\langle \mathbf{u}, x'_{\mathbf{u}}(\mathbf{u})\mathbf{v} \right\rangle = \left\langle \nabla^2 f(\mathbf{u})\mathbf{v}, \mathbf{u} \right\rangle \tag{A.1}$$

$$\left\langle \nabla^2 f^*(x(\mathbf{u}))x'_{\mathbf{u}}(\mathbf{u})\mathbf{v}, x'_{\mathbf{u}}(\mathbf{u})\mathbf{w} \right\rangle = \left\langle \mathbf{v}, x'_{\mathbf{u}}(\mathbf{u})\mathbf{w} \right\rangle = \left\langle \nabla^2 f(\mathbf{u})\mathbf{v}, \mathbf{w} \right\rangle, \tag{A.2}$$

where $\mathbf{u} \in \mathrm{dom}(f)$, and $\mathbf{v}, \mathbf{w} \in \mathbb{R}^p$. Using (A.1) and (A.2), we can compute the third-order derivative of f^* with respect to $x(\mathbf{u})$ as

$$\begin{aligned}
& \langle \nabla^3 f^*(x(\mathbf{u}))[x'_{\mathbf{u}}(\mathbf{u})\mathbf{w}]x'_{\mathbf{u}}(\mathbf{u})\mathbf{v}, x'_{\mathbf{u}}(\mathbf{u})\mathbf{v} \rangle \\
= & \left\langle \left(\left\langle \nabla^2 f^*(x(\mathbf{u}))x'_{\mathbf{u}}(\mathbf{u})\mathbf{v}, x'_{\mathbf{u}}(\mathbf{u})\mathbf{v} \right\rangle \right)'_{\mathbf{u}}, \mathbf{w} \right\rangle - 2 \left\langle \nabla^2 f^*(x(\mathbf{u}))x'_{\mathbf{u}}(\mathbf{u})\mathbf{v}, (x'_{\mathbf{u}}(\mathbf{u})\mathbf{v})'_{\mathbf{u}}\mathbf{w} \right\rangle \\
\overset{(A.1)}{=} & \langle (\langle x'_{\mathbf{u}}(\mathbf{u})\mathbf{v}, \mathbf{v} \rangle)'_{\mathbf{u}}, \mathbf{w} \rangle - 2 \left\langle \nabla^2 f^*(x(\mathbf{u}))x'_{\mathbf{u}}(\mathbf{u})\mathbf{v}, (x'_{\mathbf{u}}(\mathbf{u})\mathbf{v})'_{\mathbf{u}}\mathbf{w} \right\rangle \\
\overset{(A.2)}{=} & \left\langle \nabla^3 f(\mathbf{u})[\mathbf{w}]\mathbf{v}, \mathbf{v} \right\rangle - 2 \left\langle (\mathbf{x}'_{\mathbf{u}}(\mathbf{u})\mathbf{v})'_{\mathbf{u}}\mathbf{w}, \mathbf{v} \right\rangle \\
= & - \left\langle \nabla^3 f(\mathbf{u})[\mathbf{w}]\mathbf{v}, \mathbf{v} \right\rangle.
\end{aligned} \tag{A.3}$$

Denote $\xi := x'_{\mathbf{u}}(\mathbf{u})\mathbf{w}$ and $\eta := x'_{\mathbf{u}}(\mathbf{u})\mathbf{v}$. Note that since $x'_{\mathbf{u}}(\mathbf{u}) = \nabla^2 f(\mathbf{u})$, we have $\xi = \nabla^2 f(\mathbf{u})\mathbf{w}$, $\eta = \nabla^2 f(\mathbf{u})\mathbf{v}$, and $\mathbf{w} = \nabla^2 f(\mathbf{u})^{-1}\xi$. Using these relations and $\nabla^2 f^*(x(\mathbf{u}))x'_{\mathbf{u}}(\mathbf{u}) = \mathbb{I}$, we can derive

$$\begin{aligned}
|\langle \nabla^3 f^*(x(\mathbf{u}))[\xi]\eta, \eta \rangle| & \overset{(A.3)}{=} |\left\langle \nabla^3 f(\mathbf{u})[\mathbf{w}]\mathbf{v}, \mathbf{v} \right\rangle| \overset{(3.2)}{\leq} M_f \|\mathbf{v}\|_{\mathbf{u}}^2 \|\mathbf{w}\|_{\mathbf{u}}^{\nu-2} \|\mathbf{w}\|_2^{3-\nu} \\
& = M_f \left\langle \nabla^2 f(\mathbf{u})\mathbf{v}, \mathbf{v} \right\rangle \left\langle \nabla^2 f(\mathbf{u})\mathbf{w}, \mathbf{w} \right\rangle^{\frac{\nu-2}{2}} \|\mathbf{w}\|_2^{3-\nu} \\
& = M_f \left\langle \eta, \nabla^2 f^*(x(\mathbf{u}))x'(\mathbf{u})\mathbf{v} \right\rangle \left\langle \xi, \nabla^2 f^*(x(\mathbf{u}))x'(\mathbf{u})\mathbf{w} \right\rangle^{\frac{\nu-2}{2}} \|\nabla^2 f(\mathbf{u})^{-1}\xi\|^{3-\nu} \\
& = M_f \left\langle \nabla^2 f^*(x(\mathbf{u}))\eta, \eta \right\rangle \left\langle \nabla^2 f^*(x(\mathbf{u}))\xi, \xi \right\rangle^{\frac{\nu-2}{2}} \left\langle \nabla^2 f^*(x(\mathbf{u}))\xi, \nabla^2 f^*(x(\mathbf{u}))\xi \right\rangle^{3-\nu}.
\end{aligned}$$

For any $\mathbf{H} \in \mathcal{S}^p_{++}$, we have $\langle \mathbf{H}\xi, \xi \rangle \leq \|\mathbf{H}\xi\|_2 \|\xi\|_2$. For any $\nu \geq 3$, this inequality leads to

$$\langle \mathbf{H}\xi, \xi \rangle^{\frac{\nu-2}{2}} \|\mathbf{H}\xi\|^{3-\nu} \leq \langle \mathbf{H}\xi, \xi \rangle^{\frac{4-\nu}{2}} \|\xi\|_2^{\nu-3}.$$

Using this inequality with $\mathbf{H} = \nabla^2 f^*(x(\mathbf{u}))$ into the last expression, we obtain

$$\begin{aligned}
|\left\langle \nabla^3 f^*(x(\mathbf{u}))[\xi]\eta, \eta \right\rangle| & \leq M_f \left\langle \nabla^2 f^*(x(\mathbf{u}))\eta, \eta \right\rangle \left\langle \nabla^2 f^*(x(\mathbf{u}))\xi, \xi \right\rangle^{\frac{4-\nu}{2}} \|\xi\|_2^{\nu-3} \\
& = M_f \|\eta\|^2_{x(\mathbf{u})} \|\xi\|^{4-\nu}_{x(\mathbf{u})} \|\xi\|_2^{\nu-3}.
\end{aligned}$$

The above inequality shows that $f^* \in \widetilde{\mathcal{F}}_{M_{f^*},\nu_*}$ with $M_{f^*} = M_f$ and $\nu_* = 6 - \nu$. However, to guarantee $\nu - 3 \geq 0$ and $6 - \nu > 0$, we require $3 \leq \nu < 6$.

Finally, we prove the case of univariate functions, i.e., $p = 1$. Indeed, we have

$$x(\mathbf{u}) = f'(\mathbf{u}), \quad (f^*)'(x(\mathbf{u})) = \mathbf{u}, \quad \text{and} \quad (f^*)''(x(\mathbf{u}))x'(\mathbf{u}) = 1. \tag{A.4}$$

Here, f' is the derivative of f with respect to $\mathbf{u}$. Taking the derivative of the last equation on both sides with respect to $\mathbf{u}$, we obtain

$$(f^*)'''(x(\mathbf{u}))(x'(\mathbf{u}))^2 + (f^*)''(x(\mathbf{u}))x''(\mathbf{u}) = 0.$$

Solving this equation for $(f^*)'''(x(\mathbf{u}))$ and then using (A.4) and $x''(\mathbf{u}) = f'''(\mathbf{u})$, we get

$$\begin{aligned} |(f^*)'''(x(\mathbf{u}))| = & \ |\tfrac{(f^*)''(x(\mathbf{u}))x''(\mathbf{u})}{(x'(\mathbf{u}))^2}| = |((f^*)''(x(\mathbf{u})))^3 f'''(\mathbf{u})| \\ \leq & \ M_f|((f^*)''(x(\mathbf{u})))^3 (f''(\mathbf{u}))^{\frac{\nu}{2}}| = M_f((f^*)''(x(\mathbf{u})))^{\frac{6-\nu}{2}}. \end{aligned}$$

This inequality shows that f^* is *generalized self-concordant* with $\nu_* = 6 - \nu$ for any $\nu \in (0, 6)$.

A.1.2 The proof of Corollary 3.7.3: Bound on the mean of Hessian operator

Let $\mathbf{y}_\tau := \mathbf{x} + \tau(\mathbf{y} - \mathbf{x})$. Then $d_\nu(\mathbf{x}, \mathbf{y}_\tau) = \tau d_\nu(\mathbf{x}, \mathbf{y})$. By (3.11), we have $\nabla^2 f(\mathbf{x}+\tau(\mathbf{y}-\mathbf{x})) \preceq \left(1 - \frac{\nu-2}{2}\tau d_\nu(\mathbf{x}, \mathbf{y})\right)^{\frac{-2}{\nu-2}} \nabla^2 f(\mathbf{x})$ and $\nabla^2 f(\mathbf{x}+\tau(\mathbf{y}-\mathbf{x})) \succeq \left(1 - \frac{\nu-2}{2}\tau d_\nu(\mathbf{x}, \mathbf{y})\right)^{\frac{2}{\nu-2}} \nabla^2 f(\mathbf{x})$. Hence, we have

$$\underline{I}_\nu(\mathbf{x}, \mathbf{y})\nabla^2 f(\mathbf{x}) \preceq \int_0^1 \nabla^2 f(\mathbf{x} + \tau(\mathbf{y} - \mathbf{x}))d\tau \preceq \overline{I}_\nu(\mathbf{x}, \mathbf{y})\nabla^2 f(\mathbf{x}),$$

where $\underline{I}_\nu(\mathbf{x}, \mathbf{y}) := \int_0^1 \left(1 - \frac{\nu-2}{2}\tau d_\nu(\mathbf{x}, \mathbf{y})\right)^{\frac{2}{\nu-2}} d\tau$ and $\overline{I}_\nu(\mathbf{x}, \mathbf{y}) := \int_0^1 \left(1 - \frac{\nu-2}{2}\tau d_\nu(\mathbf{x}, \mathbf{y})\right)^{\frac{-2}{\nu-2}} d\tau$ are the two integrals in the above inequality. Computing these integrals explicitly, we can show that

- If $\nu = 4$, then $\underline{I}_\nu(\mathbf{x}, \mathbf{y}) = 1 - \frac{d_4(\mathbf{x},\mathbf{y})}{2}$ and $\overline{I}_\nu(\mathbf{x}, \mathbf{y}) = \frac{-\ln(1-d_4(\mathbf{x},\mathbf{y}))}{d_4(\mathbf{x},\mathbf{y})}$.

- If $\nu \neq 4$, then we can easily compute $\underline{I}_\nu(\mathbf{x},\mathbf{y}) = \frac{2}{\nu d_\nu(\mathbf{x},\mathbf{y})}\Big(1 - \big(1 - \frac{\nu-2}{2} d_\nu(\mathbf{x},\mathbf{y})\big)^{\frac{\nu}{\nu-2}}\Big)$, and $\overline{I}_\nu(\mathbf{x},\mathbf{y}) = \frac{2}{(\nu-4) d_\nu(\mathbf{x},\mathbf{y})}\Big(1 - \big(1 - \frac{\nu-2}{2} d_\nu(\mathbf{x},\mathbf{y})\big)^{\frac{\nu-4}{\nu-2}}\Big)$.

Hence, we obtain (3.13).

Finally, we prove for the case $\nu = 2$. Indeed, by (3.11), we have $e^{-d_2(\mathbf{x},\mathbf{y}_\tau)}\nabla^2 f(\mathbf{x}) \preceq \nabla^2 f(\mathbf{y}_\tau) \preceq e^{d_2(\mathbf{x},\mathbf{y}_\tau)}\nabla^2 f(\mathbf{x})$. Since $d_2(\mathbf{x},\mathbf{y}_\tau) = \tau d_2(\mathbf{x},\mathbf{y})$, the last estimate leads to

$$\left(\int_0^1 e^{-d_2(\mathbf{x},\mathbf{y})\tau} d\tau\right)\nabla^2 f(\mathbf{x}) \preceq \int_0^1 \nabla^2 f(\mathbf{y}_\tau) d\tau \preceq \left(\int_0^1 e^{d_2(\mathbf{x},\mathbf{y})\tau} d\tau\right)\nabla^2 f(\mathbf{x}),$$

which is exactly (3.13).

A.2 Technical proofs of results in Chapter 4

A.2.1 Techical lemmas

The following lemmas will be used in our analysis. Lemma A.2.1 is elementary, but we provide its proof for completeness.

Lemma A.2.1. (a) For a fixed $r \geq 1$ and $\bar{t} \in (0,1)$, consider a function $\psi_r(t) := \frac{1-(1-t)^r - rt(1-t)^r}{rt^2(1-t)^r}$ on $t \in (0,1)$. Then, ψ is positive and increasing on $(0,\bar{t}]$ and

$$\lim_{t\to 0^+}\psi_r(t) = \tfrac{r+1}{2}, \quad \lim_{t\to 1^-}\psi_r(t) = +\infty, \quad \text{and} \quad \sup_{0\leq t\leq \bar{t}}|\psi_r(t)| \leq \bar{C}_r(\bar{t}) < +\infty,$$

where $\bar{C}_r(\bar{t}) := \frac{1-(1-\bar{t})^r - r\bar{t}(1-\bar{t})^r}{r\bar{t}^2(1-\bar{t})^r} \in (0,+\infty)$.

(b) For $t > 0$, we also have $\frac{e^t-1-t}{t} \leq \left(\frac{3}{2} + \frac{t}{3}\right) te^t$.

Proof. The statement (b) is rather elementary, we only prove (a). Since $r \geq 1$, $\lim_{t\to 0^+}(1 - (1-t)^r - rt(1-t)^r) = \lim_{t\to 0^+} rt^2(1-t)^r = 0$ and $rt^2(1-t)^r > 0$ for $t \in (0,1)$, applying L'Hôspital's rule, we have

$$\lim_{t\to 0^+}\psi_r(t) = \frac{\lim_{t\to 0^+} r(r+1)t(1-t)^{r-1}}{\lim_{t\to 0^+} rt(2-(2+r)t)(1-t)^{r-1}} = \frac{\lim_{t\to 0^+}(r+1)}{\lim_{t\to 0^+}(2-(2+r)t)} = \frac{r+1}{2}.$$

The limit $\lim_{t\to 1^-}\psi_r(t) = +\infty$ is obvious.

Next, it is easily to compute $\psi_r'(t) = \frac{(1-t)^{r+1}(rt+2)+(r+2)t-2}{rt^3(1-t)^{r+1}}$. Let $m_r(t) := (1-t)^{r+1}(rt+2)+(r+2)t-2$ be the numerator of $\psi_r'(t)$.

We have $m_r'(t) = r+2-(1-t)^r(r^2t+2rt+r+2)$, and $m_r''(t) = r(r+1)(r+2)t(1-t)^{r-1}$. Clearly, since $r \geq 1$, $m_r''(t) \geq 0$ for $t \in [0,1]$. This implies that m_r' is nondecreasing on $[0,1]$. Hence, $m_r'(t) \geq m_r'(0) = 0$ for all $t \in [0,1]$. Consequently, m_r is nondecreasing on $[0,1]$. Therefore, $m_r(t) \geq m_r(0) = 0$ for all $t \in [0,1]$. Using the formula of ψ_r', we can see that $\psi_r'(t) \geq 0$ for all $t \in (0,1)$. This implies that ψ_r is nondecreasing on $(0,1)$. Moreover, $\lim_{t\to 0^+} \psi_r(t) = \frac{r+1}{2} > 0$. Hence, $\psi_r(t) > 0$ for all $t \in (0,1)$. This implies that ψ_r is bounded on $(0,\bar{t}] \subset (0,1)$ by $\psi_r(\bar{t})$. □

Similar to Corollary 3.7.3, we prove the following lemma on the bound of the Hessian difference.

Lemma A.2.2. Given $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$, the matrix $\mathbf{H}(\mathbf{x},\mathbf{y})$ defined by

$$\mathbf{H}(\mathbf{x},\mathbf{y}) := \nabla^2 f(\mathbf{x})^{-1/2} \left[\int_0^1 \left(\nabla^2 f(\mathbf{x}+\tau(\mathbf{y}-\mathbf{x})) - \nabla^2 f(\mathbf{x}) \right) d\tau \right] \nabla^2 f(\mathbf{x})^{-1/2}, \tag{A.5}$$

satisfies

$$\|\mathbf{H}(\mathbf{x},\mathbf{y})\| \leq R_\nu \left(d_\nu(\mathbf{x},\mathbf{y}) \right) d_\nu(\mathbf{x},\mathbf{y}), \tag{A.6}$$

where $R_\nu(t)$ is defined as follows for $t \in [0,1)$:

$$R_\nu(t) := \begin{cases} \left(\frac{3}{2} + \frac{t}{3}\right) e^t & \text{if } \nu = 2 \\ \frac{2}{(4-\nu)t^2} \left[\left(1 - \frac{\nu-2}{2} t\right)^{\frac{4-\nu}{2-\nu}} - 1 \right] - \frac{1}{t} & \text{if } 2 < \nu \leq 3. \end{cases} \tag{A.7}$$

Moreover, for a fixed $\bar{t} \in (0,1)$, we have $\sup_{0\leq t\leq \bar{t}} |R_\nu(t)| \leq \bar{M}_\nu(\bar{t})$, where

$$\bar{M}_\nu(\bar{t}) := \max \left\{ \frac{2}{(4-\nu)\bar{t}^2} \left[\left(1 - \frac{\nu-2}{2}\bar{t}\right)^{\frac{4-\nu}{2-\nu}} - 1 \right] - \frac{1}{\bar{t}}, \left(\frac{3}{2} + \frac{\bar{t}}{2}\right) e^{\bar{t}} \right\} \in (0, +\infty).$$

Proof. By Corollary 3.7.3, if we define $\mathbf{G}(\mathbf{x},\mathbf{y}) := \int_0^1 \left[\nabla^2 f(\mathbf{x}+\tau(\mathbf{y}-\mathbf{x})) - \nabla^2 f(\mathbf{x}) \right] d\tau$, then

$$[\underline{\kappa}_\nu(d_\nu(\mathbf{x},\mathbf{y})) - 1] \nabla^2 f(\mathbf{x}) \preceq \mathbf{G}(\mathbf{x},\mathbf{y}) \preceq [\overline{\kappa}_\nu(d_\nu(\mathbf{x},\mathbf{y})) - 1] \nabla^2 f(\mathbf{x}). \tag{A.8}$$

Since $\mathbf{H}(\mathbf{x},\mathbf{y}) = \nabla^2 f(\mathbf{x})^{-1/2}\mathbf{G}(\mathbf{x},\mathbf{y})\nabla^2 f(\mathbf{x})^{-1/2}$, the last inequality implies

$$\|\mathbf{H}(\mathbf{x},\mathbf{y})\| \leq \max\left\{1 - \underline{\kappa}_\nu(d_\nu(\mathbf{x},\mathbf{y})), \overline{\kappa}_\nu(d_\nu(\mathbf{x},\mathbf{y})) - 1\right\}.$$

Let $C_{\max}(t) := \max\left\{1 - \underline{\kappa}_\nu(t), \overline{\kappa}_\nu(t) - 1\right\}$ be for $t \in [0,1)$. We consider two cases.

(a) For $\nu = 2$, since $e^{-t} + e^t \geq 2$, we have $\frac{1-e^{-t}}{t} + \frac{e^t-1}{t} \geq 2$, which implies $C_{\max}(t) = \overline{\kappa}_\nu(t) - 1 = \frac{e^t-1-t}{t}$. Hence, by Lemma A.2.1, we have $C_{\max}(t) \leq \left(\frac{3}{2} + \frac{t}{3}\right) te^t$, which leads to $R_\nu(t) := \left(\frac{3}{2} + \frac{t}{3}\right) e^t$.

(b) For $\nu \in (2,3]$, we have

$$\begin{aligned} C_{\max}(t) &= \max\{1 - \tfrac{2}{\nu t}\left[1 - (1 - \tfrac{\nu-2}{2}t)^{\frac{\nu}{\nu-2}}\right], \tfrac{2}{(4-\nu)t}\left[(1 - \tfrac{\nu-2}{2}t)^{\frac{4-\nu}{2-\nu}} - 1\right] - 1\} \\ &= \tfrac{2}{(4-\nu)t}\left[\left(1 - \tfrac{\nu-2}{2}t\right)^{\frac{4-\nu}{2-\nu}} - 1\right] - 1. \end{aligned}$$

Indeed, we show that $\frac{2}{(4-\nu)t}\left[(1 - \frac{\nu-2}{2}t)^{\frac{4-\nu}{2-\nu}} - 1\right] + \frac{2}{\nu t}\left[1 - (1 - \frac{\nu-2}{2}t)^{\frac{\nu}{\nu-2}}\right] \geq 2$. Let $u := \frac{4-\nu}{\nu-2} > 0, v := \frac{\nu}{\nu-2} > 0$ and $\tilde{t} := \frac{\nu-2}{2}t \in [0,1)$. The last inequality is equivalent to $\frac{1}{u}\left[\frac{1}{(1-\tilde{t})^u} - 1\right] + \frac{1}{v}\left[1 - (1-\tilde{t})^v\right] \geq 2\tilde{t}$. Consider the function $s(\tilde{t}) := \frac{1}{v} - \frac{1}{u} + \frac{1}{u(1-\tilde{t})^u} - \frac{(1-\tilde{t})^v}{v} - 2\tilde{t}$. Then it is suffices to prove that $s(\tilde{t}) \geq 0$. It is clear that $s'(\tilde{t}) = \frac{1}{(1-\tilde{t})^{u+1}} + (1-\tilde{t})^{v-1} - 2 = (1-\tilde{t})^{-\frac{2}{\nu-2}} + (1-\tilde{t})^{\frac{2}{\nu-2}} - 2 \geq 0$ for all $\tilde{t} \in [0,1)$. We obtain $s(\tilde{t}) \geq s(0) = 0$. Hence, $C_{\max}(t) = \frac{2}{(4-\nu)t}\left[\left(1 - \frac{\nu-2}{2}t\right)^{\frac{4-\nu}{2-\nu}} - 1\right] - 1$ and $R_\nu(t) = C_{\max}(t)/t$ as shown in (A.7). Let $r := u$, then $R_\nu(t) = \frac{\nu-2}{2}\psi_r(\tilde{t})$, where ψ_r is defined in Lemma A.2.1.

Putting (a) and (b) together, we obtain (A.6) with R_ν defined by (A.7). The boundedness of R_ν follows from Lemma A.2.1. □

A.2.2 The proof of Theorem 4.2.2: Convergence of damped Newton methods

The proof of this theorem is divided into two parts: computation of the step-size, and the proof the local quadratic convergence.

Computing the step-size τ_k: From Proposition 3.7.5, for any $\mathbf{x}^k, \mathbf{x}^{k+1} \in \mathrm{dom}(f)$, if $d_\nu(\mathbf{x}^k, \mathbf{x}^{k+1}) < \frac{2}{\nu-2}$, then we have

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \left\langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \right\rangle + \omega_\nu\left(d_\nu(\mathbf{x}^k, \mathbf{x}^{k+1})\right) \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2_{\mathbf{x}^k}.$$

Now, using (4.1), we have $\left\langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \right\rangle = -\tau_k \left(\|\nabla f(\mathbf{x}^k)\|^*_{\mathbf{x}^k} \right)^2 = -\tau_k \lambda_k^2$. On the other hand, we have

$$\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2_{\mathbf{x}^k} \overset{(4.1)}{=} \tau_k^2 \left\langle \nabla^2 f(\mathbf{x}^k)^{-1} \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^k) \right\rangle \overset{(4.3)}{=} \tau_k^2 \lambda_k^2,$$
$$\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2_2 \overset{(4.1)}{=} \tau_k^2 \left\langle \nabla^2 f(\mathbf{x}^k)^{-1} \nabla f(\mathbf{x}^k), \nabla^2 f(\mathbf{x}^k)^{-1} \nabla f(\mathbf{x}^k) \right\rangle \overset{(4.3)}{=} \tau_k^2 \beta_k^2.$$

Using the definition of $d_\nu(\cdot)$ in (3.8), the two last equalities, and (4.4), we can easily show that $d_\nu(\mathbf{x}^k, \mathbf{x}^{k+1}) = \tau_k d_k$. Substituting these relations into the first estimate, we obtain

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \left(\tau_k \lambda_k^2 - \omega_\nu \left(\tau_k d_k \right) \tau_k^2 \lambda_k^2 \right).$$

We consider the following cases:

(a) If $\nu = 2$, then, by (3.17), we have $\eta_k(\tau) := \lambda_k^2 \tau - \left(\frac{\lambda_k}{d_k} \right)^2 \left(e^{\tau d_k} - \tau d_k - 1 \right)$. This function attains the maximum at $\tau_k := \frac{\ln(1+d_k)}{d_k} \in (0,1)$ with

$$\eta_k(\tau_k) = \left(\frac{\lambda_k}{d_k} \right)^2 \Big[(1 + d_k) \ln(1 + d_k) - d_k \Big].$$

It is easy to check from the right-hand side that $\Delta_k := \eta_k(\tau_k) > 0$ for $\tau_k > 0$.

(b) If $\nu = 3$, by (3.17), we have $\eta_k(\tau) := \lambda_k^2 \tau + \left(\frac{\lambda_k}{d_k} \right)^2 \left[2\tau d_k + 4 \ln(1 - 0.5\tau d_k) \right]$ with $d_k = M_f \lambda_k$. We can show that $\eta_k(\tau)$ achieves the maximum at $\tau_k = \frac{1}{1+0.5 d_k} = \frac{1}{1+0.5 M_f \lambda_k} \in (0,1)$ with

$$\eta_k(\tau_k) = \frac{\lambda_k^2}{1 + 0.5 M_f \lambda_k} + \left(\frac{2}{M_f} \right)^2 \left[\frac{0.5 M_f \lambda_k}{1 + 0.5 M_f \lambda_k} + \ln \left(1 - \frac{0.5 M_f \lambda_k}{1 + 0.5 M_f \lambda_k} \right) \right].$$

We can also easily check from right-hand side that $\Delta_k := \eta_k(\tau_k) > 0$ for $\lambda_k > 0$.

(c) If $2 < \nu < 3$, then we have $d_k = M_f \lambda_k^{\nu-2} \beta_k^{3-\nu}$. By (3.17), we have

$$\eta_k(\tau) \;=\; \left(\lambda_k^2 + \frac{\lambda_k^2}{d_k} \frac{2}{4 - \nu} \right) \tau - \left(\frac{\lambda_k}{d_k} \right)^2 \frac{2}{(4-\nu)(3-\nu)} \left[\left(1 - \frac{\nu - 2}{2} \tau d_k \right)^{\frac{2(3-\nu)}{2-\nu}} - 1 \right].$$

Our aim is to find $\tau^* \in (0,1]$ by solving $\max_{\tau \in [0,1]} \eta_k(\tau)$. This problem always has a global solution. First, we compute the first- and the second-order derivatives of η_k as follows:

$$\eta_k'(\tau) = \lambda_k^2 \left[1 - \frac{2}{(\nu - 4) d_k} \left(1 - \left(1 - \frac{\nu - 2}{2} \tau d_k \right)^{\frac{\nu - 4}{\nu - 2}} \right) \right] \text{ and } \eta_k''(\tau) = -\lambda_k^2 \left(1 - \frac{\nu - 2}{2} \tau d_k \right)^{\frac{-2}{\nu - 2}}.$$

Let us set $\eta_k'(\tau_k) = 0$. Then, we get

$$\tau_k = \frac{2}{(\nu-2)d_k}\left[1 - \left(1 + \frac{4-\nu}{2}d_k\right)^{-\frac{\nu-2}{4-\nu}}\right] \in (0,1) \quad \text{(by the Bernoulli inequality)},$$

with

$$\eta_k(\tau_k) = \frac{2\lambda_k^2}{(\nu-2)d_k}\left\{1 - \frac{4-\nu}{2(3-\nu)}\left(1 + \frac{4-\nu}{2}d_k\right)^{2-\nu} + \frac{1}{(3-\nu)d_k}\left[1 - \left(1 + \frac{4-\nu}{2}d_k\right)^{2-\nu}\right]\right\}.$$

In addition, we can check that $\eta_k''(\tau_k) < 0$. Hence, the value of τ_k above achieves the maximum of $\eta_k(\cdot)$. Then, we have $\Delta_k := \eta_k(\tau_k) > \eta_k(0) = 0$.

The proof of local quadratic convergence: Let $\mathbf{x}_f^\star$ be the optimal solution of (2.1). We have

$$\begin{aligned}\|\mathbf{x}^{k+1} - \mathbf{x}_f^\star\|_{\mathbf{x}^k} &= \|\mathbf{x}^k - \tau_k \nabla^2 f(\mathbf{x}^k)^{-1}\nabla f(\mathbf{x}^k) - \mathbf{x}_f^\star\|_{\mathbf{x}^k} \\ &\leq (1-\tau_k)\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k} + \tau_k\|\mathbf{x}^k - \mathbf{x}_f^\star - \nabla^2 f(\mathbf{x}^k)^{-1}\nabla f(\mathbf{x}^k)\|_{\mathbf{x}^k}.\end{aligned}$$

Hence, we can write

$$\|\mathbf{x}^{k+1} - \mathbf{x}_f^\star\|_{\mathbf{x}^k} \leq (1-\tau_k)\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k} + \tau_k\|\nabla^2 f(\mathbf{x}^k)^{-1}\Big[\nabla f(\mathbf{x}_f^\star) - \nabla f(\mathbf{x}^k) - \nabla^2 f(\mathbf{x}^k)(\mathbf{x}_f^\star - \mathbf{x}^k)\Big]\|_{\mathbf{x}^k}. \tag{A.9}$$

Let us define $T_k := \left\|\nabla^2 f(\mathbf{x}^k)^{-1}\Big[\nabla f(\mathbf{x}_f^\star) - \nabla f(\mathbf{x}^k) - \nabla^2 f(\mathbf{x}^k)(\mathbf{x}_f^\star - \mathbf{x}^k)\Big]\right\|_{\mathbf{x}^k}$ and consider three cases as follows:

(a) For $\nu = 2$, using Corollary 3.7.3, we have $\left(\frac{1-e^{-\bar{d}_k}}{\bar{d}_k}\right)\nabla^2 f(\mathbf{x}^k) \preceq \int_0^1 \nabla^2 f(\mathbf{x}^k + t(\mathbf{x}_f^\star - \mathbf{x}^k))dt \preceq \left(\frac{e^{\bar{d}_k}-1}{\bar{d}_k}\right)\nabla^2 f(\mathbf{x}^k)$, where $\bar{d}_k := M_f\|\mathbf{x}^k - \mathbf{x}_f^\star\|_2$. Using the above inequality, we can show that

$$T_k \leq \max\left\{1 - \frac{1-e^{-\bar{d}_k}}{\bar{d}_k}, \frac{e^{\bar{d}_k}-1}{\bar{d}_k} - 1\right\}\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k} = \left(\frac{e^{\bar{d}_k} - 1 - \bar{d}_k}{\bar{d}_k^2}\right)\bar{d}_k\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}.$$

Let $\underline{\sigma}_k := \lambda_{\min}(\nabla^2 f(\mathbf{x}^k))$. We first derive

$$\begin{aligned}\|\nabla^2 f(\mathbf{x}^k)^{-1}\nabla f(\mathbf{x}^k)\|_2 &= \|\nabla^2 f(\mathbf{x}^k)^{-1}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}_f^\star))\|_2\\ &= \|\int_0^1 \nabla^2 f(\mathbf{x}^k)^{-1}\nabla^2 f(\mathbf{x}^k + t(\mathbf{x}_f^\star - \mathbf{x}^k))(\mathbf{x}^k - \mathbf{x}_f^\star)dt\|_2\\ &= \|\nabla^2 f(\mathbf{x}^k)^{-1/2}\mathbf{K}(\mathbf{x}^k, \mathbf{x}_f^\star)\nabla^2 f(\mathbf{x}^k)^{1/2}(\mathbf{x}^k - \mathbf{x}_f^\star)\|_2\\ &\leq \tfrac{1}{\sqrt{\underline{\sigma}_k}}\|\mathbf{K}(\mathbf{x}^k, \mathbf{x}_f^\star)\|\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}.\end{aligned}$$

where $\mathbf{K}(\mathbf{x}^k, \mathbf{x}_f^\star) := \int_0^1 \nabla^2 f(\mathbf{x}^k)^{-1/2}\nabla^2 f(\mathbf{x}^k + t(\mathbf{x}_f^\star - \mathbf{x}^k)\nabla^2 f(\mathbf{x}^k)^{-1/2}dt$. Using Corollary 3.7.3 and noting that $\bar{d}_k := M_f\|\mathbf{x}^k - \mathbf{x}_f^\star\|_2$, we can estimate $\|\mathbf{K}(\mathbf{x}^k, \mathbf{x}_f^\star)\| \leq \frac{e^{\bar{d}_k} - 1}{\bar{d}_k}$. Using the two last estimates, and the definition of d_k, we can derive

$$d_k \;= M_f\|\nabla^2 f(\mathbf{x}^k)^{-1}\nabla f(\mathbf{x}^k)\|_2 \leq \tfrac{M_f e^{\bar{d}_k - 1}}{\bar{d}_k\sqrt{\underline{\sigma}_k}}\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k} \leq M_f e\tfrac{\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}}{\sqrt{\underline{\sigma}_k}},$$

provided that $\bar{d}_k \leq 1$. Since, the step-size $\tau_k = \frac{1}{d_k}\ln(1 + d_k)$, we have $1 - \tau_k \leq \frac{d_k}{2} \leq \frac{M_f e\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}}{2\sqrt{\underline{\sigma}_k}}$. On the other hand, $\frac{e^{\bar{d}_k} - 1 - \bar{d}_k}{\bar{d}_k^2} \leq \frac{e}{2}$ for all $0 \leq \bar{d}_k \leq 1$. Substituting T_k into (A.9) and using these relations, we have

$$\|\mathbf{x}^{k+1} - \mathbf{x}_f^\star\|_{\mathbf{x}^k} \leq \tfrac{e}{2}\bar{d}_k\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k} + \tfrac{M_f e}{2}\tfrac{\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}^2}{\sqrt{\underline{\sigma}_k}},$$

provided that $\bar{d}_k \leq 1$. On the other hand, by Proposition 3.7.2, we have $\|\mathbf{x}^{k+1} - \mathbf{x}_f^\star\|_{\mathbf{x}^{k+1}} \leq e^{\frac{\bar{d}_{k+1} + \bar{d}_k}{2}}\|\mathbf{x}^{k+1} - \mathbf{x}_f^\star\|_{\mathbf{x}^k}$ and $\underline{\sigma}_{k+1}^{-1} \leq e^{\bar{d}_k + \bar{d}_{k+1}}\underline{\sigma}_k^{-1}$. In addition, $\bar{d}_k \leq \frac{M_f}{\sqrt{\underline{\sigma}_k}}\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}$ Combining the above inequalities, we finally get

$$\frac{\|\mathbf{x}^{k+1} - \mathbf{x}_f^\star\|_{\mathbf{x}^{k+1}}}{\sqrt{\underline{\sigma}_{k+1}}} \leq M_f e^{1 + \bar{d}_{k+1} + \bar{d}_k}\left(\frac{\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}}{\sqrt{\underline{\sigma}_k}}\right)^2,$$

which shows that $\left\{\frac{\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}}{\sqrt{\underline{\sigma}_k}}\right\}$ quadratically converges to zero locally. Since $\|\mathbf{x}^k - \mathbf{x}_f^\star\|_2 \leq \frac{\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}}{\sqrt{\underline{\sigma}_k}}$, we can also conclude that $\{\|\mathbf{x}^k - \mathbf{x}_f^\star\|_2\}$ quadratically converges to zero.

(b) For $\nu = 3$, we can follow [74]. However, for completeness, we give a short proof here. Using Corollary 3.7.3, we have $\left(1 - \frac{r_k}{2} + \frac{r_k^2}{12}\right)\nabla^2 f(\mathbf{x}^k) \preceq \int_0^1 \nabla^2 f(\mathbf{x}^k + t(\mathbf{x}_f^\star - \mathbf{x}^k))dt \preceq$

$\frac{1}{1-0.5r_k}\nabla^2 f(\mathbf{x}^k)$, where $r_k := M_f\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k} < 2$. Using the above inequality, we can show that

$$T_k \leq \max\left\{\frac{r_k}{2} - \frac{r_k^2}{12}, \frac{0.5r_k}{1-0.5r_k}\right\}\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k} = \frac{0.5M_f\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}^2}{1 - 0.5M_f\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}}.$$

Substituting T_k into (A.9) and using $\tau_k = \frac{1}{1+0.5M_f\lambda_k}$, we have

$$\|\mathbf{x}^{k+1} - \mathbf{x}_f^\star\|_{\mathbf{x}^k} \leq \frac{0.5M_f\lambda_k}{1+0.5M_f\lambda_k}\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k} + \frac{1}{1+0.5M_f\lambda_k}\left(\frac{0.5M_f\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}^2}{1-0.5M_f\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}}\right).$$

Next, we need to upper bound λ_k. Since $\nabla f(\mathbf{x}_f^\star) = 0$. Using Corollary 3.7.3, we can bound λ_k as

$$\begin{aligned}
\lambda_k &= \|\nabla f(\mathbf{x}^k)\|_{\mathbf{x}^k}^* = \|\nabla^2 f(\mathbf{x}^k)^{-1/2}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}_f^\star))\|_2 \\
&= \|\textstyle\int_0^1 \nabla^2 f(\mathbf{x}^k)^{-1/2}\nabla^2 f(\mathbf{x}^k + t(\mathbf{x}_f^\star - \mathbf{x}^k))(\mathbf{x}_f^\star - \mathbf{x}^k)dt\|_2 \\
&\leq \|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}\|\textstyle\int_0^1 \nabla^2 f(\mathbf{x}^k)^{-1/2}\nabla^2 f(\mathbf{x}^k + t(\mathbf{x}_f^\star - \mathbf{x}^k))\nabla^2 f(\mathbf{x}^k)^{-1/2}dt\|_2 \\
&\overset{\text{Corollary 3.7.3}}{\leq} \tfrac{\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}}{1-0.5M_f\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}} \leq 2\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k},
\end{aligned}$$

provided that $M_f\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k} < 1$. Overestimating the above inequality using this bound, we get

$$\begin{aligned}
\|\mathbf{x}^{k+1} - \mathbf{x}_f^\star\|_{\mathbf{x}^k} &\leq 0.5M_f\lambda_k\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k} + \tfrac{0.5M_f\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}^2}{1-0.5M_f\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}} \\
&\leq M_f\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}^2 + M_f\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}^2 = 2M_f\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}^2,
\end{aligned}$$

provided that $M_f\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k} < 1$. On the other hand, we can also estimate $\|\mathbf{x}^{k+1} - \mathbf{x}_f^\star\|_{\mathbf{x}^{k+1}} \leq \frac{\|\mathbf{x}^{k+1} - \mathbf{x}_f^\star\|_{\mathbf{x}^k}}{1-0.5M_f\left(\|\mathbf{x}^{k+1} - \mathbf{x}_f^\star\|_{\mathbf{x}^k} + \|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}\right)}$. Combining the last two inequalities, we get

$$\|\mathbf{x}^{k+1} - \mathbf{x}_f^\star\|_{\mathbf{x}^{k+1}} \leq \frac{2M_f\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}^2}{1 - 2M_f\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}^2 - 0.5M_f\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}}$$

The right-hand side function $\psi(t) = \frac{2M_f}{1-2M_f t^2 - 0.5M_f t} \leq 4M_f$ on $t \in \left[0, \frac{1}{2M_f}\right]$. Hence, if $\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k} \leq \frac{1}{2M_f}$, then $\|\mathbf{x}^{k+1} - \mathbf{x}_f^\star\|_{\mathbf{x}^{k+1}} \leq 4M_f\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}^2$. This shows that if $\mathbf{x}^0 \in \mathrm{dom}(f)$ is chosen such that $\|\mathbf{x}^0 - \mathbf{x}_f^\star\|_{\mathbf{x}^0} < \frac{1}{4M_f}$, then $\{\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}\}$ quadratically converges to zero.

(c) For $\nu \in (2, 3)$, with the same argument as in the proof of Theorem 4.2.3, we can show that

$$\|\mathbf{x}^{k+1} - \mathbf{x}_f^\star\|_{\mathbf{x}^k} \leq R_\nu(d_k)d_k\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k},$$

where R_ν is defined by (A.7) and $d_k := M_f\|\mathbf{x}^k-\mathbf{x}_f^\star\|_2^{3-\nu}\|\mathbf{x}^k-\mathbf{x}_f^\star\|_{\mathbf{x}^k}^{\nu-2}$. Using again the argument as in the proof of Theorem 4.2.3, we have

$$\frac{\|\mathbf{x}^{k+1}-\mathbf{x}_f^\star\|_{\mathbf{x}^{k+1}}}{\underline{\sigma}_{k+1}^{\frac{3-\nu}{2}}} \le C_\nu(d_k, \|\mathbf{x}^k-\mathbf{x}_f^\star\|_{\mathbf{x}^k})\left(\frac{\|\mathbf{x}^k-\mathbf{x}_f^\star\|_{\mathbf{x}^k}}{\underline{\sigma}_k^{\frac{3-\nu}{2}}}\right)^2.$$

Here, $C_\nu(\cdot,\cdot)$ is a given function deriving from R_ν. Under the condition that d_k and $\|\mathbf{x}^k-\mathbf{x}_f^\star\|_{\mathbf{x}^k}$ are sufficiently small, we can show that $C_\nu(d_k, \|\mathbf{x}^k-\mathbf{x}_f^\star\|_{\mathbf{x}^k}) \le \bar{C}_\nu$. Hence, the last inequality shows that $\left\{\frac{\|\mathbf{x}^k-\mathbf{x}_f^\star\|_{\mathbf{x}^k}}{\underline{\sigma}_k^{\frac{3-\nu}{2}}}\right\}$ quadratically converges to zero. Since $\underline{\sigma}_k^{\frac{3-\nu}{2}}\|\mathbf{x}^k-\mathbf{x}_f^\star\|_{\mathbf{H}_k} \le \|\mathbf{x}^k-\mathbf{x}_f^\star\|_{\mathbf{x}^k}$, where $\mathbf{H}_k := \nabla^2 f(\mathbf{x}^k)^{\nu-2}$, we have $\|\mathbf{x}^k-\mathbf{x}_f^\star\|_{\mathbf{H}_k} \le \frac{\|\mathbf{x}^k-\mathbf{x}_f^\star\|_{\mathbf{x}^k}}{\underline{\sigma}_k^{\frac{3-\nu}{2}}}$. Hence, we can conclude that $\{\|\mathbf{x}^k-\mathbf{x}_f^\star\|_{\mathbf{H}_k}\}$ also locally converges to zero at a quadratic rate.

A.2.3 The proof of Theorem 4.2.3: Convergence of full Newton methods

We divide this proof into two parts: the quadratic convergence of $\left\{\frac{\lambda_k}{\underline{\sigma}_k^{\frac{3-\nu}{2}}}\right\}$, and the quadratic convergence of $\left\{\|\mathbf{x}^k-\mathbf{x}_f^\star\|_{\mathbf{H}_k}\right\}$.

The quadratic convergence of $\left\{\frac{\lambda_k}{\underline{\sigma}_k^{\frac{3-\nu}{2}}}\right\}$: Since the full-step Newton scheme updates $\mathbf{x}^{k+1} := \mathbf{x}^k - \nabla^2 f(\mathbf{x}^k)^{-1}\nabla f(\mathbf{x}^k)$, if we denote by $n_{\mathrm{nt}}^k = \mathbf{x}^{k+1}-\mathbf{x}^k = -\nabla^2 f(\mathbf{x}^k)^{-1}\nabla f(\mathbf{x}^k)$, then the last expression leads to $\nabla f(\mathbf{x}^k) + \nabla^2 f(\mathbf{x}^k)n_{\mathrm{nt}}^k = 0$. In addition, $\|n_{\mathrm{nt}}^k\|_{\mathbf{x}^k} = \|\nabla f(\mathbf{x}^k)\|_{\mathbf{x}^k}^* = \lambda_k$.

First, by $\nabla f(\mathbf{x}^k) + \nabla^2 f(\mathbf{x}^k)n_{\mathrm{nt}}^k = 0$ and the mean-value theorem, we can show that

$$\nabla f(\mathbf{x}^{k+1}) = \nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^k) - \nabla^2 f(\mathbf{x}^k)n_{\mathrm{nt}}^k = \int_0^1 \left[\nabla^2 f(\mathbf{x}^k + tn_{\mathrm{nt}}^k) - \nabla^2 f(\mathbf{x}^k)\right] n_{\mathrm{nt}}^k dt.$$

Let us define

$$\mathbf{G}_k := \int_0^1 \left[\nabla^2 f(\mathbf{x}^k + tn_{\mathrm{nt}}^k) - \nabla^2 f(\mathbf{x}^k)\right] dt \text{ and } \mathbf{H}_k := \nabla^2 f(\mathbf{x}^k)^{-1/2}\mathbf{G}_k\nabla^2 f(\mathbf{x}^k)^{-1/2}.$$

Then, the above estimate implies $\nabla f(\mathbf{x}^{k+1}) = \mathbf{G}_k n_{\mathrm{nt}}^k$. Hence, we can show that

$$\begin{aligned}\left[\|\nabla f(\mathbf{x}^{k+1})\|_{\mathbf{x}^k}^*\right]^2 &= \left\langle \nabla^2 f(\mathbf{x}^k)^{-1}\mathbf{G}_k n_{\mathrm{nt}}^k, \mathbf{G}_k n_{\mathrm{nt}}^k\right\rangle = \left\langle \mathbf{H}_k\nabla^2 f(\mathbf{x}^k)^{1/2}n_{\mathrm{nt}}^k, \mathbf{H}_k\nabla^2 f(\mathbf{x}^k)^{1/2}n_{\mathrm{nt}}^k\right\rangle \\ &\le \|\mathbf{H}_k\|^2\|n_{\mathrm{nt}}^k\|_{\mathbf{x}^k}^2 = \|\mathbf{H}_k\|^2\lambda_k^2.\end{aligned}$$

By Lemma A.2.2, we have $\|\mathbf{H}_k\| \leq R_\nu(d_k)d_k$, where R_ν is defined by (A.7). Combining the two last inequalities and using Proposition 3.7.2, we consider the following cases:

(a) If $\nu = 2$, then we have $\lambda_{k+1}^2 \leq e^{d_2^k}\left[\|\nabla f(\mathbf{x}^{k+1})\|_{\mathbf{x}^k}^*\right]^2$, which implies $\lambda_{k+1} \leq e^{\frac{d_2^k}{2}} R_2(d_2^k)d_2^k\lambda_k$. Note that $\lambda_k \geq \frac{\sqrt{\underline{\sigma}_k}d_2^k}{M_f}$ and $\frac{1}{\underline{\sigma}_{k+1}} \leq \frac{e^{d_2^k}}{\underline{\sigma}_k}$. Based on the above inequality, we have

$$\frac{\lambda_{k+1}}{\sqrt{\underline{\sigma}_{k+1}}} \leq M_f R_2(d_2^k)e^{d_2^k}\left(\frac{\lambda_k}{\sqrt{\underline{\sigma}_k}}\right)^2.$$

By a numerical calculation, we can easily check that if $d_k < d_2^\star \approx 0.12964$, then

$$\frac{\lambda_{k+1}}{\sqrt{\underline{\sigma}_{k+1}}} \leq 2M_f\left(\frac{\lambda_k}{\sqrt{\underline{\sigma}_k}}\right)^2.$$

Consequently, if $\frac{\lambda_0}{\sqrt{\underline{\sigma}_0}} < \frac{1}{M_f}\min\{d_2^\star, 0.5\} = \frac{d_2^\star}{M_f}$, then we can prove

$$d_2^{k+1} \leq d_2^k \text{ and } \frac{\lambda_{k+1}}{\sqrt{\underline{\sigma}_{k+1}}} \leq \frac{\lambda_k}{\sqrt{\underline{\sigma}_k}},$$

by induction. Under the condition $\frac{\lambda_0}{\sqrt{\underline{\sigma}_0}} < \frac{d_2^\star}{M_f}$, the above inequality shows that the ratio $\{\frac{\lambda_k}{\sqrt{\underline{\sigma}_k}}\}$ converges to zero at a quadratic rate.

Now, if $\nu > 2$, then we consider different cases. Note that

$$\lambda_{k+1}^2 \leq \left(1 - \frac{\nu-2}{2}d_k\right)^{\frac{-2}{\nu-2}}\left[\|\nabla f(\mathbf{x}^{k+1})\|_{\mathbf{x}^k}^*\right]^2,$$

which follows that

$$\lambda_{k+1} \leq \left(1 - \frac{\nu-2}{2}d_k\right)^{\frac{-1}{\nu-2}} R_\nu(d_k)d_k\lambda_k. \tag{A.10}$$

Note that $d_k = M_f\beta_k^{3-\nu}\lambda_k^{\nu-2}$ and $\underline{\sigma}_{k+1}^{-1} \leq \left(1 - \frac{\nu-2}{2}d_k\right)^{\frac{-2}{\nu-2}}\underline{\sigma}_k^{-1}$. Based on these relations and (A.10) we can argue as follows:

(b) If $2 < \nu < 3$, then $\lambda_k \geq \beta_k\sqrt{\underline{\sigma}_k}$, which follows that $d_k \leq M_f\underline{\sigma}_k^{-\frac{3-\nu}{2}}\lambda_k$. Hence,

$$\frac{\lambda_{k+1}}{\underline{\sigma}_{k+1}^{\frac{3-\nu}{2}}} \leq \left(1 - \frac{\nu-2}{2}d_k\right)^{-\frac{4-\nu}{\nu-2}} R_\nu(d_k)M_f\left(\frac{\lambda_k}{\underline{\sigma}_k^{\frac{3-\nu}{2}}}\right)^2.$$

If $d_k < d_\nu^\star$, where $d_\nu^\star$ is the unique solution to the equation

$$\left(1 - \frac{\nu-2}{2} d_k\right)^{-\frac{4-\nu}{\nu-2}} R_\nu(d_k) = 2,$$

then $\underline{\sigma}_{k+1}^{-\frac{3-\nu}{2}} \lambda_{k+1} \leq 2M_f \left(\underline{\sigma}_k^{-\frac{3-\nu}{2}} \lambda_k\right)^2$. Note that it is straightforward to check that this equation always admits a positive solution. Hence, if we choose $\mathbf{x}^0 \in \mathrm{dom}(f)$ such that $\underline{\sigma}_0^{-\frac{3-\nu}{2}} \lambda_0 < \frac{1}{M_f} \min\{d_\nu^\star, 0.5\}$, then we can prove the following two inequalities together by induction:

$$d_k \leq d_{k+1} \text{ and } \underline{\sigma}_{k+1}^{-\frac{3-\nu}{2}} \lambda_{k+1} \leq \underline{\sigma}_k^{-\frac{3-\nu}{2}} \lambda_k.$$

In addition, the above inequality also shows that $\{\underline{\sigma}_k^{-\frac{3-\nu}{2}} \lambda_k\}$ quadratically converges to zero.

(c) If $\nu = 3$, then $d_k = M_f \lambda_k$, and

$$\lambda_{k+1} \leq (1 - 0.5 d_k)^{-1} R_3(d_k) d_k \lambda_k = M_f \frac{R_3(d_k)}{1 - 0.5 d_k} \lambda_k^2.$$

Directly checking the right-hand side of the above estimate, one can show that if $d_k < d_3^\star = 1$, then $\lambda_{k+1} \leq 2M_f \lambda_k^2$. Hence, if $\lambda_0 < \frac{1}{M_f} \min\{d_3^\star, 0.5\} = \frac{1}{2M_f}$, then we can prove the following two inequalities together by induction:

$$d_{k+1} \leq d_k \text{ and } \lambda_{k+1} \leq \lambda_k.$$

Moreover, the first inequality above also shows that $\{\lambda_k\}$ converges to zero quadratically.

The quadratic convergence of $\left\{\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{H}_k}\right\}$: First, using Proposition 3.7.4 with $\mathbf{x} := \mathbf{x}^k$ and $\mathbf{y} = \mathbf{x}_f^\star$, and noting that $\nabla f(\mathbf{x}_f^\star) = 0$, we have

$$\bar{\kappa}_\nu(-d_\nu(\mathbf{x}^k, \mathbf{x}_f^\star)) \|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}^2 \leq \left\langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}_f^\star \right\rangle \leq \|\nabla f(\mathbf{x}^k)\|_{\mathbf{x}^k}^* \|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k},$$

where the last inequality follows from the Cauchy-Schwarz inequality. Hence, we obtain

$$\bar{\kappa}_\nu(-d_\nu(\mathbf{x}^k, \mathbf{x}_f^\star)) \|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k} \leq \|\nabla f(\mathbf{x}^k)\|_{\mathbf{x}^k}^* = \lambda_k. \tag{A.11}$$

We consider three cases:

(1) When $\nu = 2$, we have $\bar{\kappa}_\nu(\tau) = \frac{e^\tau - 1}{\tau}$. Hence, $\bar{\kappa}_\nu(-d_\nu(\mathbf{x}^k, \mathbf{x}_f^\star)) = \frac{1 - e^{-d_\nu(\mathbf{x}^k, \mathbf{x}_f^\star)}}{d_\nu(\mathbf{x}^k, \mathbf{x}_f^\star)} \geq 1 - \frac{d_\nu(\mathbf{x}^k, \mathbf{x}_f^\star)}{2} \geq \frac{1}{2}$ whenever $d_\nu(\mathbf{x}^k, \mathbf{x}_f^\star) \leq 1$. Using this inequality in (A.11), we have $\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k} \leq 2\|\nabla f(\mathbf{x}^k)\|_{\mathbf{x}^k}^* = 2\lambda_k$ provided that $d_\nu(\mathbf{x}^k, \mathbf{x}_f^\star) \leq 1$. One the other hand, by the definition of $\underline{\sigma}_k$, we have $\sqrt{\underline{\sigma}_k}\|\mathbf{x}^k - \mathbf{x}_f^\star\|_2 \leq \|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}$. Combining the two last inequalities, we obtain $\|\mathbf{x}^k - \mathbf{x}_f^\star\|_2 \leq \frac{2\lambda_k}{\sqrt{\underline{\sigma}_k}}$ provided that $d_\nu(\mathbf{x}^k, \mathbf{x}_f^\star) \leq 1$. Since $\{\frac{\lambda_k}{\sqrt{\underline{\sigma}_k}}\}$ locally converges to zero at a quadratic rate, the last relation also shows that $\left\{\|\mathbf{x}^k - \mathbf{x}_f^\star\|_2\right\}$ also locally converges to zero at a quadratic rate.

(2) For $\nu = 3$, we have $\bar{\kappa}_\nu(-d_\nu(\mathbf{x}^k, \mathbf{x}_f^\star)) = \frac{1}{1 + 0.5 d_\nu(\mathbf{x}^k, \mathbf{x}_f^\star)}$ and $d_\nu(\mathbf{x}^k, \mathbf{x}_f^\star) = M_f\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}$. Hence, from (A.11), we obtain $\frac{\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}}{1 + 0.5 M_f \|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}} \leq \lambda_k$. This implies $\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k} \leq \frac{\lambda_k}{1 - 0.5 M_f \lambda_k}$ as long as $0.5 M_f \lambda_k < 1$. Clearly, since λ_k locally converges to zero at a quadratic rate, $\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k}$ also locally converges to zero at a quadratic rate.

(3) For $2 < \nu < 3$, we have $\bar{\kappa}_\nu(-d_\nu(\mathbf{x}^k, \mathbf{x}_f^\star)) = \left(\frac{2}{\nu - 4}\right) \frac{\left(1 + \frac{\nu-2}{2} d_\nu(\mathbf{x}^k, \mathbf{x}_f^\star)\right)^{\frac{\nu-4}{\nu-2}} - 1}{d_\nu(\mathbf{x}^k, \mathbf{x}_f^\star)} \geq 1 - 0.5 d_\nu(\mathbf{x}^k, \mathbf{x}_f^\star) \geq 0.5$ provided that $d_\nu(\mathbf{x}^k, \mathbf{x}_f^\star) < 1$. Similar to the case $\nu = 2$, we have $\underline{\sigma}_k^{\frac{3-\nu}{2}} \|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{H}_k} \leq \|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{x}^k} \leq 2\lambda_k$, where $\mathbf{H}_k := \nabla^2 f(\mathbf{x}^k)^{\nu-2}$. Hence, $\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{H}_k} \leq \frac{2\lambda_k}{\underline{\sigma}_k^{\frac{3-\nu}{2}}}$. Since $\left\{\frac{\lambda_k}{\underline{\sigma}_k^{\frac{3-\nu}{2}}}\right\}$ locally converges to zero at a quadratic rate, $\left\{\|\mathbf{x}^k - \mathbf{x}_f^\star\|_{\mathbf{H}_k}\right\}$ also locally converges to zero at a quadratic rate.

A.2.4 The proof of Theorem 4.3.1: Solution existence and uniqueness

Consider a sublevel set $\mathcal{L}_F(\mathbf{x}) := \{\mathbf{y} \in \mathrm{dom}(F) \mid F(\mathbf{y}) \leq F(\mathbf{x})\}$ of F in (2.7). For any $\mathbf{y} \in \mathcal{L}_F(\mathbf{x})$ and $\mathbf{v} \in \partial g(\mathbf{x})$, by (3.16) and the convexity of g, we have

$$F(\mathbf{x}) \geq F(\mathbf{y}) \geq F(\mathbf{x}) + \langle \nabla f(\mathbf{x}) + \mathbf{v}, \mathbf{y} - \mathbf{x} \rangle + \omega_\nu\left(-d_\nu(\mathbf{x}, \mathbf{y})\right) \|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}}^2.$$

By the Cauchy-Schwarz inequality, we have

$$\omega_\nu\left(-d_\nu(\mathbf{x}, \mathbf{y})\right) \|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}} \leq \|\nabla f(\mathbf{x}) + \mathbf{v}\|_{\mathbf{x}}^*. \tag{A.12}$$

Now, using the assumption $\nabla^2 f(\mathbf{x}) \succ 0$ for some $\mathbf{x} \in \mathrm{dom}(f)$, we have $\sigma_{\min}(\mathbf{x}) := \lambda_{\min}(\nabla^2 f(\mathbf{x})) > 0$, the smallest eigenvalue of $\nabla^2 f(\mathbf{x})$.

(a) If $\nu = 2$, then $d_2(\mathbf{x}, \mathbf{y}) = M_f \|\mathbf{y} - \mathbf{x}\|_2 \leq \frac{M_f}{\sqrt{\sigma_{\min}(\mathbf{x})}} \|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}}$. This estimate together with (A.12) imply

$$\omega_2 \left(-d_2(\mathbf{x}, \mathbf{y})\right) d_2(\mathbf{x}, \mathbf{y}) \leq \frac{M_f}{\sqrt{\sigma_{\min}(\mathbf{x})}} \|\nabla f(\mathbf{x}) + \mathbf{v}\|_{\mathbf{x}}^* = \frac{M_f}{\sqrt{\sigma_{\min}(\mathbf{x})}} \lambda(\mathbf{x}). \tag{A.13}$$

We consider the function $s_2(t) := \omega_2(-t)t = 1 - \frac{1-e^{-t}}{t}$. Clearly, $s_2'(t) = \frac{e^t - t - 1}{t^2 e^t} > 0$ for all $t \in \mathbb{R}_+$. Hence, $s_2(t)$ is increasing on $\mathbb{R}_+$. However, $s_2(t) < 1$ and $\lim\limits_{t \to +\infty} s_2(t) = 1$. Therefore, if $\frac{M_f}{\sqrt{\sigma_{\min}(\mathbf{x})}} \lambda(\mathbf{x}) < 1$, then the equation $s_2(t) - \frac{M_f}{\sqrt{\sigma_{\min}(\mathbf{x})}} \lambda(\mathbf{x}) = 0$ has a unique solution $t^* \in (0, +\infty)$. In this case, for $0 \leq d_2(\mathbf{x}, \mathbf{y}) \leq t^*$, (A.13) holds. This condition leads to $M_f \|\mathbf{y} - \mathbf{x}\|_2 \leq t^* < +\infty$, which implies that the sublevel set $\mathcal{L}_F(\mathbf{x})$ is bounded. Consequently, solution $\mathbf{x}^\star$ of (2.7) exists.

(b) If $2 < \nu \leq 3$, then

$$d_\nu(\mathbf{x}, \mathbf{y}) \leq \frac{M_f}{\sigma_{\min}(\mathbf{x})^{\frac{3-\nu}{2}}} \|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}}.$$

This inequality together with (A.12) imply

$$\omega_\nu \left(-d_\nu(\mathbf{x}, \mathbf{y})\right) d_\nu(\mathbf{x}, \mathbf{y}) \leq \frac{M_f}{\sigma_{\min}(\mathbf{x})^{\frac{3-\nu}{2}}} \|\nabla f(\mathbf{x}) + \mathbf{v}\|_{\mathbf{x}}^* = \frac{M_f}{\sigma_{\min}(\mathbf{x})^{\frac{3-\nu}{2}}} \lambda(\mathbf{x}).$$

We consider $s_\nu(t) := \omega_\nu(-t)t$. After a few elementary calculations, we can easily check that s_ν is increasing on $\mathbb{R}_+$ and $s_\nu(t) < \frac{2}{4-\nu}$ for all $t > 0$, and $\lim\limits_{t \to +\infty} s_\nu(t) = \frac{2}{4-\nu}$. Hence, if $\frac{M_f}{\sigma_{\min}(\mathbf{x})^{\frac{3-\nu}{2}}} \lambda(\mathbf{x}) < \frac{2}{4-\nu}$, then, similar to Case (a), we can show that solution $\mathbf{x}^\star$ of (2.7) exists. This condition implies that $\lambda(\mathbf{x}) < \frac{2\sigma_{\min}(\mathbf{x})^{\frac{3-\nu}{2}}}{(4-\nu)M_f}$. Especially, when $\nu = 3$, this condition becomes $\lambda(\mathbf{x}) < \frac{2}{M_f}$.

Note that the condition on $\lambda(\mathbf{x})$ in both cases (a) and (b) can be unified. The uniqueness of the solution $\mathbf{x}^\star$ in these cases follows from the strict convexity of F.

A.2.5 The proof of Theorem 4.3.2: Convergence of the damped PN method

Given $\mathbf{H} \in \mathcal{S}^p_{++}$ and a proper, closed, and convex function $g : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$, slightly different from (2.6), we define

$$\mathcal{P}^g_{\mathbf{H}}(\mathbf{u}) := (\mathbf{H} + \partial g)^{-1}(\mathbf{u}) = \operatorname{argmin}_{\mathbf{x}}\{g(\mathbf{x}) + \tfrac{1}{2}\langle \mathbf{H}\mathbf{x}, \mathbf{x}\rangle - \langle \mathbf{u}, \mathbf{x}\rangle\}.$$

If $\mathbf{H} = \nabla^2 f(\mathbf{x})$ is the Hessian mapping of a strictly convex function f, then we can also write $\mathcal{P}_{\nabla^2 f(\mathbf{x})}(\mathbf{u})$ shortly as $\mathcal{P}_{\mathbf{x}}(\mathbf{u})$ for our notational convenience. The following lemma will be used in the sequel, whose proof can be found in [102].

Lemma A.2.3. Let $g : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ be a proper, closed, and convex function, and $\mathbf{H} \in \mathcal{S}^p_{++}$. Then, the mapping $\mathcal{P}^g_{\mathbf{H}}$ defined above is non-expansive with respect to the weighted norm defined by $\mathbf{H}$, i.e., for any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$, we have

$$\|\mathcal{P}^g_{\mathbf{H}}(\mathbf{u}) - \mathcal{P}^g_{\mathbf{H}}(\mathbf{v})\|_{\mathbf{H}} \leq \|\mathbf{u} - \mathbf{v}\|^*_{\mathbf{H}}. \tag{A.14}$$

Let us define

$$S_{\mathbf{x}}(\mathbf{u}) := \nabla^2 f(\mathbf{x})\mathbf{u} - \nabla f(\mathbf{u}) \quad \text{and} \quad e_{\mathbf{x}}(\mathbf{u}, \mathbf{v}) := [\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{u})](\mathbf{v} - \mathbf{u}), \tag{A.15}$$

for any vectors $\mathbf{x}, \mathbf{u} \in \operatorname{dom}(f)$ and $\mathbf{v} \in \mathbb{R}^p$. We now prove Theorem 4.3.2 in the main text.

The proof of Theorem 4.3.2 is divided into two parts: computation of the step-size, and the proof the local quadratic convergence.

Computing the step-size τ_k: Since $\mathbf{z}^k$ satisfies the optimality condition (4.10), we have

$$-\nabla f(\mathbf{x}^k) - \nabla^2 f(\mathbf{x}^k) n^k_{\text{pnt}} \in \partial g(\mathbf{z}^k).$$

Using Proposition 3.7.5 we obtain

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \tau_k \left\langle \nabla f(\mathbf{x}^k), n^k_{\text{pnt}} \right\rangle + \omega_\nu(\tau_k d_k)\tau_k^2 \lambda_k^2.$$

Since $\mathbf{x}^{k+1} = (1-\tau_k)\mathbf{x}^k + \tau_k \mathbf{z}^k$, using this relation and the convexity of g, we have

$$g(\mathbf{x}^{k+1}) \leq g(\mathbf{x}^k) - \tau_k \left\langle \nabla f(\mathbf{x}^k) + \nabla^2 f(\mathbf{x}^k) n^k_{\text{pnt}}, n^k_{\text{pnt}} \right\rangle.$$

Summing up the last two inequalities, we obtain the following estimate

$$F(\mathbf{x}^{k+1}) \leq F(\mathbf{x}^k) - \eta_k(\tau_k).$$

With the same argument as in the proof of Theorem 4.2.2, we obtain the conclusion of Theorem 4.3.2.

The proof of local quadratic convergence: We consider the distance between $\mathbf{x}^{k+1}$ and $\mathbf{x}^\star$ measured by $\|\mathbf{x}^{k+1} - \mathbf{x}^\star\|_{\mathbf{x}^\star}$. By the definition of $\mathbf{x}^{k+1}$, we have

$$\|\mathbf{x}^{k+1} - \mathbf{x}^\star\|_{\mathbf{x}^\star} \leq (1-\tau_k)\|\mathbf{x}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star} + \tau_k \|\mathbf{z}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star}. \tag{A.16}$$

Using the new notations in (A.15), it follows from the optimality condition (4.8) and (4.10) that $\mathbf{z}^k = \mathcal{P}^g_{\mathbf{x}^\star}(S_{\mathbf{x}^\star}(\mathbf{x}^k) + e_{\mathbf{x}^\star}(\mathbf{x}^k, \mathbf{z}^k))$ and $\mathbf{x}^\star = \mathcal{P}^g_{\mathbf{x}^\star}(S_{\mathbf{x}^\star}(\mathbf{x}^\star))$. By Lemma A.2.3 and the triangle inequality, we can show that

$$\|\mathbf{z}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star} \leq \|S_{\mathbf{x}^\star}(\mathbf{x}^k) - S_{\mathbf{x}^\star}(\mathbf{x}^\star)\|^*_{\mathbf{x}^\star} + \|e_{\mathbf{x}^\star}(\mathbf{x}^k, \mathbf{z}^k)\|^*_{\mathbf{x}^\star}. \tag{A.17}$$

By following the same argument as in [102], if we apply Lemma A.2.2, then we can derive

$$\|S_{\mathbf{x}^\star}(\mathbf{x}^k) - S_{\mathbf{x}^\star}(\mathbf{x}^\star)\|^*_{\mathbf{x}^\star} \leq R_\nu(d_\nu(\mathbf{x}^\star, \mathbf{x}^k)) d_\nu(\mathbf{x}^\star, \mathbf{x}^k) \|\mathbf{x}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star}, \tag{A.18}$$

where $R_\nu(\cdot)$ is defined by (A.7).

Next, using the same argument as the proof of (A.25) in Theorem 4.3.3 below, we can bound the second term $\|e_{\mathbf{x}^\star}(\mathbf{x}^k, \mathbf{z}^k)\|^*_{\mathbf{x}^\star}$ of (A.17) as

$$\|e_{\mathbf{x}^\star}(\mathbf{x}^k, \mathbf{z}^k)\|^*_{\mathbf{x}^\star} \leq \begin{cases} \left[\left(1 - \frac{\nu-2}{2} d_\nu(\mathbf{x}^\star, \mathbf{x}^k)\right)^{\frac{-2}{\nu-2}} - 1 \right] \|\mathbf{z}^k - \mathbf{x}^k\|_{\mathbf{x}^\star}, & \text{if } \nu > 2 \\ \left(e^{d_\nu(\mathbf{x}^\star, \mathbf{x}^k)} - 1\right) \|\mathbf{z}^k - \mathbf{x}^k\|_{\mathbf{x}^\star} & \text{if } \nu = 2. \end{cases}$$

Combining this inequality, (A.17) (A.18), and the triangle inequality

$$\|\mathbf{z}^k - \mathbf{x}^k\|_{\mathbf{x}^\star} \leq \|\mathbf{z}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star} + \|\mathbf{x}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star},$$

we obtain

$$\|\mathbf{z}^k - \mathbf{x}^k\|_{\mathbf{x}^\star} \leq \hat{R}_\nu(d_\nu(\mathbf{x}^\star, \mathbf{x}^k))\|\mathbf{x}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star} \tag{A.19}$$

and

$$\|\mathbf{z}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star} \leq \tilde{R}_\nu(d_\nu(\mathbf{x}^\star, \mathbf{x}^k))d_\nu(\mathbf{x}^\star, \mathbf{x}^k)\|\mathbf{x}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star}, \tag{A.20}$$

where $\hat{R}_\nu$ and $\tilde{R}_\nu$ are defined as

$$\hat{R}_\nu(t) := \begin{cases} \frac{tR_\nu(t)+1}{2-\left(1-\frac{\nu-2}{2}t\right)^{\frac{-2}{\nu-2}}}, & \text{if } \nu > 2 \\ \frac{tR_\nu(t)+1}{2-e^t} & \text{if } \nu = 2 \end{cases} \text{ and } \tilde{R}_\nu(t) := \begin{cases} \frac{tR_\nu(t)-1+\left(1-\frac{\nu-2}{2}t\right)^{\frac{-2}{\nu-2}}}{t\left(2-\left(1-\frac{\nu-2}{2}t\right)^{\frac{-2}{\nu-2}}\right)}, & \text{if } \nu > 2 \\ \frac{tR_\nu(t)-1+e^t}{t(2-e^t)} & \text{if } \nu = 2 \end{cases}$$

respectively.

By using Lemma A.2.2 and after some simple calculations, one can show that there exists a constant $c_\nu \in (0, +\infty)$ such that if $t \in [0, \bar{d}_\nu]$, then both $\hat{R}_\nu(t)$ and $\tilde{R}_\nu(t) \in [0, c_\nu]$, where $\bar{d}_\nu := \frac{2}{\nu-2}(1 - 0.6^{\frac{\nu-2}{2}})$ for $\nu \geq 2$ (when $t \to 0+$ or $\nu = 2$, we consider the limit). Using this bound, (A.16) (A.20) and the fact that $\tau_k \leq 1$, we can bound

$$\|\mathbf{x}^{k+1} - \mathbf{x}^\star\|_{\mathbf{x}^\star} \leq \left[(1 - \tau_k) + c_\nu d_\nu(\mathbf{x}^\star, \mathbf{x}^k)\right] \|\mathbf{x}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star}. \tag{A.21}$$

Let $\underline{\sigma}^\star := \sigma_{\min}(\nabla^2 f(\mathbf{x}^\star))$ be the smallest eigenvalue of $\nabla^2 f(\mathbf{x}^\star)$. We consider the following cases:

(a) If $\nu = 2$, for $0 \leq d_\nu(\mathbf{x}^\star, \mathbf{x}^k) \leq \bar{d}_\nu$, we can bound $1 - \tau_k$ as

$$1 - \tau_k \;\; = 1 - \tfrac{\ln(1+d_k)}{d_k} \leq \tfrac{d_k}{2} = \tfrac{M_f}{2}\|\mathbf{z}^k - \mathbf{x}^k\|_2 \leq \tfrac{M_f}{2}\tfrac{\|\mathbf{z}^k - \mathbf{x}^k\|_{\mathbf{x}^\star}}{\sqrt{\underline{\sigma}^\star}} \overset{(A.19)}{\leq} \tfrac{c_\nu M_f}{2\sqrt{\underline{\sigma}^\star}}\|\mathbf{x}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star}.$$

On the other hand, we have $d_\nu(\mathbf{x}^\star, \mathbf{x}^k) = M_f\|\mathbf{x}^k - \mathbf{x}^\star\|_2 \le \frac{M_f}{\sqrt{\underline{\sigma}^\star}}\|\mathbf{x}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star}$. Using these estimates into (A.21), we get

$$\|\mathbf{x}^{k+1} - \mathbf{x}^\star\|_{\mathbf{x}^\star} \le \left(\frac{c_\nu M_f}{2\sqrt{\underline{\sigma}^\star}}\|\mathbf{x}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star} + \frac{c_\nu M_f}{\sqrt{\underline{\sigma}^\star}}\|\mathbf{x}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star}\right)\|\mathbf{x}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star} = \frac{3c_\nu M_f}{2\sqrt{\underline{\sigma}^\star}}\|\mathbf{x}^k - \mathbf{x}^\star\|^2_{\mathbf{x}^\star}.$$

Let $c^\star_\nu := \frac{3c_\nu M_f}{2\sqrt{\underline{\sigma}^\star}}$. The last estimate shows that if $\|\mathbf{x}^0 - \mathbf{x}^\star\|_{\mathbf{x}^\star} \le \min\{\frac{\bar{d}_\nu\sqrt{\underline{\sigma}^\star}}{M_f}, \frac{1}{c^\star_\nu}\}$, then $\{\|\mathbf{x}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star}\}$ quadratically converges to zero.

(b) If $2 < \nu \le 3$, then we first show that

$$d_\nu(\mathbf{x}^\star, \mathbf{x}^k) = M_f\|\mathbf{x}^k - \mathbf{x}^\star\|_2^{3-\nu}\|\mathbf{x}^k - \mathbf{x}^\star\|^{\nu-2}_{\mathbf{x}^\star} \le \frac{M_f}{(\underline{\sigma}^\star)^{\frac{3-\nu}{2}}}\|\mathbf{x}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star}. \tag{A.22}$$

Hence, if $\|\mathbf{x}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star} \le m_\nu\bar{d}_\nu$, where $m_\nu := \frac{(\underline{\sigma}^\star)^{\frac{3-\nu}{2}}}{M_f}$, then $d_\nu(\mathbf{x}^\star, \mathbf{x}^k) \le \bar{d}_\nu$. Next, using the definition of d_k in (4.4), we can bound it as

$$\begin{aligned} d_k &= M_f\|\mathbf{z}^k - \mathbf{x}^k\|^{\nu-2}_{\mathbf{x}^k}\|\mathbf{z}^k - \mathbf{x}^k\|_2^{3-\nu} \overset{(3.11)}{\le} M_f\left[\frac{\|\mathbf{z}^k - \mathbf{x}^k\|_{\mathbf{x}^\star}}{\big(1 - \frac{\nu-2}{2}d_\nu(\mathbf{x}^\star, \mathbf{x}^k)\big)^{\frac{1}{\nu-2}}}\right]^{\nu-2}\frac{\|\mathbf{z}^k - \mathbf{x}^k\|^{3-\nu}_{\mathbf{x}^\star}}{(\underline{\sigma}^\star)^{\frac{3-\nu}{2}}} \\ &\le \frac{M_f}{\big(1 - \frac{\nu-2}{2}\bar{d}_\nu\big)(\underline{\sigma}^\star)^{\frac{3-\nu}{2}}}\|\mathbf{z}^k - \mathbf{x}^k\|_{\mathbf{x}^\star} \overset{(A.19)}{\le} \frac{M_f}{\big(1 - \frac{\nu-2}{2}\bar{d}_\nu\big)(\underline{\sigma}^\star)^{\frac{3-\nu}{2}}}c_\nu\|\mathbf{x}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star}. \end{aligned}$$

Using this estimate, letting $\hat{d}_k := \frac{\nu-2}{2}d_k$, then we can bound $1 - \tau_k$ as follows:

$$\begin{aligned} 1 - \tau_k &= 1 - \tfrac{1}{\hat{d}_k} + \tfrac{1}{\hat{d}_k}\left(1 - \frac{\frac{4-\nu}{\nu-2}\hat{d}_k}{1 + \frac{4-\nu}{\nu-2}\hat{d}_k}\right)^{\frac{\nu-2}{4-\nu}} \overset{\text{Bernoulli's inequality}}{\le} 1 - \tfrac{1}{\hat{d}_k} + \tfrac{1}{\hat{d}_k}\left(1 - \frac{\nu-2}{4-\nu}\frac{\frac{4-\nu}{\nu-2}\hat{d}_k}{1 + \frac{4-\nu}{\nu-2}\hat{d}_k}\right) \\ &= \frac{\frac{4-\nu}{\nu-2}\hat{d}_k}{1 + \frac{4-\nu}{\nu-2}\hat{d}_k} \le \tfrac{4-\nu}{\nu-2}\hat{d}_k \le \frac{(4-\nu)M_f}{2\big(1 - \frac{\nu-2}{2}\bar{d}_\nu\big)(\underline{\sigma}^\star)^{\frac{3-\nu}{2}}}c_\nu\|\mathbf{x}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star} = n_\nu\|\mathbf{x}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star}, \end{aligned}$$

where $n_\nu := \frac{(4-\nu)M_f}{2\big(1 - \frac{\nu-2}{2}\bar{d}_\nu\big)(\underline{\sigma}^\star)^{\frac{3-\nu}{2}}}c_\nu > 0$. Substituting this estimate and (A.22) into (A.21), we get

$$\|\mathbf{x}^{k+1} - \mathbf{x}^\star\|_{\mathbf{x}^\star} \le \left(n_\nu + \frac{c_\nu}{m_\nu}\right)\|\mathbf{x}^k - \mathbf{x}^\star\|^2_{\mathbf{x}^\star} = c^*_\nu\|\mathbf{x}^k - \mathbf{x}^\star\|^2_{\mathbf{x}^\star}.$$

Hence, if $\|\mathbf{x}^0 - \mathbf{x}^\star\|_{\mathbf{x}^\star} \le \min\{m_\nu\bar{d}_\nu, \frac{1}{c^\star_\nu}\}$, then the last estimate shows that the sequence $\{\|\mathbf{x}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star}\}$ quadratically converges to zero.

In summary, there exists a neighborhood $\mathcal{N}(\mathbf{x}^\star)$ of $\mathbf{x}^\star$, such that if $\mathbf{x}^0 \in \mathcal{N}(\mathbf{x}^\star) \cap \mathrm{dom}(f)$, then the whole sequence $\{\|\mathbf{x}^k - \mathbf{x}^\star\|_{\mathbf{x}^\star}\}$ quadratically converges to zero. $\square$

A.2.6 The proof of Theorem 4.3.3: Quadratic convergence of the PN method

Since $\mathbf{z}^k$ is the optimal solution to (4.9), which satisfies (4.10), we have $\nabla^2 f(\mathbf{x}^k)\mathbf{x}^k - \nabla f(\mathbf{x}^k) \in (\nabla^2 f(\mathbf{x}^k) + \partial g)(\mathbf{z}^k)$. Using this optimality condition, we get

$$\begin{aligned} \mathbf{x}^{k+1} &= \mathbf{z}^k \qquad = \mathcal{P}^g_{\mathbf{x}^k}(S_{\mathbf{x}^k}(\mathbf{x}^k) + e_{\mathbf{x}^k}(\mathbf{x}^k, \mathbf{z}^k)) \qquad \text{and} \\ \mathbf{x}^{k+2} &= \mathbf{z}^{k+1} \quad = \mathcal{P}^g_{\mathbf{x}^k}(S_{\mathbf{x}^k}(\mathbf{x}^{k+1}) + e_{\mathbf{x}^k}(\mathbf{x}^{k+1}, \mathbf{z}^{k+1})). \end{aligned}$$

Let us define $\tilde{\lambda}_{k+1} := \|n^{k+1}_{\mathrm{pnt}}\|_{\mathbf{x}^k}$. Then, by Lemma (A.2.3) and the triangular inequality, we have

$$\begin{aligned} \tilde{\lambda}_{k+1} &\leq \|S_{\mathbf{x}^k}(\mathbf{x}^{k+1}) - S_{\mathbf{x}^k}(\mathbf{x}^k)\|^*_{\mathbf{x}^k} + \|e_{\mathbf{x}^k}(\mathbf{x}^{k+1}, \mathbf{z}^{k+1}) - e_{\mathbf{x}^k}(\mathbf{x}^k, \mathbf{z}^k)\|^*_{\mathbf{x}^k} \\ &= \|S_{\mathbf{x}^k}(\mathbf{x}^{k+1}) - S_{\mathbf{x}^k}(\mathbf{x}^k)\|^*_{\mathbf{x}^k} + \|e_{\mathbf{x}^k}(\mathbf{x}^{k+1}, \mathbf{z}^{k+1})\|^*_{\mathbf{x}^k}. \end{aligned} \tag{A.23}$$

Let us first bound the term $\|S_{\mathbf{x}^k}(\mathbf{x}^{k+1}) - S_{\mathbf{x}^k}(\mathbf{x}^k)\|^*_{\mathbf{x}^k}$ as follows:

$$\|S_{\mathbf{x}^k}(\mathbf{x}^{k+1}) - S_{\mathbf{x}^k}(\mathbf{x}^k)\|^*_{\mathbf{x}^k} \leq R_\nu(d_k) d_k \lambda_k, \tag{A.24}$$

where $R_\nu(t)$ is defined as (A.7). Indeed, from the mean-value theorem, we have

$$\|S_{\mathbf{x}^k}(\mathbf{x}^{k+1}) - S_{\mathbf{x}^k}(\mathbf{x}^k)\|^*_{\mathbf{x}^k} = \| \int_0^1 [\nabla^2 f(\mathbf{x}^k + t n^k_{\mathrm{pnt}}) - \nabla^2 f(\mathbf{x}^k)] n^k_{\mathrm{pnt}} \mathrm{d}t \|_{\mathbf{x}^k} \leq \|\mathbf{H}(\mathbf{x}^k, \mathbf{x}^{k+1})\| \lambda_k,$$

where $\mathbf{H}$ is defined as (A.5). Combining the above inequality and (A.7) in Lemma A.2.2, we get (A.24).

Next we bound the term $\|e_{\mathbf{x}^k}(\mathbf{x}^{k+1}, \mathbf{z}^{k+1})\|^*_{\mathbf{x}^k}$ as follows:

$$\|e_{\mathbf{x}^k}(\mathbf{x}^{k+1}, \mathbf{z}^{k+1})\|_{\mathbf{x}^k} \leq \begin{cases} \big[\big(1 - \frac{\nu-2}{2} d_k\big)^{\frac{-2}{\nu-2}} - 1 \big] \tilde{\lambda}_{k+1}, & \text{if } \nu > 2 \\ (e^{d_k} - 1) \tilde{\lambda}_{k+1} & \text{if } \nu = 2. \end{cases} \tag{A.25}$$

Note that

$$\|e_{\mathbf{x}^k}(\mathbf{x}^{k+1},\mathbf{z}^{k+1})\|^*_{\mathbf{x}^k} = \|[\nabla^2 f(\mathbf{x}^k) - \nabla^2 f(\mathbf{x}^{k+1})](\mathbf{z}^{k+1}-\mathbf{x}^{k+1})\|^*_{\mathbf{x}^k} \leq \|\widetilde{\mathbf{H}}(\mathbf{x}^k,\mathbf{x}^{k+1})\|\tilde{\lambda}_{k+1},$$

where

$$\begin{aligned}\widetilde{\mathbf{H}}(\mathbf{x},\mathbf{y}) &:= \nabla^2 f(\mathbf{x})^{-1/2}\left(\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\right)\nabla^2 f(\mathbf{x})^{-1/2}\\ &= \mathbb{I} - \nabla^2 f(\mathbf{x})^{-1/2}\nabla^2 f(\mathbf{y})\nabla^2 f(\mathbf{x})^{-1/2}.\end{aligned}$$

By Proposition 3.7.2, we have

$$\|\widetilde{\mathbf{H}}(\mathbf{x},\mathbf{y})\| \leq \begin{cases} \max\{1-\left(1-\frac{\nu-2}{2}d_\nu(\mathbf{x},\mathbf{y})\right)^{\frac{2}{\nu-2}}, \left(1-\frac{\nu-2}{2}d_\nu(\mathbf{x},\mathbf{y})\right)^{\frac{-2}{\nu-2}}-1\}, & \text{if } \nu>2\\ \max\{1-e^{-d_\nu(\mathbf{x},\mathbf{y})}, e^{d_\nu(\mathbf{x},\mathbf{y})}-1\} & \text{if } \nu=2.\end{cases}$$

This inequality can be simplified as

$$\|\widetilde{\mathbf{H}}(\mathbf{x},\mathbf{y})\| \leq \begin{cases} \left(1-\frac{\nu-2}{2}d_\nu(\mathbf{x},\mathbf{y})\right)^{\frac{-2}{\nu-2}}-1, & \text{if } \nu>2\\ e^{d_\nu(\mathbf{x},\mathbf{y})}-1 & \text{if } \nu=2.\end{cases} \tag{A.26}$$

Hence, the inequality (A.25) holds.

Now, we combine (A.23)(A.24), and (A.25), if $\nu=2$, and assuming that $d_k < \ln 2$, then we get

$$\tilde{\lambda}_{k+1} \leq \frac{R_2(d_k)d_k}{2-e^{d_k}}\lambda_k.$$

By Proposition 3.7.2, we have $\lambda^2_{k+1} \leq e^{d_k}\tilde{\lambda}^2_{k+1}$. Combining this estimate and the last inequality, we get

$$\lambda_{k+1} \leq \frac{R_2(d_k)d_k e^{\frac{d_k}{2}}}{2-e^{d_k}}\lambda_k. \tag{A.27}$$

Note that $\lambda_k \geq \frac{\sqrt{\underline{\sigma}_k}d_k}{M_f}$ and $\underline{\sigma}^{-1}_{k+1} \leq e^{d_k}\underline{\sigma}^{-1}_k$. It follows from (A.27) that

$$\frac{\lambda_{k+1}}{\sqrt{\underline{\sigma}_{k+1}}} \leq M_f\frac{R_2(d_k)e^{d_k}}{2-e^{d_k}}\left(\frac{\lambda_k}{\sqrt{\underline{\sigma}_k}}\right)^2.$$

By a numerical calculation, we can check that if $d_k \le d_2^\star \approx 0.35482$, then

$$\frac{\lambda_{k+1}}{\sqrt{\underline{\sigma}_{k+1}}} \le 2M_f \left(\frac{\lambda_k}{\sqrt{\underline{\sigma}_k}}\right)^2.$$

Hence, if we choose $\mathbf{x}^0 \in \mathrm{dom}(f)$ such that $\frac{\lambda_0}{\sqrt{\underline{\sigma}_0}} \le \frac{1}{M_f}\min\{d_2^\star, 0.5\} = \frac{d_2^\star}{M_f}$, then we can prove the following two inequalities together by induction:

$$d_{k+1} \le d_k \quad \text{and} \quad \frac{\lambda_{k+1}}{\sqrt{\underline{\sigma}_{k+1}}} \le \frac{\lambda_k}{\sqrt{\underline{\sigma}_k}}.$$

These inequalities show the nonincreasing monotonicity of $\{d_k\}$ and $\{\lambda_k\}$. The above inequality also shows the local quadratic convergence of the sequence $\{\frac{\lambda_k}{\sqrt{\underline{\sigma}_k}}\}$.

Now, if $\nu > 2$ and assume that $d_k < \frac{\nu-2}{2}\left(1 - \left(\frac{1}{2}\right)^{\frac{\nu-2}{2}}\right)$, then

$$\tilde{\lambda}_{k+1} \le \frac{R_\nu(d_k)d_k}{2 - \left(1 - \frac{\nu-2}{2}d_k\right)^{\frac{-2}{\nu-2}}}\lambda_k.$$

By Proposition 3.7.2, we have $\lambda_{k+1}^2 \le \left(1 - \frac{\nu-2}{2}d_k\right)^{\frac{-2}{\nu-2}} \tilde{\lambda}_{k+1}^2$. Hence, combining these inequalities, we get

$$\lambda_{k+1} \le \frac{R_\nu(d_k)d_k\left(1 - \frac{\nu-2}{2}d_k\right)^{\frac{-1}{\nu-2}}}{2 - \left(1 - \frac{\nu-2}{2}d_k\right)^{\frac{-2}{\nu-2}}}\lambda_k. \tag{A.28}$$

Note that $d_k = M_f\beta_k^{3-\nu}\lambda_k^{\nu-2}$, $\underline{\sigma}_{k+1}^{-1} \le \left(1 - \frac{\nu-2}{2}d_k\right)^{\frac{-2}{\nu-2}}\underline{\sigma}_k^{-1}$ and $\sigma_{k+1}^{-1} \le \left(1 - \frac{\nu-2}{2}d_k\right)^{\frac{-2}{\nu-2}}\sigma_k^{-1}$. Using these relations and (A.28), we consider two cases:

(a) If $\nu = 3$, then $d_k = M_f\lambda_k$, and

$$\lambda_{k+1} \le \frac{R_3(d_k)(1-0.5d_k)^{-1}}{2-(1-0.5d_k)^{-2}}d_k\lambda_k = M_f\frac{R_3(d_k)(1-0.5d_k)^{-1}}{2-(1-0.5d_k)^{-2}}\lambda_k^2.$$

By a simple numerical calculation, we can show that if $d_k \le d_3^\star \approx 0.41886$, then $\lambda_{k+1} \le 2M_f\lambda_k^2$. Hence, if $\lambda_0 < \frac{1}{M_f}\min\{d_3^\star, 0.5\} = \frac{d_3^\star}{M_f}$, then we can prove the following two inequalities together by induction

$$d_{k+1} \le d_k \text{ and } \lambda_{k+1} \le \lambda_k.$$

These inequalities show the non-increasing monotonicity of $\{d_k\}$ and $\{\lambda_k\}$. The above inequality also shows the quadratic convergence of the sequence $\{\lambda_k\}$.

(b) If $2 < \nu < 3$, then $\lambda_k \geq \beta_k \sqrt{\underline{\sigma}_k}$, which implies that $d_k \leq M_f \underline{\sigma}_k^{-\frac{3-\nu}{2}} \lambda_k$. Hence, we have

$$\frac{\lambda_{k+1}}{\underline{\sigma}_{k+1}^{\frac{3-\nu}{2}}} \leq \frac{R_\nu(d_k)\left(1 - \frac{\nu-2}{2} d_k\right)^{-\frac{4-\nu}{\nu-2}}}{2 - \left(1 - \frac{\nu-2}{2} d_k\right)^{\frac{-2}{\nu-2}}} M_f \left(\frac{\lambda_k}{\underline{\sigma}_k^{\frac{3-\nu}{2}}}\right)^2.$$

If $d_k < d_\nu^\star$, then $\underline{\sigma}_{k+1}^{-\frac{3-\nu}{2}} \lambda_{k+1} \leq 2 M_f \left(\underline{\sigma}_k^{-\frac{3-\nu}{2}} \lambda_k\right)^2$, where $d_\nu^\star$ is the unique solution to the equation

$$\frac{R_\nu(d_k)\left(1 - \frac{\nu-2}{2} d_k\right)^{-\frac{4-\nu}{\nu-2}}}{2 - \left(1 - \frac{\nu-2}{2} d_k\right)^{\frac{-2}{\nu-2}}} = 2.$$

Note that it is straightforward to check that this equation always admits a positive solution. Therefore, if $\underline{\sigma}_0^{-\frac{3-\nu}{2}} \lambda_0 \leq \frac{1}{M_f} \min\{d_\nu^\star, 0.5\}$, then we can prove the following two inequalities together by induction:

$$d_k \leq d_{k+1} \text{ and } \underline{\sigma}_{k+1}^{-\frac{3-\nu}{2}} \lambda_{k+1} \leq \underline{\sigma}_k^{-\frac{3-\nu}{2}} \lambda_k.$$

These inequalities show the non-increasing monotonicity of $\{d_k\}$ and $\{\lambda_k\}$. The above inequality also shows the quadratic convergence of the sequence $\left\{\frac{\lambda_k}{\underline{\sigma}_k^{\frac{3-\nu}{2}}}\right\}$.

Finally, to prove the local quadratic convergence of $\{\mathbf{x}^k\}$ to $\mathbf{x}^\star$, we use the same argument as in the proof of Theorem 4.2.3 and Theorem 4.3.2, where we omit the details here.

A.3 Technical proofs of results in Chapter 5

A.3.1 The proof of Lemma 5.2.1: Properties of global inexact oracle

(a) Substituting $\mathbf{x} = \mathbf{y}$ into (5.2), we obtain (5.4) for all $\mathbf{x} \in \mathrm{dom}(f)$.

(b) Clearly, if $\langle g(\bar{\mathbf{x}}), \mathbf{y} - \bar{\mathbf{x}} \rangle \geq 0$ for all $\mathbf{y} \in \mathrm{dom}(f)$, then $\langle g(\bar{\mathbf{x}}), \mathbf{x}^\star - \bar{\mathbf{x}} \rangle \geq 0$ for a minimizer $\mathbf{x}^\star$ of f. Using this relation into (5.2), we have

$$f^\star = f(\mathbf{x}^\star) \geq \tilde{f}(\bar{\mathbf{x}}) + \omega((1-\delta_0)\|\mathbf{x}^\star - \bar{\mathbf{x}}\|_{\bar{\mathbf{x}}}) \geq \tilde{f}(\bar{\mathbf{x}}) \overset{(5.4)}{\geq} f(\bar{\mathbf{x}}) - \delta_1.$$

This implies $f^\star \leq f(\bar{\mathbf{x}}) \leq f^\star + \delta_1$.

(c) Let $\nabla f(\mathbf{x})$ be a (sub)gradient of f at $\mathbf{x}$. For $\mathbf{y} \in \mathrm{dom}(f)$, it follows from (5.2) and (5.4) that

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq \tilde{f}(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle .$$

Subtracting this estimate from the second inequality of (5.2), we have

$$\langle \nabla f(\mathbf{x}) - g(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \omega_* \left((1 + \delta_0) \|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}} \right) + \delta_1, \tag{A.29}$$

provided that $\|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}} < \frac{1}{1+\delta_0}$. Let us consider an arbitrary $z \in \mathbb{R}^p$ such that

$$\|\nabla f(\mathbf{x}) - g(\mathbf{x})\|_{\mathbf{x}}^* = | \langle \nabla f(\mathbf{x}) - g(\mathbf{x}), \mathbf{z} \rangle | \text{ and } \|\mathbf{z}\|_{\mathbf{x}} = 1.$$

Then, by choosing $\mathbf{y} \in \mathrm{dom}(f)$ such that $\mathbf{y} = y_\tau(\mathbf{x}) := \mathbf{x} + \tau \mathrm{sign}(\langle \nabla f(\mathbf{x}) - g(\mathbf{x}), \mathbf{z} \rangle)\mathbf{z}$ for some $\tau > 0$, (A.29) becomes

$$\tau \|\nabla f(\mathbf{x}) - g(\mathbf{x})\|_{\mathbf{x}}^* \leq \omega_* \left((1 + \delta_0)\tau \right) + \delta_1,$$

which is equivalent to

$$\|\nabla f(\mathbf{x}) - g(\mathbf{x})\|_{\mathbf{x}}^* \leq s(\tau; \delta_0, \delta_1) := \tfrac{\omega_*((1+\delta_0)\tau) + \delta_1}{\tau}. \tag{A.30}$$

Let us take $\tau := \frac{\bar{c}}{(1+\delta_0+\bar{c})(1+\delta_0)}$ for some $\bar{c} > 0$. Then, we can check that $\|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}} = \tau < \frac{1}{1+\delta_0}$. In this case, the right-hand side of (A.30) becomes

$$s(\bar{c}; \delta_0, \delta_1) = \tfrac{(1+\delta_0)(1+\delta_0+\bar{c})}{\bar{c}} \left[\delta_1 + \ln\left(1 + \tfrac{\bar{c}}{1+\delta_0}\right) \right] - (1 + \delta_0), \tag{A.31}$$

for any $\bar{c} > 0$. Minimizing (A.31) over $\bar{c}$, we can show that the minimum is attained at $\bar{c} := \bar{c}(\delta_0, \delta_1) > 0$ which is the unique solution of $\omega\big(\frac{\bar{c}}{1+\delta_0}\big) = \delta_1$ in $\bar{c}$. Substituting $\bar{c} = \bar{c}(\delta_0, \delta_1)$ in $s(\bar{c}; \delta_0, \delta_1)$, we can see that the minimum value of (A.31) is $\bar{c}(\delta_0, \delta_1)$.

(d) Let us consider the function $\varphi(\mathbf{y}) := f(\mathbf{y}) - \big\langle \nabla f(\mathbf{x}^0), \mathbf{y} \big\rangle$ for some $\mathbf{x}^0 \in \mathrm{dom}(f)$. It is clear that $\nabla \varphi(\mathbf{x}^0) = 0$, which shows that $\mathbf{x}^0$ is a minimizer of φ. Hence, we have $\varphi(\mathbf{x}^0) \leq \varphi(\mathbf{x} - tH(\mathbf{x})^{-1}h(\mathbf{x}))$ for some $t > 0$ such that $\mathbf{x} - tH(\mathbf{x})^{-1}h(\mathbf{x}) \in \mathrm{dom}(f)$. If we define

$\tilde{\varphi}(\mathbf{x}) := \tilde{f}(\mathbf{x}) - \langle \nabla f(\mathbf{x}^0), \mathbf{x} \rangle$, and $h(\mathbf{x}) := g(\mathbf{x}) - \nabla f(\mathbf{x}^0)$, then, by using (5.2), we can further derive

$$\varphi(\mathbf{x}^0) \leq \varphi(\mathbf{x} - tH(\mathbf{x})^{-1}h(\mathbf{x})) \leq \tilde{\varphi}(\mathbf{x}) - t(\|h(\mathbf{x})\|_{\mathbf{x}}^*)^2 + \omega_* \left((1+\delta_0)t\|h(\mathbf{x})\|_{\mathbf{x}}^*\right) + \delta_1.$$

Minimizing the right-hand side w.r.t $t > 0$ and note that $\mathrm{dom}(f)$ is open, we obtain

$$\varphi(\mathbf{x}^0) \leq \tilde{\varphi}(\mathbf{x}) - \omega\left(\frac{\|h(\mathbf{x})\|_{\mathbf{x}}^*}{1+\delta_0}\right) + \delta_1,$$

given $t = \frac{1}{(1+\delta_0)(1+\delta_0+\|h(\mathbf{x})\|_{\mathbf{x}}^*)}$. Using the definition of φ, we can show that

$$\begin{aligned}
\omega\left(\tfrac{\|h(\mathbf{x})\|_{\mathbf{x}}^*}{1+\delta_0}\right) &\leq \tilde{f}(\mathbf{x}) - f(\mathbf{x}^0) - \langle \nabla f(\mathbf{x}^0), \mathbf{x} - \mathbf{x}^0 \rangle + \delta_1 \\
&\overset{(5.4)}{\leq} -\omega((1-\delta_0)\|\mathbf{x} - \mathbf{x}^0\|_{\mathbf{x}}) + \langle g(\mathbf{x}) - \nabla f(\mathbf{x}^0), \mathbf{x} - \mathbf{x}^0 \rangle + \delta_1 \\
&\leq \|h(\mathbf{x})\|_{\mathbf{x}}^* \|\mathbf{x} - \mathbf{x}^0\|_{\mathbf{x}} + \delta_1,
\end{aligned}$$

where the last inequality is by the Cauchy-Schwarz inequality and $\omega(\cdot) \geq 0$. By letting $\mathbf{x}^0 = \mathbf{y}$ into this inequality, we obtain

$$\omega\left(\tfrac{\|g(\mathbf{x}) - \nabla f(\mathbf{y})\|_{\mathbf{x}}^*}{1+\delta_0}\right) \leq \|g(\mathbf{x}) - \nabla f(\mathbf{y})\|_{\mathbf{x}}^* \|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}} + \delta_1,$$

which is exactly (5.6). □

A.3.2 The proof of Lemma 5.2.2: Properties of local inexact oracle

The estimates in (5.7) are direct consequences of (5.3). We prove (5.8). From [74, Theorem 4.1.6], for all $\mathbf{x} \in \mathrm{dom}(f)$ and $\mathbf{y}$ satisfying $\|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}} < 1$, we have

$$(1 - \|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}})^2 \nabla^2 f(\mathbf{x}) \preceq \nabla^2 f(\mathbf{y}) \preceq \tfrac{1}{(1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^2} \nabla^2 f(\mathbf{x}),$$

provided that $\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < 1$. Moreover, by using the second inequality of (5.3), we can easily show that, for any $\mathbf{x} \in \mathcal{X}$, one has

$$(1+\delta_3)^{-1}\|\!|\mathbf{y}-\mathbf{x}|\!\|_{\mathbf{x}} \leq \|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} \leq (1-\delta_3)^{-1}\|\!|\mathbf{y}-\mathbf{x}|\!\|_{\mathbf{x}}, \tag{A.32}$$

for all $\mathbf{y} \in \mathrm{dom}(f)$. Combining these two inequalities we can further derive

$$\begin{aligned} H(\mathbf{y}) &\succeq (1-\delta_3)\nabla^2 f(\mathbf{y}) \succeq (1-\delta_3)\left(1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right)^2 \nabla^2 f(\mathbf{x}) \\ &\succeq \frac{1-\delta_3}{1+\delta_3}\left(1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right)^2 H(\mathbf{x}) \succeq \tfrac{(1-\delta_3-\|\!|\mathbf{y}-\mathbf{x}|\!\|_{\mathbf{x}})^2}{1-\delta_3^2} H(\mathbf{x}), \end{aligned}$$

and

$$\begin{aligned} H(\mathbf{y}) &\preceq (1+\delta_3)\nabla^2 f(\mathbf{y}) \preceq \tfrac{1+\delta_3}{(1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^2}\nabla^2 f(\mathbf{x}) \\ &\preceq \tfrac{1+\delta_3}{1-\delta_3}\tfrac{1}{(1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^2} H(\mathbf{x}) \preceq \tfrac{1+\delta_3}{1-\delta_3}\tfrac{1}{(1-(1-\delta_3)^{-1}\|\!|\mathbf{y}-\mathbf{x}|\!\|_{\mathbf{x}})^2} H(\mathbf{x}), \end{aligned}$$

which is the first estimate of (5.8).

To prove the last relation, let $G_{\mathbf{x}} = [\nabla^2 f(\mathbf{x})]^{-1/2}(\nabla^2 f(\mathbf{x}) - H(\mathbf{x}))[\nabla^2 f(\mathbf{x})]^{-1/2}$. Then, from local oracle definition we have $\|G_{\mathbf{x}}\| \leq \delta_3$. Moreover, one can show that

$$\begin{aligned} \|\!|(\nabla^2 f(\mathbf{x}) - H(\mathbf{x}))\mathbf{v}|\!\|_{\mathbf{y}}^* &\leq \tfrac{1}{1-\delta_3}\|(\nabla^2 f(\mathbf{x}) - H(\mathbf{x}))\mathbf{v}\|_{\mathbf{y}}^* \\ &\leq \tfrac{1}{1-\delta_3}\left(\mathbf{v}^\top(\nabla^2 f(\mathbf{x}) - H(\mathbf{x}))\tfrac{1}{(1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^2}[\nabla^2 f(\mathbf{x})]^{-1}(\nabla^2 f(\mathbf{x}) - H(\mathbf{x}))\mathbf{v}\right)^{1/2} \\ &\leq \tfrac{1}{(1-\delta_3)(1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})}\|G_{\mathbf{x}}[\nabla^2 f(\mathbf{x})]^{1/2}\mathbf{v}\| \leq \tfrac{1}{(1-\delta_3)(1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})}\|G_{\mathbf{x}}\|\|\mathbf{v}\|_{\mathbf{x}} \\ &\leq \tfrac{\delta_3}{(1-\delta_3)^2((1-(1-\delta_3)^{-1}\|\!|\mathbf{y}-\mathbf{x}|\!\|_{\mathbf{x}}))}\|\!|\mathbf{v}|\!\|_{\mathbf{x}} \\ &= \tfrac{\delta_3}{(1-\delta_3)(1-\delta_3-\|\!|\mathbf{y}-\mathbf{x}|\!\|_{\mathbf{x}})}\|\!|\mathbf{v}|\!\|_{\mathbf{x}}. \end{aligned}$$

This is exactly the second estimate of (5.8). □

A.3.3 The proof of Lemma 5.3.1: Computational inexact oracle

For any $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$ and $\alpha \in (0,1)$, we have:

$$\begin{aligned} f(\mathbf{y}) &\geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \omega(\|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}}) \\ &\geq \hat{f}(\mathbf{x}) + \langle g(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle - \varepsilon + \langle \nabla f(\mathbf{x}) - g(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \omega(\|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}}) \\ &\geq \hat{f}(\mathbf{x}) + \langle g(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle - \varepsilon - \|\nabla f(\mathbf{x}) - g(\mathbf{x})\|_{\mathbf{x}}^* \|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}} + \omega((1-\delta_3)\|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}}) \\ &\geq \hat{f}(\mathbf{x}) + \langle g(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \omega(\alpha(1-\delta_3)\|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}}) \\ &\quad - \varepsilon - \delta_2 \|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}} + \omega((1-\delta_3)\|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}}) - \omega(\alpha(1-\delta_3)\|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}}), \end{aligned} \tag{A.33}$$

where the first inequality is from [74, Theorem 4.1.7], the second and the last are from oracle setting, and the third is from Cauchy-Schwarz inequality. Now we consider the function $\psi(t) := -\delta_2 t + \omega(\gamma t) - \omega(\alpha\gamma t)$ where $\gamma := 1 - \delta_3$. Clearly, we can write $\psi(t) = \gamma t - \ln(1+\gamma t) - \delta_2 t - \alpha\gamma t + \ln(1+\alpha\gamma t)$. We have $\psi'(t) = (1-\alpha)\gamma - \delta_2 - \frac{\gamma}{1+\gamma t} + \frac{\alpha\gamma}{1+\alpha\gamma t}$, and $\psi''(t) = \frac{\gamma^2}{(1+\gamma t)^2} - \frac{(\alpha\gamma)^2}{(1+\alpha\gamma t)^2} \geq 0$. Hence, it is convex. It attains the minimum at $t^* > 0$ as a solution of $(1-\alpha)\gamma - \delta_2 - \frac{\gamma}{1+\gamma t} + \frac{\alpha\gamma}{1+\alpha\gamma t} = 0$. Solving this equation, we get

$$t^* = \frac{1}{2\alpha\gamma}\left(\sqrt{(1+\alpha)^2 + \frac{4\alpha\delta_2}{(1-\alpha)\gamma - \delta_2}} - (1+\alpha)\right) > 0,$$

is the minimum point, provided that $(1-\alpha)\gamma > \delta_2$. Substituting this into (A.33), we obtain

$$f(\mathbf{y}) \geq \tilde{f}(\mathbf{x}) + \langle g(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \omega\left(\alpha(1-\delta_3)\|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}}\right),$$

where $\tilde{f}(\mathbf{x}) := \hat{f}(\mathbf{x}) - \varepsilon + \psi(t^*)$. It remains to compute $\psi(t^*)$. For this $t = t^*$, using first-order optimal condition we get

$$\begin{aligned} \psi(t^*) &= \gamma t^* - \ln(1+\gamma t^*) - \delta_2 t^* - \alpha\gamma t^* + \ln(1+\alpha\gamma t^*) \\ &= \frac{1}{1+\alpha\gamma t^*}\frac{(1-\alpha)\gamma t^*}{1+\gamma t^*} + \ln\left(1 - \frac{(1-\alpha)\gamma t^*}{1+\gamma t^*}\right). \end{aligned}$$

Substituting the expression of t^*, we have

$$\begin{aligned}\frac{(1-\alpha)\gamma t^*}{1+\gamma t^*} &= \frac{(1-\alpha)[(1+\alpha)(1-\delta_3)+\delta_2]}{2(1-\delta_3)} \\ &- \frac{(1-\alpha)(1-\delta_3)-\delta_2}{2(1-\delta_3)}\sqrt{(1+\alpha)^2+\frac{4\alpha\delta_2}{(1-\alpha)(1-\delta_3)-\delta_2}},\end{aligned}$$

and

$$\begin{aligned}\frac{1}{1+\alpha\gamma t^*} &= \frac{(1-\alpha)(1-\delta_3)-\delta_2}{2(1-\alpha)\alpha(1-\delta_3)}\sqrt{(1+\alpha)^2+\frac{4\alpha\delta_2}{(1-\alpha)(1-\delta_3)-\delta_2}} \\ &- \frac{(1-\alpha)(1-\delta_3)-\delta_2}{2\alpha(1-\delta_3)}.\end{aligned}$$

By computing $\psi(t^*)$ directly with $\alpha = 1-\frac{2\delta_2}{1-\delta_3} > 0$, we obtain the first inequality of (5.13). To prove the second inequality of (5.13), we also have

$$\begin{aligned}f(\mathbf{y}) &\leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \omega_*(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}) \\ &\leq \hat{f}(\mathbf{x}) + \langle g(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \omega_*(\beta(1+\delta_3)\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}) \\ &+ \varepsilon + \|g(\mathbf{x})-\nabla f(\mathbf{x})\|_{\mathbf{x}}^*\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} + \omega_*((1+\delta_3)\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}) \\ &- \omega_*(\beta(1+\delta_3)\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}),\end{aligned} \tag{A.34}$$

where the first inequality is from [74, Theorem 4.1.8], while the second is from oracle setting and Cauchy-Schwarz inequality. Let us consider the function

$$\begin{aligned}\bar{\psi}(t) &= \delta_2 t + \omega_*(\bar{\gamma}t) - \omega_*(\beta\bar{\gamma}t) \\ &= \delta_2 t - \bar{\gamma}t - \ln(1-\bar{\gamma}t) + \beta\gamma t + \ln(1-\beta\bar{\gamma}t) \\ &= (\beta-1)\bar{\gamma}t + \delta_2 t + \ln(1-\beta\bar{\gamma}t) - \ln(1-\bar{\gamma}t),\end{aligned}$$

where $\bar{\gamma} = 1+\delta_3 \geq 1$. We have $\bar{\psi}'(t) = (\beta-1)\bar{\gamma} + \delta_2 - \frac{\beta\bar{\gamma}}{1-\beta\bar{\gamma}t} + \frac{\bar{\gamma}}{1-\bar{\gamma}t}$ and $\bar{\psi}''(t) = -\frac{(\beta\bar{\gamma})^2}{(1-\beta\bar{\gamma}t)^2} + \frac{\bar{\gamma}^2}{(1-\bar{\gamma}t)^2} \leq 0$ for $\beta \geq 1$. Letting $\psi'(t) = 0$ we get

$$\bar{t}^* = \frac{1}{2\beta\bar{\gamma}}\left(1+\beta-\sqrt{(1+\beta)^2-\frac{4\beta\delta_2}{(\beta-1)\bar{\gamma}+\delta_2}}\right) > 0$$

is the maximum point, provided that $\delta_2 > 0$ For this $t = \bar{t}^*$, using first-order optimal condition we get

$$\begin{aligned}\bar{\psi}(\bar{t}^*) &= (\beta-1)\bar{\gamma}\bar{t}^* + \delta_2\bar{t}^* + \ln(1-\beta\bar{\gamma}\bar{t}^*) - \ln(1-\bar{\gamma}\bar{t}^*) \\ &= \frac{1}{1-\beta\bar{\gamma}\bar{t}^*}\frac{(\beta-1)\bar{\gamma}\bar{t}^*}{1-\bar{\gamma}\bar{t}^*} + \ln\left(1-\frac{(\beta-1)\bar{\gamma}t}{1-\bar{\gamma}\bar{t}^*}\right).\end{aligned}$$

Substituting $\bar{t}^*$, we get

$$\begin{aligned}\frac{(\beta-1)\bar{\gamma}\bar{t}^*}{1-\bar{\gamma}\bar{t}^*} &= \frac{(\beta-1)(1+\delta_3)+\delta_2}{2(1+\delta_3)}\sqrt{(1+\beta)^2 - \frac{4\beta\delta_2}{(\beta-1)(1+\delta_3)+\delta_2}} \\ &- \frac{(\beta-1)[(\beta+1)(1+\delta_3)+\delta_2]}{2(1+\delta_3)},\end{aligned}$$

and

$$\begin{aligned}\frac{1}{1-\beta\bar{\gamma}\bar{t}^*} &= \frac{(\beta-1)(1+\delta_3)+\delta_2}{2\beta(1+\delta_3)} \\ &- \frac{(\beta-1)(1+\delta_3)+\delta_2}{2(\beta-1)\beta(1+\delta_3)}\sqrt{(1+\beta)^2 - \frac{4\beta\delta_2}{(\beta-1)(1+\delta_3)+\delta_2}}.\end{aligned}$$

Substituting above formulations back to (A.34), and using the increasing property of ω and ω_*, we obtain the second inequality in (5.13) by letting $\delta_1 := 2\varepsilon - \psi(t^*) + \bar{\psi}(\bar{t}^*) \geq 0$ and $\delta_0 := \max\{1-(1-\delta_3)\alpha, (1+\delta_3)\beta - 1\}$. Finally the lemma is proven by taking $\alpha = 1 - \frac{2\delta_2}{1-\delta_3} > 0$ and β as shown in equation (5.14). □

A.3.4 The proof of Lemma 5.3.2: Inexact oracle of dual problem

Since φ is self-concordant, by [74, Theorem 4.1.6] we have

$$(1-\delta(\mathbf{x}))^2[\nabla^2\varphi(u^*(\mathbf{x}))]^{-1} \preceq [\nabla^2\varphi(\tilde{u}^*(\mathbf{x}))]^{-1} \preceq (1-\delta(\mathbf{x}))^{-2}[\nabla^2\varphi(u^*(\mathbf{x}))]^{-1}.$$

Multiplying this inequality by $\mathbf{A}$ and $\mathbf{A}^\top$ on the left and right of each item, we obtain

$$(1-\delta(\mathbf{x}))^2\nabla^2 f(\mathbf{x}) \preceq H(\mathbf{x}) \preceq (1-\delta(\mathbf{x}))^{-2}\nabla^2 f(\mathbf{x}). \tag{A.35}$$

Since $\delta(\mathbf{x}) \leq \delta$, then imposing that $(1-\delta)^2 \geq (1-\delta_3)^2$ and $(1-\delta)^{-2} \leq (1+\delta_3)^2$, we get $\delta_3 \geq \delta/(1-\delta)$, which proves the second bound of (5.18).

Next, by definition of $g(\mathbf{x})$ and $\nabla f(\mathbf{x})$, we can derive that

$$\begin{aligned}
[\|\!|g(\mathbf{x}) - \nabla f(\mathbf{x})|\!\|_{\mathbf{x}}^*]^2 &= (\tilde{u}^*(\mathbf{x}) - u^*(\mathbf{x}))^\top \mathbf{A}^\top \left(\mathbf{A}\nabla^2\varphi(\tilde{u}^*(\mathbf{x}))^{-1}\mathbf{A}^\top\right)^{-1}\mathbf{A}(\tilde{u}^*(\mathbf{x}) - u^*(\mathbf{x})) \\
&\leq (\tilde{u}^*(\mathbf{x}) - u^*(\mathbf{x}))^\top \nabla^2\varphi(\tilde{u}^*(\mathbf{x}))(\tilde{u}^*(\mathbf{x}) - u^*(\mathbf{x})) \\
&= \|\tilde{u}^*(\mathbf{x}) - u^*(\mathbf{x})\|^2_{\tilde{u}^*(\mathbf{x})} \leq \delta^2(\mathbf{x}) \leq \delta^2,
\end{aligned}$$

which implies the first estimate of (5.18). Here, the inequality in this chain follows from the fact that $\mathbf{A}^\top(\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^\top)^{-1}\mathbf{A} \preceq \mathbf{Q}$ for a symmetric positive definite matrix $\mathbf{Q} = \nabla^2\varphi(u^*(\mathbf{x}))$ (see [31] for a detailed proof of this fact).

Finally, by definition of $f(\cdot)$ and $\tilde{f}(\cdot)$, we can derive

$$\begin{aligned}
f(\mathbf{x}) - \tilde{f}(\mathbf{x}) &= \left[\left\langle u^*(\mathbf{x}), \mathbf{A}^\top\mathbf{x}\right\rangle - \varphi(u^*(\mathbf{x}))\right] - \left[\left\langle \tilde{u}^*(\mathbf{x}), \mathbf{A}^\top\mathbf{x}\right\rangle - \varphi(\tilde{u}^*(\mathbf{x}))\right] \\
&= \varphi(\tilde{u}^*(\mathbf{x})) - \varphi(u^*(\mathbf{x})) - \left\langle \mathbf{A}^\top\mathbf{x}, \tilde{u}^*(\mathbf{x}) - u^*(\mathbf{x})\right\rangle \\
&= \varphi(\tilde{u}^*(\mathbf{x})) - \varphi(u^*(\mathbf{x})) - \left\langle \nabla\varphi(u^*(\mathbf{x})), \tilde{u}^*(\mathbf{x}) - u^*(\mathbf{x})\right\rangle,
\end{aligned}$$

where the last equality follows from the optimality condition $\nabla\varphi(u^*(\mathbf{x})) = \mathbf{A}^\top\mathbf{x}$ for $u^*(\mathbf{x})$. Since φ is self-concordant, using [74, Theorem 4.1.7, 4.1.8] we get

$$\omega(\|\tilde{u}^*(\mathbf{x}) - u^*(\mathbf{x})\|_{u^*(\mathbf{x})}) \leq f(\mathbf{x}) - \tilde{f}(\mathbf{x}) \leq \omega_*(\|\tilde{u}^*(\mathbf{x}) - u^*(\mathbf{x})\|_{u^*(\mathbf{x})}),$$

which leads to

$$0 \leq \omega\left(\frac{\delta(\mathbf{x})}{1+\delta(\mathbf{x})}\right) \leq f(\mathbf{x}) - \tilde{f}(\mathbf{x}) \leq \omega_*\left(\frac{\delta(\mathbf{x})}{1-\delta(\mathbf{x})}\right) \leq \omega_*\left(\frac{\delta}{1-\delta}\right), \tag{A.36}$$

given $\delta(\mathbf{x}) < 1$. The proof is completed using Lemma 5.3.1 by letting $\varepsilon := \omega_*\left(\frac{\delta}{1-\delta}\right)$ and δ_2, δ_3 defined above. Since $2\delta_2 + \delta_3 < 1$ is required in Lemma 5.3.1, we have $\delta \in [0, 0.292]$.

From the optimality condition of (5.16) we have $\nabla\varphi(u^*(\mathbf{x})) - \mathbf{A}^\top\mathbf{x} = 0$. Let $r(\mathbf{x}) := \nabla\varphi(\tilde{u}^*(\mathbf{x})) - \mathbf{A}^\top\mathbf{x}$. Then, using self-concordance of φ, we have

$$\frac{\|\tilde{u}^*(\mathbf{x}) - u^*(\mathbf{x})\|^2_{u^*(\mathbf{x})}}{1 + \|\tilde{u}^*(\mathbf{x}) - u^*(\mathbf{x})\|_{u^*(\mathbf{x})}} \leq \langle \nabla\varphi(\tilde{u}^*(\mathbf{x})) - \nabla\varphi(u^*(\mathbf{x})), \tilde{u}^*(\mathbf{x}) - u^*(\mathbf{x})\rangle = \langle r(\mathbf{x}), \tilde{u}^*(\mathbf{x}) - u^*(\mathbf{x})\rangle\,.$$

Since $\delta(\mathbf{x}) := \|\tilde{u}^*(\mathbf{x}) - u^*(\mathbf{x})\|_{\tilde{u}^*(\mathbf{x})}$, by the Cauchy-Schwarz inequality, we can show that $\frac{\delta(\mathbf{x})^2}{1+\delta(\mathbf{x})} \leq \|r(\mathbf{x})\|^*_{\tilde{u}^*(\mathbf{x})}\delta(\mathbf{x})$. Therefore, we obtain the last statement of Lemma 5.3.2 provided that $\delta \in (0, 1)$. □

A.3.5 The proof of Lemma 5.4.2: Key estimate for local convergence

From the definition of ν^k in (5.20) we have

$$H(\mathbf{x}^k)\mathbf{x}^k + \nu^k - g(\mathbf{x}^k) \in \partial R(\bar{\mathbf{z}}^k) + H(\mathbf{x}^k)\bar{\mathbf{z}}^k, \quad \text{or} \quad \bar{\mathbf{z}}^k \in \mathcal{P}_{\mathbf{x}^k}(\mathbf{x}^k + [H(\mathbf{x}^k)]^{-1}(\nu^k - g(\mathbf{x}^k))). \tag{A.37}$$

On the other hand, if we denote $r_{\mathbf{x}^k}(\bar{\mathbf{z}}^k) := g(\mathbf{x}^k) + H(\mathbf{x}^k)(\bar{\mathbf{z}}^k - \mathbf{x}^k)$, then

$$\begin{aligned}
& \nu^k - r_{\mathbf{x}^k}(\bar{\mathbf{z}}^k) && \in \partial R(\bar{\mathbf{z}}^k) \\
\iff \quad & \bar{\mathbf{z}}^k + [H(\mathbf{x}^{k+1})]^{-1}(\nu^k - r_{\mathbf{x}^k}(\bar{\mathbf{z}}^k)) && \in \bar{\mathbf{z}}^k + [H(\mathbf{x}^{k+1})]^{-1}\partial R(\bar{\mathbf{z}}^k) \\
\iff \quad & \bar{\mathbf{z}}^k && = \mathcal{P}_{\mathbf{x}^{k+1}}(\bar{\mathbf{z}}^k + [H(\mathbf{x}^{k+1})]^{-1}(\nu^k - r_{\mathbf{x}^k}(\bar{\mathbf{z}}^k))).
\end{aligned} \tag{A.38}$$

For simplicity of notation, we define $H_k := H(\mathbf{x}^k)$, $f'_k := \nabla f(\mathbf{x}^k)$ and $g_k := g(\mathbf{x}^k)$. By the triangle inequality, it is obvious that

$$\lambda_{k+1} = \|\mathbf{x}^{k+1} - \bar{\mathbf{z}}^{k+1}\|_{\mathbf{x}^{k+1}} \leq \|\bar{\mathbf{z}}^{k+1} - \bar{\mathbf{z}}^k\|_{\mathbf{x}^{k+1}} + \|\mathbf{x}^{k+1} - \bar{\mathbf{z}}^k\|_{\mathbf{x}^{k+1}}. \tag{A.39}$$

For the second item, by (5.7) and iPNA, we have

$$\begin{aligned}\|\mathbf{x}^{k+1}-\mathbf{z}^k\|_{\mathbf{x}^{k+1}} &\le (1+\delta_3^{k+1})\|\mathbf{x}^{k+1}-\mathbf{z}^k\|_{\mathbf{x}^{k+1}} \le \tfrac{1+\delta_3^{k+1}}{1-\|\mathbf{x}^{k+1}-\mathbf{x}^k\|_{\mathbf{x}^k}}\|\mathbf{x}^{k+1}-\mathbf{z}^k\|_{\mathbf{x}^k}\\ &\le \tfrac{1+\delta_3^{k+1}}{1-\frac{\|\mathbf{x}^{k+1}-\mathbf{x}^k\|_{\mathbf{x}^k}}{1-\delta_3^k}}\tfrac{\|\mathbf{x}^{k+1}-\mathbf{z}^k\|_{\mathbf{x}^k}}{1-\delta_3^k} = \tfrac{1+\delta_3^{k+1}}{1-\delta_3^k-\|\mathbf{x}^{k+1}-\mathbf{x}^k\|_{\mathbf{x}^k}}\|\mathbf{x}^{k+1}-\mathbf{z}^k\|_{\mathbf{x}^k}\\ &= \tfrac{(1+\delta_3^{k+1})(1-\alpha_k)\lambda_k}{1-\delta_3^k-\alpha_k\lambda_k}.\end{aligned} \tag{A.40}$$

For the first item, by the inexact subproblem setting (5.20) and nonexpansiveness of the proximal operator $\mathcal{P}_{\mathbf{x}}$, we have

$$\begin{aligned}&\|\bar{\mathbf{z}}^{k+1}-\bar{\mathbf{z}}^k\|_{\mathbf{x}^{k+1}}\\ =\;&\|\mathcal{P}_{\mathbf{x}^{k+1}}(\mathbf{x}^{k+1}+H_{k+1}^{-1}(\nu^{k+1}-g_{k+1}))-\mathcal{P}_{\mathbf{x}^{k+1}}(\bar{\mathbf{z}}^k+H_{k+1}^{-1}(\nu^k-r_{\mathbf{x}^k}(\bar{\mathbf{z}}^k)))\|_{\mathbf{x}^{k+1}}\\ \le\;&\|(\mathbf{x}^{k+1}+H_{k+1}^{-1}(\nu^{k+1}-g_{k+1}))-(\bar{\mathbf{z}}^k+H_{k+1}^{-1}(\nu^k-r_{\mathbf{x}^k}(\bar{\mathbf{z}}^k)))\|_{\mathbf{x}^{k+1}}\\ =\;&\|(H_{k+1}-H_k)(\mathbf{x}^{k+1}-\bar{\mathbf{z}}^k)-(g_{k+1}-g_k-H_k(\mathbf{x}^{k+1}-\mathbf{x}^k))+(\nu^{k+1}-\nu^k)\|_{\mathbf{x}^{k+1}}^*.\end{aligned} \tag{A.41}$$

For the last item of (A.41), by triangle inequality we have

$$\begin{aligned}\|\nu^{k+1}-\nu^k\|_{\mathbf{x}^{k+1}}^* &\le \|\nu^{k+1}\|_{\mathbf{x}^{k+1}}^*+\|\nu^k\|_{\mathbf{x}^{k+1}}^*\\ &\overset{(5.8)}{\le} \|\nu^{k+1}\|_{\mathbf{x}^{k+1}}^*+\tfrac{(1-(\delta_3^k)^2)}{(1-\delta_3^{k+1})(1-\delta_3^k-\alpha_k\lambda_k)}\|\nu^k\|_{\mathbf{x}^k}^*\\ &\overset{\text{below}(5.20)}{\le} \delta_4^{k+1}\lambda_{k+1}+\tfrac{1-(\delta_3^k)^2}{(1-\delta_3^{k+1})(1-\delta_3^k-\alpha_k\lambda_k)}\delta_4^k\lambda_k.\end{aligned} \tag{A.42}$$

Next, using triangle inequality, we can derive that

$$\begin{aligned}&\|(H_{k+1}-H_k)(\mathbf{x}^{k+1}-\bar{\mathbf{z}}^k)-(g_{k+1}-g_k-H_k(\mathbf{x}^{k+1}-\mathbf{x}^k))\|_{\mathbf{x}^{k+1}}^*\\ \le\;&\|H_{k+1}(\mathbf{x}^{k+1}-\mathbf{z}^k)\|_{\mathbf{x}^{k+1}}^*+\|H_k(\mathbf{x}^{k+1}-\mathbf{z}^k)\|_{\mathbf{x}^{k+1}}^*\\ &+\|f'_{k+1}-g_{k+1}\|_{\mathbf{x}^{k+1}}^*+\|f'_k-g_k\|_{\mathbf{x}^{k+1}}^*\\ &+\|f'_{k+1}-f'_k-\nabla^2 f(\mathbf{x}^k)(\mathbf{x}^{k+1}-\mathbf{x}^k)\|_{\mathbf{x}^{k+1}}^*+\|(H_k-\nabla^2 f(\mathbf{x}^k))(\mathbf{x}^{k+1}-\mathbf{x}^k)\|_{\mathbf{x}^{k+1}}^*\\ \le\;&\tfrac{(1+\delta_3^{k+1})(1-\alpha_k)}{1-\delta_3^k-\alpha_k\lambda_k}\lambda_k+\tfrac{(1-(\delta_3^k)^2)(1-\alpha_k)}{(1-\delta_3^{k+1})(1-\delta_3^k-\alpha_k\lambda_k)}\lambda_k+\delta_2^{k+1}+\tfrac{(1-(\delta_3^k)^2)\delta_2^k}{(1-\delta_3^{k+1})(1-\delta_3^k-\alpha_k\lambda_k)}\\ &+\tfrac{1}{1-\delta_3^{k+1}}\left(\tfrac{\alpha_k\lambda_k}{1-\delta_3^k-\alpha_k\lambda_k}\right)^2+\tfrac{(2+\delta_3^k)\delta_3^k}{1-\delta_3^{k+1}}\tfrac{\alpha_k\lambda_k}{1-\delta_3^k-\alpha_k\lambda_k}\end{aligned} \tag{A.43}$$

where the last inequality follows from (5.3)(5.7) and (iPNA). Using triangle inequality for (A.41), adding (A.40)(A.42), and (A.43) together back to (A.39), and rearranging the result, we get the desired inequality. □

A.3.6 Implementation details: Approximate proximal-Newton directions

When solving the subproblem in iPNA to compute a proximal-Newton direction, we use FISTA [4]. At the j-th iteration, a new estimate $\mathbf{d}^j$ is computed through the following update

$$\mathbf{d}^j = \text{prox}_{\alpha R}\left(\mathbf{x}^k + \mathbf{w} - \alpha(g(\mathbf{x}^k) + H(\mathbf{x}^k)\mathbf{w})\right) - \mathbf{x}^k,$$

where w is calculated as

$$\mathbf{w} = \mathbf{d}^{j-1} + \frac{t_{j-1} - 1}{t_j}(\mathbf{d}^{j-1} - \mathbf{d}^{j-2}).$$

By definition of proximal operator $\text{prox}_{\alpha R}$, the following relation holds:

$$\frac{1}{\alpha}(\mathbf{w} - \mathbf{d}^j) \in g(\mathbf{x}^k) + H(\mathbf{x}^k)\mathbf{w} + \partial R(\mathbf{x}^k + \mathbf{d}^j),$$

which yields that the vector

$$\nu := \frac{\mathbf{w} - \mathbf{d}^j}{\alpha} + H(\mathbf{x}^k)(\mathbf{d}^j - \mathbf{w}) = \left(\frac{\mathbb{I}_p}{\alpha} - H(\mathbf{x}^k)\right)(\mathbf{w} - \mathbf{d}^j)$$

satisfies the condition $\nu \in g(\mathbf{x}^k) + H(\mathbf{x}^k)(\mathbf{d}^j) + \partial R(\mathbf{x}^k + \mathbf{d}^j)$. In our implementation, ν was used in the condition (5.20) to determine whether to accept this $\mathbf{d}^j$ as an inexact direction at iteration k in iPNA.

BIBLIOGRAPHY

[1] Francis Bach et al. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

[2] Francis R Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1):595–627, 2014.

[3] Heinz H Bauschke, Patrick L Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 2011. Springer, 2017.

[4] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[5] Stephen Becker, Jérôme Bobin, and Emmanuel J Candès. Nesta: A fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39, 2011.

[6] Stephen Becker and Jalal Fadili. A quasi-newton proximal splitting method. In *Advances in Neural Information Processing Systems*, pages 2618–2626, 2012.

[7] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton University Press, 2009.

[8] José M Bioucas-Dias and Mário AT Figueiredo. A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Transactions on Image processing*, 16(12):2992–3004, 2007.

[9] Raghu Bollapragada, Richard H Byrd, and Jorge Nocedal. Exact and inexact subsampled newton methods for optimization. *IMA Journal of Numerical Analysis*, 2016.

[10] J Frédéric Bonnans. Local analysis of newton-type methods for variational inequalities and nonlinear programming. *Applied mathematics & optimization*, 29(2):161–186, 1994.

[11] Allan Borodin, Ran El-Yaniv, and Vincent Gogan. Can we learn to beat the best stock. In *Advances in Neural Information Processing Systems*, pages 345–352, 2004.

[12] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[13] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[14] Richard H Byrd, Samantha L Hansen, Jorge Nocedal, and Yoram Singer. A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016.

[15] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

[16] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

[17] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[18] Tsung-Hui Chang, Mingyi Hong, Wei-Cheng Liao, and Xiangfeng Wang. Asynchronous distributed admm for large-scale optimizationpart i: algorithm and convergence analysis. *IEEE Transactions on Signal Processing*, 64(12):3118–3130, 2016.

[19] Tzu-Yi Chen and James W Demmel. Balancing sparse matrices for computing eigenvalues. *Linear algebra and its applications*, 309(1-3):261–287, 2000.

[20] Radek Cibulka, A Dontchev, and Michel H Geoffroy. Inexact newton methods and dennis–moré theorems for nonsmooth generalized equations. *SIAM Journal on Control and Optimization*, 53(2):1003–1019, 2015.

[21] Michael B Cohen, Aleksander Madry, Dimitris Tsipras, and Adrian Vladu. Matrix scaling and balancing via box constrained newton's method and interior point methods. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*, pages 902–913. IEEE, 2017.

[22] Patrick L Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

[23] Andrew R Conn, Katya Scheinberg, and Luis N Vicente. *Introduction to derivative-free optimization*, volume 8. Siam, 2009.

[24] Ron S Dembo, Stanley C Eisenstat, and Trond Steihaug. Inexact newton methods. *SIAM Journal on Numerical analysis*, 19(2):400–408, 1982.

[25] Wei Deng, Ming-Jun Lai, Zhimin Peng, and Wotao Yin. Parallel multi-block admm with o (1/k) convergence. *Journal of Scientific Computing*, 71(2):712–736, 2017.

[26] John E Dennis and Jorge J Moré. A characterization of superlinear convergence and its application to quasi-newton methods. *Mathematics of computation*, 28(126):549–560, 1974.

[27] Peter Deuflhard. *Newton methods for nonlinear problems: affine invariance and adaptive algorithms*, volume 35. Springer Science & Business Media, 2011.

[28] Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1-2):37–75, 2014.

[29] Quoc Tran Dinh, Anastasios Kyrillidis, and Volkan Cevher. A proximal newton framework for composite minimization: Graph learning without cholesky decompositions and matrix inversions. In *International Conference on Machine Learning*, pages 271–279, 2013.

[30] Quoc Tran Dinh, Ion Necoara, and Moritz Diehl. Path-following gradient-based decomposition algorithms for separable convex optimization. *Journal of Global Optimization*, 59(1):59–80, 2014.

[31] Quoc Tran Dinh, Ion Necoara, Carlo Savorgnan, and Moritz Diehl. An inexact perturbed path-following method for lagrangian decomposition in large-scale separable convex optimization. *SIAM Journal on Optimization*, 23(1):95–125, 2013.

[32] Elizabeth D Dolan and Jorge J Moré. Benchmarking optimization software with performance profiles. *Mathematical programming*, 91(2):201–213, 2002.

[33] Jonathan Eckstein and Dimitri P Bertsekas. On the douglasrachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1):293–318, 1992.

[34] Murat A Erdogdu and Andrea Montanari. Convergence rates of sub-sampled newton methods. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 3052–3060. MIT Press, 2015.

[35] Werner Fenchel. On conjugate convex functions. *Canad. J. Math*, 1(73-77), 1949.

[36] Olivier Fercoq and Zheng Qu. Restarting accelerated gradient methods with a rough strong convexity estimate. *arXiv preprint arXiv:1609.07358*, 2016.

[37] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics (NRL)*, 3(1-2):95–110, 1956.

[38] Michael P Friedlander and Gabriel Goh. Efficient evaluation of scaled proximal operators. *Electronic Transactions on Numerical Analysis*, 46:1–22, 2017.

[39] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[40] Wenbo Gao and Donald Goldfarb. Quasi-newton methods: Superlinear convergence without line search for self-concordant functions. *arXiv preprint arXiv:1612.06965*, 2016.

[41] Pontus Giselsson and Stephen Boyd. Monotonicity and restart in fast gradient methods. In *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, pages 5058–5063. IEEE, 2014.

[42] Vikas Goel and Ignacio E Grossmann. A class of stochastic programs with decision dependent uncertainty. *Mathematical programming*, 108(2):355–394, 2006.

[43] Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

[44] Michael Grant, Stephen Boyd, and Yinyu Ye. Disciplined convex programming. In *Global optimization*, pages 155–210. Springer, 2006.

[45] L Grippo, F Lampariello, and S Lucidi. A truncated newton method with nonmonotone line search for unconstrained optimization. *Journal of Optimization Theory and Applications*, 60(3):401–419, 1989.

[46] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. 2009.

[47] ZT Harmany, RF Marcia, and RM Willett. This is spiral-tap: Sparse poisson intensity reconstruction algorithms–theory and practice. *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, 21(3):1084–1096, 2012.

[48] Elad Hazan and Sanjeev Arora. *Efficient algorithms for online convex optimization and their applications*. Princeton University, 2006.

[49] Niao He, Zaid Harchaoui, Yichen Wang, and Le Song. Fast and simple optimization for poisson likelihood models. *arXiv preprint arXiv:1608.01264*, 2016.

[50] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[51] Cho-Jui Hsieh, Inderjit S Dhillon, Pradeep K Ravikumar, and Mátyás A Sustik. Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in neural information processing systems*, pages 2330–2338, 2011.

[52] Cho-Jui Hsieh, Mátyás A Sustik, Inderjit S Dhillon, Pradeep K Ravikumar, and Russell Poldrack. Big & quic: Sparse inverse covariance estimation for a million variables. In *Advances in neural information processing systems*, pages 3165–3173, 2013.

[53] DB Iudin and Arkadi S Nemirovskii. Informational complexity and efficient methods for solving complex extremal problems. *Matekon*, 13(3):25–45, 1977.

[54] Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435, 2013.

[55] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

[56] Nikos Komodakis and Jean-Christophe Pesquet. Playing with duality: An overview of recent primal dual approaches for solving large-scale optimization problems. *IEEE Signal Processing Magazine*, 32(6):31–54, 2015.

[57] Anastasios Kyrillidis, Rabeeh Karimi Mahabadi, Quoc Tran-Dinh, and Volkan Cevher. Scalable sparse covariance estimation via self-concordance. In *AAAI*, pages 1946–1952, 2014.

[58] Simon Lacoste-Julien, Fredrik Lindsten, and Francis Bach. Sequential kernel herding: Frank-wolfe optimization for particle filtering. In *Artificial Intelligence and Statistics*, pages 544–552, 2015.

[59] Guy Lebanon and John D Lafferty. Boosting and maximum likelihood for exponential models. In *Advances in neural information processing systems*, pages 447–454, 2002.

[60] Jason D Lee, Yuekai Sun, and Michael Saunders. Proximal newton-type methods for convex optimization. In *Advances in Neural Information Processing Systems*, pages 827–835, 2012.

[61] Stamatios Lefkimmiatis and Michael Unser. Poisson image reconstruction with hessian schatten-norm regularization. *IEEE transactions on image processing*, 22(11):4314–4327, 2013.

[62] Bernard Lemaire. The proximal algorithm. *International series of numerical mathematics*, 87:73–87, 1989.

[63] Jinchao Li, Martin S Andersen, and Lieven Vandenberghe. Inexact proximal newton methods for self-concordant functions. *Mathematical Methods of Operations Research*, 85(1):19–41, 2017.

[64] Lu Li and Kim-Chuan Toh. An inexact interior point method for l 1-regularized sparse covariance selection. *Mathematical Programming Computation*, 2(3-4):291–315, 2010.

[65] Chih-Jen Lin and Jorge J Moré. Newton's method for large bound-constrained optimization problems. *SIAM Journal on Optimization*, 9(4):1100–1127, 1999.

[66] Zhaosong Lu. Randomized block proximal damped newton method for composite self-concordant minimization. *SIAM Journal on Optimization*, 27(3):1910–1942, 2017.

[67] James Stephen Marron, Michael J Todd, and Jeongyoun Ahn. Distance-weighted discrimination. *Journal of the American Statistical Association*, 102(480):1267–1271, 2007.

[68] Renato DC Monteiro, Mauricio R Sicre, and Benar Fux Svaiter. A hybrid proximal extragradient self-concordant primal barrier method for monotone variational inequalities. *SIAM Journal on Optimization*, 25(4):1965–1996, 2015.

[69] Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France*, 93(2):273–299, 1965.

[70] Arkadii Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[71] Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

[72] Yu Nesterov. Accelerating the cubic regularization of newtons method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.

[73] Yu. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, Aug 2013.

[74] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[75] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13. Siam, 1994.

[76] Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[77] Jorge Nocedal and Stephen J Wright. *Numerical optimization 2nd*. Springer, 2006.

[78] Peter Ochs, Yunjin Chen, Thomas Brox, and Thomas Pock. ipiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014.

[79] Brendan Odonoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.

[80] Gergely Odor, Yen-Huan Li, Alp Yurtsever, Ya-Ping Hsieh, Quoc Tran-Dinh, Marwa El Halabi, and Volkan Cevher. Frank-wolfe works for non-lipschitz continuous gradient objectives: Scalable poisson phase retrieval. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 6230–6234. Ieee, 2016.

[81] Francesco Orabona, Andreas Argyriou, and Nathan Srebro. Prisma: Proximal iterative smoothing algorithm. *arXiv preprint arXiv:1206.2372*, 2012.

[82] Anton Osokin, Jean-Baptiste Alayrac, Isabella Lukasewitz, Puneet Dokania, and Simon Lacoste-Julien. Minding the gaps for block frank-wolfe optimization of structured svms. In *International Conference on Machine Learning*, pages 593–602, 2016.

[83] Figen Oztoprak, Jorge Nocedal, Steven Rennie, and Peder A Olsen. Newton-like methods for sparse inverse covariance estimation. In *Advances in neural information processing systems*, pages 755–763, 2012.

[84] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[85] BN Parlett and TL Landis. Methods for scaling to doubly stochastic form. *Linear Algebra and its Applications*, 48:53–79, 1982.

[86] Jiming Peng, Cornelis Roos, and Tamás Terlaky. *Self-regularity: a new paradigm for primal-dual interior-point algorithms*. Princeton University Press, 2009.

[87] Mert Pilanci and Martin J Wainwright. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245, 2017.

[88] Boris T Polyak. Introduction to optimization. translations series in mathematics and engineering. *Optimization Software*, 1987.

[89] Roman A Polyak. Regularized newton method for unconstrained convex optimization. *Mathematical programming*, 120(1):125–145, 2009.

[90] Stephen M Robinson. Strongly regular generalized equations. *Mathematics of Operations Research*, 5(1):43–62, 1980.

[91] Stephen M Robinson. Newton's method for a class of nonsmooth functions. *Set-Valued Analysis*, 2(1):291–305, 1994.

[92] Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 2015.

[93] Farbod Roosta-Khorasani and Michael W Mahoney. Sub-sampled newton methods i: globally convergent algorithms. *arXiv preprint arXiv:1601.04737*, 2016.

[94] Farbod Roosta-Khorasani and Michael W Mahoney. Sub-sampled newton methods ii: Local convergence rates. *arXiv preprint arXiv:1601.04738*, 2016.

[95] Ernest K Ryu and Stephen Boyd. Stochastic proximal iteration: a non-asymptotic improvement upon stochastic gradient descent. *Author website, early draft*, 2014.

[96] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

[97] Tianxiao Sun and Quoc Tran-Dinh. Generalized self-concordant functions: a recipe for newton-type methods. *Mathematical Programming*, May 2018.

[98] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[99] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[100] Kim-Chuan Toh, Michael J Todd, and Reha H Tütüncü. On the implementation and usage of sdpt3–a matlab software package for semidefinite-quadratic-linear programming, version 4.0. In *Handbook on semidefinite, conic and polynomial optimization*, pages 715–754. Springer, 2012.

[101] Q Tran Dinh, I Necoara, and M Diehl. A dual decomposition algorithm for separable non-convex optimization using the penalty framework. In *Proceedings of the 52nd Conference on Decision and Control*, 2013.

[102] Quoc Tran-Dinh, Anastasios Kyrillidis, and Volkan Cevher. Composite self-concordant minimization. *The Journal of Machine Learning Research*, 16(1):371–416, 2015.

[103] Quoc Tran-Dinh, Yen-Huan Li, and Volkan Cevher. Composite convex minimization involving self-concordant-like cost functions. In *Modelling, Computation and Optimization in Information Systems and Management Sciences*, pages 155–168. Springer, 2015.

[104] Quoc Tran-Dinh, Tianxiao Sun, and Shu Lu. Self-concordant inclusions: a unified framework for path-following generalized newton-type algorithms. *Mathematical Programming*, Mar 2018.

[105] Diederik Verscheure, Bram Demeulenaere, Jan Swevers, Joris De Schutter, and Moritz Diehl. Time-optimal path tracking for robots: A convex optimization approach. *IEEE Transactions on Automatic Control*, 54(10):2318–2327, 2009.

[106] Makoto Yamashita, Katsuki Fujisawa, and Masakazu Kojima. Implementation and evaluation of sdpa 6.0 (semidefinite programming algorithm 6.0). *Optimization Methods and Software*, 18(4):491–505, 2003.

[107] Tianbao Yang and Qihang Lin. Rsg: Beating subgradient method without smoothness and strong convexity. *arXiv preprint arXiv:1512.03107*, 2015.

[108] Richard Y Zhang, Salar Fattahi, and Somayeh Sojoudi. Linear-time algorithm for learning large-scale sparse graphical models. *arXiv preprint arXiv:1802.04911*, 2018.

[109] Yuchen Zhang and Xiao Lin. Disco: Distributed optimization for self-concordant empirical loss. In *International conference on machine learning*, pages 362–370, 2015.