Math. Program., Ser. A (2019) 178:145–213
https://doi.org/10.1007/s10107-018-1282-4

FULL LENGTH PAPER

Generalized self-concordant functions: a recipe for Newton-type methods

Tianxiao Sun[1] · **Quoc Tran-Dinh**[1]

Received: 4 April 2017 / Accepted: 25 April 2018 / Published online: 8 May 2018

Abstract We study the smooth structure of convex functions by generalizing a powerful concept so-called *self-concordance* introduced by Nesterov and Nemirovskii in the early 1990s to a broader class of convex functions which we call *generalized self-concordant functions*. This notion allows us to develop a unified framework for designing Newton-type methods to solve convex optimization problems. The proposed theory provides a mathematical tool to analyze both local and global convergence of Newton-type methods without imposing unverifiable assumptions as long as the underlying functionals fall into our class of generalized self-concordant functions. First, we introduce the class of generalized self-concordant functions which covers the class of standard self-concordant functions as a special case. Next, we establish several properties and key estimates of this function class which can be used to design numerical methods. Then, we apply this theory to develop several Newton-type methods for solving a class of smooth convex optimization problems involving generalized self-concordant functions. We provide an explicit step-size for a 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 mapping. 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 global and local convergence without

✉ Quoc Tran-Dinh
quoctd@email.unc.edu

Tianxiao Sun
tianxias@email.unc.edu

[1] Department of Statistics and Operations Research, University of North Carolina at Chapel Hill (UNC), 318 Hanes Hall, CB# 3260, Chapel Hill, NC 27599-3260, USA

additional assumptions. Finally, we verify our theoretical results via several numerical examples, and compare them with existing methods.

Keywords Generalized self-concordance · Newton-type method · Proximal Newton method · Quadratic convergence · Local and global convergence · Convex optimization

Mathematics Subject Classification 90C25 · 90-08

1 Introduction

The Newton method is a classical numerical scheme for solving systems of nonlinear equations and smooth optimization [47,50]. However, there are at least two reasons that prevent the use of such methods from solving large-scale problems. Firstly, while these methods often have a fast local convergence rate which can be up to a quadratic rate, their global convergence has not been well-understood [46]. In practice, one can use a damped-step scheme utilizing the Lipschitz constant of the objective derivatives to compute a suitable step-size as often seen in gradient-type methods, or incorporate the algorithm with a globalization strategy such as line-search, trust-region, or filter to guarantee a descent property [47]. Both strategies allow us to prove a global convergence of the underlying Newton-type method in some sense. Unfortunately, in practice, there exist several problems whose objective function does not have global Lipschitz gradient or Hessian such as logarithmic or reciprocal functions. This class of problems does not provide us some uniform bounds to obtain a constant step-size in optimization algorithms. On the other hand, using a globalization strategy for determining step-sizes often requires centralized computation such as function evaluations, which prevent us from using distributed computation and stochastic descent methods. Secondly, Newton algorithms are second-order methods which often require a high per-iteration complexity due to the operations on the Hessian mapping of the objective function or its approximations. In addition, these methods require the underlying functionals to be smooth up to a given smoothness levels which does not often hold in many practical models.

Motivation In recent years, there has been a great 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 [6,11,16,18,34,42,43,54,55,57,58,61]. Several combinations of Newton-type methods and other techniques such as proximal operators [8], cubic regularization [42], gradient regularization [55], randomized algorithms such as sketching [54], subsampling [18], and fast eigen-decomposition [26] have opened up a new research direction and attracted a great attention in solving nonsmooth and large-scale problems. Hitherto, research in this direction remains focusing on specific classes of problems where standard assumptions such as nonsingularity and Hessian Lipschitz continuity are preserved. However, such assumptions do not hold for many other examples as shown in [62]. Moreover, if they are satisfied, then we often get a lower bound of possible step-sizes for our algorithm which may lead to a poor performance, especially in large-scale problems.

In the seminal work [45], Nesterov and Nemirovskii showed that the class of log-barriers does not satisfy the standard assumptions of the Newton method if the solution of the underlying problem is closed to the boundary of the domain of a barrier function. They introduced a powerful concept called "self-concordance" to overcome this drawback and developed new Newton schemes to achieve global and local convergence without requiring any additional assumption, or a globalization strategy. While the self-concordance notion was initially invented to study interior-point methods, it is less well-known in other communities. Recent works [1,14,38,62,67,72] have popularized this concept to solve other problems arising from machine learning, statistics, image processing, scientific computing, and variational inequalities.

Our goals In this paper, motivated by [1,63,72], we aim at generalizing the self-concordance concept in [45] to a broader class of smooth and convex functions. To illustrate our idea, we consider a univariate smooth and convex function $\varphi : \mathbb{R} \to \mathbb{R}$. If φ satisfies the inequality $|\varphi'''(t)| \le M_\varphi \varphi''(t)^{3/2}$ for all t in the domain of φ and for a given constant $M_\varphi \ge 0$, then we say that φ is self-concordant (in Nesterov and Nemirovskii's sense [45]). We instead generalize this inequality to

$$|\varphi'''(t)| \le M_\varphi \varphi''(t)^{\frac{\nu}{2}}, \tag{1}$$

for all t in the domain of φ, and for given constants $\nu > 0$ and $M_\varphi \ge 0$.

We emphasize that generalizing from univariate to multivariate functions in the standard self-concordant case (i.e., $\nu = 3$) [45] preserves several important properties including the multilinear symmetry [40, Lemma 4.1.2], while, unfortunately, they do not hold for the case $\nu \ne 3$. Therefore, we modify the definition in [45] to overcome this drawback. Note that a similar idea has been also studied in [1,63] for a class of logistic-type functions. Nevertheless, the definition using in these papers is limited, and still creates certain difficulty for developing further theory in general cases.

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

$$x^{k+1} := x^k - s_k F'(x^k)^{-1} F(x^k), \tag{2}$$

where F can be represented as the right-hand side of a smooth monotone equation $F(x) = 0$, or the optimality condition of a convex optimization or a convex–concave saddle-point problem, F' is the Jacobian map of F, and $s_k \in (0, 1]$ is a given step-size. Despite the Newton scheme (2) is invariant to a change of variables [16], its convergence property relies on the growth of the Hessian mapping along the Newton iterative process. In classical settings, the Lipschitz continuity and the non-degeneracy of the Hessian mapping in a neighborhood of a given solution are key assumptions to achieve local quadratic convergence rate [16]. These assumptions have been considered to be standard, but they are often very difficult to check in practice, especially the second requirement. A natural idea is to classify the functionals of the underlying problem into a known class of functions to choose a suitable method for minimizing it. While first-order methods for convex optimization essentially rely on the Lipschitz gradient continuity, Newton schemes usually use the Lipschitz continuity of the Hessian mapping and its non-degeneracy to obtain a well-defined Newton direction as we have previously mentioned. For self-concordant functions, the second condition automatically holds, but the first assumption fails to satisfy. 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 convergence of Newton-type methods, see [16] for an intensive study of generic Newton-type methods, and [40,45] for the self-concordant function class.

Our approach We attempt to develop some background theory for a broad class of smooth and convex functions under the structure (1). By adopting the local norm defined via the Hessian mapping of such a convex function from [45], we can prove some lower and upper bound estimates for the local norm distance between two points in the domain as well as for the growth of the Hessian mapping. Together with this background theory, we also identify a class of functions using in generalized linear models [37,39] as well as in empirical risk minimization [68] that falls into our generalized self-concordance class for many well-known loss-type functions as listed in Table 2.

Applying our generalized self-concordant theory, we develop a class of Newton-type methods to solve the following composite convex minimization problem:

$$F^\star := \min_{x\in\mathbb{R}^p} \Big\{ F(x) := f(x) + g(x) \Big\}, \tag{3}$$

where f is a generalized self-concordant function in our context, and g is a proper, closed, and convex function that can be referred to as a regularization term. We consider two cases. The first case is a non-composite convex problem in which g is vanished (i.e., $g = 0$). In the second case, we assume that g is equipped with a "tractably" proximal operator [see (34) for the definition].

Our contribution To this end, our main contribution can be summarized as follows.

(a) We generalize the self-concordant notion in [40] to a more broader class of smooth convex functions which we call generalized self-concordance. We identify several loss-type functions that can be cast into our generalized self-concordant class. We also prove several fundamental properties and show that the sum and linear transformation of *generalized self-concordant* functions are *generalized self-concordant* for a given range of ν or under suitable assumptions.

(b) We develop lower and upper bounds on the Hessian mapping, the gradient mapping, and the function values for generalized self-concordant functions. These estimates are key to analyze several numerical optimization methods including Newton-type methods.

(c) We propose a class of Newton methods including full-step and damped-step schemes to minimize a generalized self-concordant function. We explicitly show 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 the damped-step and the full-step schemes using a suitable metric.

(d) We also extend our Newton schemes to handle the composite setting (3). We develop both full-step and damped-step proximal Newton methods to solve this problem and provide a rigorous theoretical convergence guarantee in both local and global sense.

(e) We also study a quasi-Newton variant of our Newton scheme to minimize a generalized self-concordant function. Under a modification of the well-known Dennis–Moré condition [15] or a BFGS update, we show that our quasi-Newton method locally converges at a superlinear rate to the solution of the underlying problem.

Let us emphasize the following aspects of our contribution. Firstly, we observe that the self-concordance notion is a powerful concept and has widely been used in interior-point methods as well as in other optimization schemes [28,35,62,72], generalizing it to a broader class of smooth convex functions can substantially cover a number of new applications or can develop new methods for solving old problems including logistic and multimonomial logistic regression, optimization involving exponential objectives, and distance-weighted discrimination problems in support vector machine (see Table 2 below). Secondly, 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 in order to choose suitable algorithms to solve the corresponding optimization problem. Thirdly, the theory developed in this paper can potentially apply to other optimization methods such as gradient-type, sketching and sub-sampling Newton, and Frank–Wolfe's algorithms as done in the literature [49,54,57,58,62]. Finally, we also show that it is possible to impose additional structure such as self-concordant barrier to develop path-following scheme or interior-point-type methods for solving a subclass of composite convex minimization problems of the form (3). We believe that our 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 [67].

Summary of generalized self-concordant properties We provide a short summary on the main properties of generalized self-concordant (gsc) functions in Table 1.

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 [45].

Related work Since the self-concordance concept was introduced in 1990s [45], its first extension is perhaps proposed by [1] for a class of logistic regression. In [63], 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 in [72] showed that the regularized logistic loss function is indeed standard self-concordant. In [2] Bach continued exploiting his result in [1] to show that the averaging stochastic gradient method can achieve the same best-known con-

Table 1 A summary of generalized self-concordant properties

Result	Property	Range of ν
Definitions 1 and 2	Definitions of gsc functions	$\nu > 0$
Proposition 1	Sum of gsc functions	$\nu \geq 2$
Proposition 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(a)	Non-degenerate property	$\nu \geq 2$
Proposition 3(b)	Unboundedness	$\nu > 0$
Proposition 4(a)	gsc and strong convexity	$\nu \in (0, 3]$
Proposition 4(b)	gsc and Lipschitz gradient continuity	$\nu \geq 2$
Proposition 6	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 7, 8, 9, and 10	Local norm, Hessian, gradient, and function value bounds	$\nu \geq 2$

vergence rate as in strongly convex case without adding a regularizer. In [62], the authors exploited standard self-concordance theory in [45] to develop several classes of optimization algorithms including proximal Newton, proximal quasi-Newton, and proximal gradient methods to solve composite convex minimization problems. In [35], Lu extended [62] to study randomized block coordinate descent methods. In a recent paper [22], Gao and Goldfarb investigated quasi-Newton methods for self-concordant problems. As another example, [53] proposed an alternative to the standard self-concordance, called self-regularity. The authors applied this theory to develop a new paradigm for interior-point methods. The theory developed in this paper, on the one hand, is a generalization of the well-known self-concordance notion developed in [45]; on the other hand, it also covers the work in [1,61,72] as specific examples. Several concrete applications and extensions of self-concordance notion can also be found in the literature including [28,32,49,53]. Recently, [14] exploited smooth structures of exponential functions to design interior-point methods for solving two fundamental problems in scientific computing called matrix scaling and balancing.

Paper organization The rest of this paper is organized as follows. Section 2 develops the foundation theory for our generalized self-concordant functions including definitions, examples, basic properties, Fenchel's conjugate, smoothing technique, and key bounds. Section 3 is devoted to studying full-step and damped-step Newton schemes to minimize a generalized self-concordant function including their global and local convergence guarantees. Section 4 considers the composite setting (3) and studies proximal Newton-type methods, and investigates their convergence guarantees. Section 5 deals with a quasi-Newton scheme for solving the noncomposite problem of (3). Numerical examples are provided in Sect. 6 to illustrate advantages of our theory.

Finally, for clarity of presentation, several technical results and proofs are moved to the appendix.

2 Theory of generalized self-concordant functions

We generalize the class of self-concordant functions introduced by Nesterov and Nemirovskii in [40] to a broader class of smooth and convex functions. We identify several examples of such functions. Then, we develop several properties of this function class by utilizing our new definitions.

Notation Given a proper, closed, and convex function $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$, we denote by $\mathrm{dom}(f) := \{x \in \mathbb{R}^p \mid f(x) < +\infty\}$ the domain of f, and by $\partial f(x) := \left\{w \in \mathbb{R}^p \mid f(y) \geq f(x) + \langle w, y - x\rangle, \ \forall y \in \mathrm{dom}(f)\right\}$ the subdifferential of f at $x \in \mathrm{dom}(f)$. We use $\mathcal{C}^3(\mathrm{dom}(f))$ to denote the class of three times continuously differentiable functions on its open domain $\mathrm{dom}(f)$. We denote by ∇f its gradient map, by $\nabla^2 f$ its Hessian map, and by $\nabla^3 f$ its third-order derivative. For a twice continuously differentiable convex function f, $\nabla^2 f$ is symmetric positive semidefinite, and can be written as $\nabla^2 f(\cdot) \succeq 0$. If it is positive definite, then we write $\nabla^2 f(\cdot) \succ 0$.

Let $\mathbb{R}_+$ and $\mathbb{R}_{++}$ denote the sets of nonnegative and positive real numbers, respectively. We use $\mathcal{S}^p_+$ and $\mathcal{S}^p_{++}$ to denote the sets of symmetric positive semidefinite and symmetric positive definite matrices of the size $p \times p$, respectively. Given a $p \times p$ matrix $H \succ 0$, we define a weighted norm with respect to H as $\|u\|_H := \langle Hu, u\rangle^{1/2}$ for $u \in \mathbb{R}^p$. The corresponding dual norm is $\|v\|^*_H := \left\langle H^{-1}v, v\right\rangle^{1/2}$. If $H = \mathbb{I}$, the identity matrix, then $\|u\|_H = \|u\|^*_H = \|u\|_2$, where $\|\cdot\|_2$ is the standard Euclidean norm. Note that $\|\cdot\|^*_2 = \|\cdot\|_2$.

We say that f is strongly convex with the strong convexity parameter $\mu_f > 0$ if $f(\cdot) - \frac{\mu_f}{2}\|\cdot\|^2$ is convex. We also say that f has Lipschitz gradient if ∇f is Lipschitz continuous with the Lipschitz constant $L_f \in [0, +\infty)$, i.e., $\|\nabla f(x) - \nabla f(y)\|^* \leq L_f\|x - y\|$ for all $x, y \in \mathrm{dom}(f)$.

For $f \in \mathcal{C}^3(\mathrm{dom}(f))$, if $\nabla^2 f(x) \succ 0$ at a given $x \in \mathrm{dom}(f)$, then we define a local norm $\|u\|_x := \langle\nabla^2 f(x)u, u\rangle^{1/2}$ as a weighted norm of u with respect to $\nabla^2 f(x)$. The corresponding dual norm $\|v\|^*_x$, is defined as $\|v\|^*_x := \max\left\{\langle v, u\rangle \mid \|u\|_x \leq 1\right\} = \left\langle\nabla^2 f(x)^{-1}v, v\right\rangle^{1/2}$ for $v \in \mathbb{R}^p$.

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 $\varphi \in \mathcal{C}^3\,(\mathrm{dom}(\varphi))$. 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 1 Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a $\mathcal{C}^3\,(\mathrm{dom}(\varphi))$ and univariate function with open domain $\mathrm{dom}(\varphi)$, and $\nu > 0$ and $M_\varphi \geq 0$ be two constants. We say that φ is (M_φ, ν)-generalized self-concordant if

$$|\varphi'''(t)| \leq M_\varphi\varphi''(t)^{\frac{\nu}{2}}, \quad \forall t \in \mathrm{dom}(\varphi). \tag{4}$$

The inequality (4) 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 $b \in \mathbb{R}$, then we have $\varphi''(t) = a$ and $\varphi'''(t) = 0$. The inequality (4) 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 is $(0, \nu)$-*generalized self-concordant* for any $\nu > 0$. While (4) holds for any other constant $\hat{M}_\varphi \geq M_\varphi$, we often require that M_φ is the smallest constant satisfying (4).

Example 1 Let us now provide some common examples satisfying Definition 1.

(a) *Standard self-concordant functions* If we choose $\nu = 3$, then (4) becomes $|\varphi'''(t)| \leq M_\varphi \varphi''(t)^{3/2}$ which is the standard self-concordant functions in $\mathbb{R}$ introduced in [45].
(b) *Logistic functions* In [1], Bach modified the standard self-concordant inequality in [45] to obtain $\left|\varphi'''(t)\right| \leq M_\varphi \varphi''(t)$, and showed that the well-known logistic loss $\varphi(t) := \log(1+e^{-t})$ satisfies this definition. In [63] the authors also exploited this definition, and developed a class of first-order and second-order methods to solve composite convex minimization problems. Hence, $\varphi(t) := \log(1 + e^{-t})$ is a generalized self-concordant function with $M_\varphi = 1$ and $\nu = 2$.
(c) *Exponential functions* The exponential function $\varphi(t) := e^{-t}$ also belongs to (4) with $M_\varphi = 1$ and $\nu = 2$. This function is often used, e.g., in Ada-boost [33], or in matrix scaling [14].
(d) *Distance-weighted discrimination (DWD)* We consider a more general function $\varphi(t) := \frac{1}{t^q}$ on $\mathrm{dom}(\varphi) = \mathbb{R}_{++}$ and $q \geq 1$ studied in [36] for DWD using in support vector machine. As shown in Table 2, this function satisfies Definition 1 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, it is generalized self-concordant with $\nu = 4$ and $M_\varphi = 1$ in the sense of Definition 1.
(f) *Arcsine distribution* We 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 it satisfies Definition 1 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 [71] for robust regression using in statistics. Then, $M_\varphi = \frac{2-q}{\sqrt[(2-q)]{q(q-1)}}$ and $\nu = \frac{2(3-q)}{2-q} \in (4, +\infty)$.

As concrete examples, the following table, Table 2, provides a non-exhaustive list of generalized self-concordant functions used in the literature.

Remark 1 All examples given in Table 2 fall into the case $\nu \geq 2$. However, we note that Definition 1 also covers [72, 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 2 Examples of univariate generalized self-concordant functions ($\mathcal{F}_L^{1,1}$ means that $\nabla\varphi$ is Lipschitz continuous)

Function name	Form of $\varphi(t)$	ν	M_f	dom(φ)	Application	$\mathcal{F}_L^{1,1}$	References
Log-barrier	$-\ln(t)$	3	2	$\mathbb{R}_{++}$	Poisson	No	[10,40,45]
Entropy-barrier	$t\ln(t) - \ln(t)$	3	2	$\mathbb{R}_{++}$	Interior-point	No	[40]
Logistic	$\ln(1 + e^{-t})$	2	1	$\mathbb{R}$	Classification	yes	[29]
Exponential	e^{-t}	2	1	$\mathbb{R}$	AdaBoost, etc	No	[14,33]
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	[36]
Arcsine distribution	$\frac{1}{\sqrt{1-t^2}}$	$\frac{14}{5}$	< 3.25	$(-1, 1)$	Random walks	No	[24]
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	[71]
Entropy	$t\ln(t)$	4	1	$\mathbb{R}_{+}$	KL divergence	No	[10]

Table 2 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 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 $-\log(t)$ which are generalized self-concordant with $\nu = 4$ and $\nu = 3$, respectively.

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, $x \in \mathrm{dom}(f)$, and $u, v \in \mathbb{R}^p$, we consider the function $\psi(t) := \langle \nabla^2 f(x + tv)u, u\rangle$. Then, it is obvious to show that

$$\psi'(t) := \langle \nabla^3 f(x + tv)[v]u, u\rangle.$$

for $t \in \mathbb{R}$ such that $x + tv \in \mathrm{dom}(f)$, where $\nabla^3 f$ is the third-order derivative of f. It is clear that $\psi(0) = \langle \nabla^2 f(x)u, u\rangle = \|u\|_x^2$. By using the local norm, we generalize Definition 1 to multivariate functions $f : \mathbb{R}^p \to \mathbb{R}$ as follows.

Definition 2 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 $x \in \mathrm{dom}(f)$ and $u, v \in \mathbb{R}^p$, it holds

$$\left|\langle \nabla^3 f(x)[v]u, u\rangle\right| \leq M_f \, \|u\|_x^2 \, \|v\|_x^{\nu-2} \, \|v\|_2^{3-\nu}. \tag{5}$$

Here, we use a convention that $\frac{0}{0} = 0$ for the case $\nu < 2$ or $\nu > 3$. We denote this class of functions by $\widetilde{\mathcal{F}}_{M_f,\nu}(\mathrm{dom}(f))$ (shortly, $\widetilde{\mathcal{F}}_{M_f,\nu}$ when $\mathrm{dom}(f)$ is explicitly defined).

Let us consider the following two extreme cases:

1. If $\nu = 2$, (5) leads to $\left|\langle \nabla^3 f(x)[v]u, u\rangle\right| \le M_f \,\|u\|_x^2\, \|v\|_2$ which collapses to the definition introduced in [1] by letting $u = v$.
2. If $\nu = 3$ and $u = v$, (5) reduces to $\left|\langle \nabla^3 f(x)[u]u, u\rangle\right| \le M_f \,\|u\|_x^3$, Definition 2 becomes the standard self-concordant definition introduced in [40,45].

We emphasize that Definition 2 is not symmetric, but can avoid the use of multilinear mappings as required in [1,45]. However, by [45, Proposition 9.1.1] or [40, Lemma 4.1.2], Definition 2 with $\nu = 3$ is equivalent to [40, Definition 4.1.1] for standard self-concordant functions.

2.3 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 a *generalized self-concordant* for any $\beta_1, \beta_2 > 0$ according to Definition 2.

Proposition 1 (Sum of *generalized self-concordant* functions) *Let f_i be (M_{f_i}, ν)-generalized self-concordant functions satisfying* (5)*, where $M_{f_i} \ge 0$ and $\nu \ge 2$ for $i = 1, \ldots, m$. Then, for $\beta_i > 0$, $i = 1, 2, \ldots, m$, the function $f(x) := \sum_{i=1}^m \beta_i f_i(x)$ is well-defined on $\mathrm{dom}(f) = \bigcap_{i=1}^m \mathrm{dom}(f_i)$, and is (M_f, ν)-generalized self-concordant with the same order $\nu \ge 2$ and the constant*

$$M_f := \max\left\{\beta_i^{1-\frac{\nu}{2}} M_{f_i} \mid 1 \le i \le m\right\} \ge 0.$$

Proof It is sufficient to prove for $m = 2$. For $m > 2$, it follows from $m = 2$ by induction. By [40, 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 $x \in \mathrm{dom}(f)$ and $u, v \in \mathbb{R}^p$. Then, by Definition 2, we have

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

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

$$\begin{aligned} \frac{\left|\langle \nabla^3 f(x)[v]u, u\rangle\right|}{\langle \nabla^2 f(x)u, u\rangle \langle \nabla^2 f(x)v, v\rangle^{\frac{\nu-2}{2}}} &\le \frac{\beta_1 \left|\langle \nabla^3 f_1(x)[v]u, u\rangle\right| + \beta_2 \left|\langle \nabla^3 f_2(x)[v]u, u\rangle\right|}{\langle \nabla^2 f(x)u, u\rangle \langle \nabla^2 f(x)v, v\rangle^{\frac{\nu-2}{2}}} \\ &\le \left[\frac{M_{f_1}\beta_1 w_1 s_1^{\frac{\nu-2}{2}} + M_{f_2}\beta_2 w_2 s_2^{\frac{\nu-2}{2}}}{(\beta_1 w_1 + \beta_2 w_2)(\beta_1 s_1 + \beta_2 s_2)^{\frac{\nu-2}{2}}}\right]_{[T]} \|v\|_2^{3-\nu}. \end{aligned} \tag{6}$$

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, $\frac{\beta_2 w_2}{\beta_1 w_1 + \beta_2 w_2} = 1 - \xi \ge 0$ and $\frac{\beta_2 s_2}{\beta_1 s_1 + \beta_2 s_2} = 1 - \eta \ge 0$. Hence, the term $[T]$ in the square brackets of (6) 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

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

Using this estimate into (6), we can show that $f(\cdot) := \beta_1 f_1(\cdot) + \beta_2 f_2(\cdot)$ is (M_f,ν)-generalized self-concordant with $M_f := \max\left\{\beta_1^{1-\frac{\nu}{2}} M_{f_1}, \beta_2^{1-\frac{\nu}{2}} M_{f_2}\right\}$. □

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

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

Proposition 2 (Affine transformation) *Let $\mathcal{A}(x) := Ax + b$ be an affine transformation from $\mathbb{R}^p$ to $\mathbb{R}^q$, and f be an (M_f,ν)-generalized self-concordant function with $\nu > 0$. Then, the following statements hold:*

(a) *If $\nu \in (0,3]$, then $g(x) := f(\mathcal{A}(x))$ is (M_g,ν)-generalized self-concordant with $M_g := M_f\,\|A\|^{3-\nu}$.*

(b) *If $\nu > 3$ and $\lambda_{\min}(A^\top A) > 0$, then $g(x) := f(\mathcal{A}(x))$ is (M_g,ν)-generalized self-concordant with $M_g := M_f\lambda_{\min}(A^\top A)^{\frac{3-\nu}{2}}$, where $\lambda_{\min}(A^\top A)$ is the smallest eigenvalue of $A^\top A$.*

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

$$\begin{aligned} |\langle\nabla^3 g(x)[v]u, u\rangle| &= |\langle A^\top(\nabla^3 f(\tilde{x})[\tilde{v}])Au, u\rangle| = |\langle\nabla^3 f(\tilde{x})[\tilde{v}]\tilde{u}, \tilde{u}\rangle| \\ &\overset{(5)}{\leq} M_f\langle\nabla^2 f(\tilde{x})\tilde{u}, \tilde{u}\rangle\langle\nabla^2 f(\tilde{x})\tilde{v}, \tilde{v}\rangle^{\frac{\nu}{2}-1}\,\|\tilde{v}\|_2^{3-\nu} \\ &= M_f\langle A^\top\nabla^2 f(\mathcal{A}(x))Au, u\rangle\langle A^\top\nabla^2 f(\mathcal{A}(x))Av, v\rangle^{\frac{\nu}{2}-1}\|Av\|_2^{3-\nu} \\ &= M_f\langle\nabla^2 g(x)u, u\rangle\langle\nabla^2 g(x)v, v\rangle^{\frac{\nu}{2}-1}\|Av\|_2^{3-\nu}. \end{aligned} \tag{7}$$

(a) If $\nu \in (0,3]$, then we have $\|Av\|_2^{3-\nu} \leq \|A\|^{3-\nu}\|v\|_2^{3-\nu}$. Hence, the last inequality (7) implies

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

which shows that g is (M_g, ν)-generalized self-concordant with $M_g := M_f\|A\|^{3-\nu}$. (b) Note that $\|Av\|_2^2 = v^\top A^\top Av \ge \lambda_{\min}(A^\top A)\|v\|_2^2 \ge 0$, where $\lambda_{\min}(A^\top A)$ is the smallest eigenvalue of $A^\top A$. If $\lambda_{\min}(A^\top A) > 0$ and $\nu > 3$, then we have $\|Av\|_2^{3-\nu} \le \lambda_{\min}(A^\top A)^{\frac{3-\nu}{2}}\|v\|_2^{3-\nu}$. Combining this estimate and (7), we can show that g is (M_g, ν)-generalized self-concordant with $M_g := M_f\lambda_{\min}(A^\top A)^{\frac{3-\nu}{2}}$. □

Remark 2 Proposition 2 shows that generalized self-concordance is preserved via an affine transformations if $\nu \in (0, 3]$. If $\nu > 3$, then it requires A to be over-completed, i.e., $\lambda_{\min}(A^\top A) > 0$. Hence, the theory developed in the sequel remains applicable for $\nu > 3$ if A is over-completed.

The following result is an extension of standard self-concordant functions ($\nu = 3$), whose proof is very similar to [40, Theorems 4.1.3, 4.1.4] by replacing the parameters $M_f = 2$ and $\nu = 3$ with the general parameters $M_f \ge 0$ and $\nu > 0$ (or $\nu \ge 2$), respectively. We omit the detailed proof.

Proposition 3 *Let f be an (M_f, ν)-generalized self-concordant function with $\nu > 0$. Then:*

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

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

Note that Proposition 3(a) only holds for $\nu \ge 2$. If we consider $g(x) := f(\mathcal{A}(x))$ for a given affine operator $\mathcal{A}(x) = Ax + b$, then the non-degenerateness of $\nabla^2 g$ is only guaranteed if A is full-rank. Otherwise, it is non-degenerated in a given subspace of A.

2.4 Generalized self-concordant functions with special structures

We first show that if a generalized self-concordant function is strongly convex or has a Lipschitz gradient, then it can be cast into the special case $\nu = 2$ or $\nu = 3$.

Proposition 4 *Let $f \in \widetilde{\mathcal{F}}_{M_f,\nu}$ be an (M_f, ν)-generalized self-concordant with $\nu > 0$. Then:*

(a) *If $\nu \in (0, 3]$ and f is also strongly convex on* $\mathrm{dom}(f)$ *with the strong convexity parameter $\mu_f > 0$ in ℓ_2-norm, then f is also $(\hat{M}_f, \hat{\nu})$-generalized self-concordant with $\hat{\nu} = 3$ and $\hat{M}_f := \frac{M_f}{(\sqrt{\mu_f})^{3-\nu}}$.*

(b) *If $\nu \ge 2$ and ∇f is Lipschitz continuous with the Lipschitz constant $L_f \in [0, +\infty)$ in ℓ_2-norm, then f is also $(\hat{M}_f, \hat{\nu})$-generalized self-concordant with $\hat{\nu} = 2$ and $\hat{M}_f := M_f L_f^{\frac{\nu}{2}-1}$.*

Proof (a) If f is strongly convex with the strong convexity parameter $\mu_f > 0$ in ℓ_2-norm, then we have $\langle \nabla^2 f(x)v, v\rangle \geq \mu_f \|v\|_2^2$ for any $v \in \mathbb{R}^p$. Hence, $\frac{\|v\|_2}{\|v\|_x} \leq \frac{1}{\sqrt{\mu_f}}$. In this case, (5) leads to

$$\left|\langle \nabla^3 f(x)[v]u, u\rangle\right| \leq M_f \|u\|_x^2 \left(\frac{\|v\|_2}{\|v\|_x}\right)^{3-\nu} \|v\|_x \leq \frac{M_f}{(\sqrt{\mu_f})^{3-\nu}} \|u\|_x^2 \|v\|_x.$$

Hence, f is $(\hat{M}_f, \hat{\nu})$-generalized self-concordant with $\hat{\nu} = 3$ and $\hat{M}_f := \frac{M_f}{(\sqrt{\mu_f})^{3-\nu}}$.

(b) Since ∇f is Lipschitz continuous with the Lipschitz constant $L_f \in [0, +\infty)$ in ℓ_2-norm, we have $\|v\|_x^2 = \langle \nabla^2 f(x)v, v\rangle \leq L_f \|v\|_2^2$ for all $v \in \mathbb{R}^p$ which leads to $\frac{\|v\|_x}{\|v\|_2} \leq \sqrt{L_f}$ for all $v \in \mathbb{R}^p$. On the other hand, $f \in \widetilde{\mathcal{F}}_{M_f,\nu}$ with $\nu \geq 2$, we can show that

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

Hence, f is also $(\hat{M}_f, \hat{\nu})$-generalized self-concordant with $\hat{\nu} = 2$ and $\hat{M}_f := M_f L_f^{\frac{\nu-2}{2}}$. □

Proposition 4 provides two important properties. If the gradient map ∇f of a generalized self-concordant function f is Lipschitz continuous, 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. This idea is also applied to generalized self-concordant and strongly convex functions.

Given n smooth convex univariate functions $\varphi_i : \mathbb{R} \to \mathbb{R}$ satisfying (4) 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 form:

$$f(x) := \frac{1}{n}\sum_{i=1}^{n} \varphi_i(a_i^\top x + b_i), \tag{8}$$

where $a_i \in \mathbb{R}^p$ and $b_i \in \mathbb{R}$ are given vectors and numbers, respectively for $i = 1, \ldots, n$. This convex function is called a finite sum and widely used in machine learning and statistics. The decomposable structure in (8) often appears in generalized linear models [7,11], and empirical risk minimization [72], where φ_i is referred to as a loss function as can be found, e.g., in Table 2.

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 Propositions 1 and 2.

Corollary 1 *If φ_i in (8) satisfies (4) for $i = 1, \ldots, n$ with the same order $\nu \in [2, 3]$ and $M_{\varphi_i} \geq 0$, then f defined by (8) is also (M_f, ν)-generalized self-concordant in the sense of Definition 2 with the same order ν and the constant $M_f := n^{\frac{\nu}{2}-1} \max\left\{M_{\varphi_i} \|a_i\|_2^{3-\nu} \mid 1 \leq i \leq n\right\}$.*

Finally, we show that if we regularize f in (8) by a strongly convex quadratic term, then the resulting function becomes self-concordant. The proof can follow the same path as [72, Lemma 2].

Proposition 5 *Let $f(x) := \frac{1}{n}\sum_{i=1}^{n}\varphi_i(a_i^{\top}x+b_i)+\psi(x)$, where $\psi(x) := \frac{1}{2}\langle Qx, x\rangle + c^{\top}x$ is strongly convex quadratic function with $Q \in \mathcal{S}^p_{++}$. If φ_i satisfies (4) for $i = 1, \ldots, n$ with the same order $\nu \in (0, 3]$ and a constant $M_{\varphi_i} > 0$, then f is $(\hat{M}_f, 3)$-generalized self-concordant in the sense of Definition 2 with $\hat{M}_f := \lambda_{\min}(Q)^{\frac{\nu-3}{2}} \max\left\{M_{\varphi_i}\|a_i\|_2^{3-\nu} \mid 1 \le i \le n\right\}$.*

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

The 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

$$f^*(x) = \sup_u \left\{\langle x, u\rangle - f(u) \mid u \in \mathrm{dom}(f)\right\}. \tag{9}$$

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 $u^*(x)$ satisfies $\nabla f(u^*(x)) = x$, then f^* is well-defined at x. This shows that $\mathrm{dom}(f^*) = \{x \in \mathbb{R}^p \mid \nabla f(u^*(x)) = x \text{ is solvable}\}$.

Example 2 Let us look at some univariate functions. By using (9), we can directly show that:

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

Intuitively, these examples show that if φ is generalized self-concordant, then its conjugate φ^* is also generalized self-concordant. For more examples, we refer to [3, Chapter 13]. Let us generalize this result in the following proposition whose proof is given in Appendix A.1.

Proposition 6 *If f is (M_f, ν)-generalized self-concordant in $\mathrm{dom}(f) \subseteq \mathbb{R}^p$ such that $\nabla^2 f(x) \succ 0$ for $x \in \mathrm{dom}(f)$, then the conjugate function f^* of f given by (9) is well-defined, and (M_{f^*}, ν_*)-generalized self-concordant on*

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

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 $\nabla f^*(x) = u^*(x)$ and $\nabla^2 f^*(x) = \nabla^2 f(u^*(x))^{-1}$, where $u^*(x)$ is a unique solution of the maximization problem $\max_u \{\langle x, u\rangle - f(u) \mid u \in \mathrm{dom}(f)\}$ in (9) for any $x \in \mathrm{dom}(f^*)$.

Proposition 6 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., [65,66].

2.6 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.

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 [41]

$$f_\gamma(x) := \sup_{u \in \mathrm{dom}(f^*)} \left\{ \langle x, u \rangle - f^*(u) - \gamma \omega(u) \right\}, \tag{10}$$

where f^* is the Fenchel conjugate of f, $\omega : \mathrm{dom}(\omega) \subseteq \mathbb{R}^p \to \mathbb{R}$ is a smooth convex function such that $\mathrm{dom}(f^*) \subseteq \mathrm{dom}(\omega)$, and $\gamma > 0$ is called a smoothness parameter. In particular, if f is Lipschitz continuous, then $\mathrm{dom}(f^*)$ is bounded [3]. Hence, the sup operator in (10) reduces to the max operator.

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 Let us provide a few examples of well-known nonsmooth convex functions:

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

$$\begin{aligned} \|x\|_1 &= \max_u \{\langle x, u \rangle \mid \|u\|_\infty \le 1\} \\ &= \max_{u,v} \left\{ \langle x, u - v \rangle \mid \sum_{i=1}^p (u_i + v_i) = 1,\ u, v \in \mathbb{R}^p_+ \right\}. \end{aligned}$$

We can smooth this function by f_γ by choosing $\omega(u, v) := \ln(2p) + \sum_{i=1}^p (u_i \ln(u_i) + v_i \ln(v_i))$. In this case, we obtain $f_\gamma(x) = \gamma \ln \left(\sum_{i=1}^p \left(e^{x_i/\gamma} + e^{-x_i/\gamma} \right) \right) - \gamma \ln(2p)$. This function is clearly generalized self-concordant with $\nu = 2$, see [63, Lemma 4].

However, if we choose $\omega(u) := p - \sum_{i=1}^p \sqrt{1 - u_i^2}$, then we get $f_\gamma(x) = \sum_{i=1}^p \sqrt{x_i^2 + \gamma^2} - \gamma p$. In this case, f_γ is also generalized self-concordant 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 $\nu = 2$.

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 (10) becomes

$$f_\gamma(x) := \sup_u \left\{\langle x, u\rangle - \varphi(u) - \gamma\omega(u) \mid u \in \mathcal{U}\right\}. \tag{11}$$

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

2.7 Key bounds on Hessian map, gradient map, and function values

Now, we develop some key bounds on the local norms, Hessian map, gradient map, and function values of generalized self-concordant functions. In this subsection, we assume that the Hessian map $\nabla^2 f$ of f is nondegenerate at any point in its domain.

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

$$d_\nu(x, y) := \begin{cases} M_f \|y - x\|_2 & \text{if } \nu = 2 \\ \left(\frac{\nu}{2} - 1\right) M_f \|y - x\|_2^{3-\nu} \|y - x\|_x^{\nu-2} & \text{if } \nu > 2. \end{cases} \tag{12}$$

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

$$\bar{\bar{\omega}}_\nu(\tau) := \begin{cases} \frac{1}{(1-\tau)^{\frac{2}{\nu-2}}} & \text{if } \nu > 2 \\ e^\tau & \text{if } \nu = 2, \end{cases} \tag{13}$$

with $\mathrm{dom}(\bar{\bar{\omega}}_\nu) = (-\infty, 1)$ if $\nu > 2$, and $\mathrm{dom}(\bar{\bar{\omega}}_\nu) = \mathbb{R}$ if $\nu = 2$. We also adopt the Dikin ellipsoidal notion from [40] as $W^0(x; r) := \{y \in \mathbb{R}^p \mid d_\nu(x, y) < r\}$.

The next proposition provides a bound on the local norm defined by a *generalized self-concordant* function f. This bound is given for the local distances $\|y - x\|_x$ and $\|y - x\|_y$ between two points x and y in $\mathrm{dom}(f)$.

Proposition 7 (Bound of local norms) *If $\nu > 2$, then, for any $x \in \mathrm{dom}(f)$, we have $W^0(x; 1) \subseteq \mathrm{dom}(f)$. For any $x, y \in \mathrm{dom}(f)$, let $d_\nu(x, y)$ be defined by* (12)*, and $\bar{\bar{\omega}}_\nu(\cdot)$ be defined by* (13)*. Then, we have*

$$\bar{\bar{\omega}}_\nu\left(-d_\nu(x, y)\right)^{\frac{1}{2}} \|y - x\|_x \leq \|y - x\|_y \leq \bar{\bar{\omega}}_\nu\left(d_\nu(x, y)\right)^{\frac{1}{2}} \|y - x\|_x. \tag{14}$$

If $\nu > 2$, then the right-hand side inequality of (14) *holds if $d_\nu(x, y) < 1$.*

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

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

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

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

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

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

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

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

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

given that $d_\nu(x, y) < 1$. Taking the power of $\frac{1}{\nu-2} > 0$ in both sides, we get (14) for the case $\nu > 2$.

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

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

Clearly, it is easy to show that $\phi'(t) = \frac{\langle \nabla^3 f(x+tu)[u]u,u\rangle}{\langle \nabla^2 f(x+tu)u,u\rangle} = \frac{\langle \nabla^3 f(x+tu)[u]u,u\rangle}{\|u\|_{x+tu}^2}$. Using again Definition 2 with $u = v$ and $x + tu$ instead of x, we obtain $\left|\phi'(t)\right| \leq M_f \|u\|_2$.

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

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

Substituting $u = y - x$ into this inequality, we get $\left|\ln\|y-x\|_y - \ln\|y-x\|_x\right| \leq \frac{M_f}{2}\|y-x\|_2$. Hence, $\ln\|y-x\|_x - \frac{M_f}{2}\|y-x\|_2 \leq \ln\|y-x\|_y \leq \ln\|y-x\|_x + \frac{M_f}{2}\|y-x\|_2$. This inequality leads to (14) for the case $\nu = 2$. □

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

Proposition 8 (Bounds of Hessian map) *For any $x, y \in \mathrm{dom}(f)$, let $d_\nu(x, y)$ be defined by (12), and $\bar{\bar{\omega}}_\nu(\cdot)$ be defined by (13). Then, we have*

$$[1 - d_\nu(x,y)]^{\frac{2}{\nu-2}}\,\nabla^2 f(x) \preceq \nabla^2 f(y) \preceq [1 - d_\nu(x,y)]^{\frac{-2}{\nu-2}}\,\nabla^2 f(x) \qquad \textit{if}\, \nu > 2, \qquad (15)$$

$$e^{-d_\nu(x,y)}\nabla^2 f(x) \preceq \nabla^2 f(y) \preceq e^{d_\nu(x,y)}\nabla^2 f(x) \qquad \textit{if}\ \nu = 2, \qquad (16)$$

where (15) holds if $d_\nu(x, y) < 1$ for the case $\nu > 2$.

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

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

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

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

which implies

$$\left|\frac{d\ln\psi(t)}{dt}\right| \leq M_f\left[\frac{\|y_t-x\|_{y_t}}{t}\right]^{\nu-2}\|y-x\|_2^{3-\nu}. \qquad (17)$$

Assume that $d_\nu(x, y) < 1$. Then, by the definition of y_t and $d_\nu(\cdot)$, we have $d_\nu(x, y_t) = td_\nu(x, y)$ and $\|y_t - x\|_x = t\,\|y - x\|_x$. Using Proposition 7, we can derive

$$\begin{aligned}\tfrac{1}{t}\|y_t - x\|_{y_t} &\leq \tfrac{1}{t}\left[1 - \left(\tfrac{\nu}{2} - 1\right)\|y_t - x\|_2^{3-\nu}\|y_t - x\|_x^{\nu-2}\right]^{-\frac{1}{\nu-2}}\|y_t - x\|_x \\ &= \tfrac{1}{t}\left[1 - d_\nu(x, y_t)\right]^{-\frac{1}{\nu-2}}\|y_t - x\|_x \\ &= \left[1 - d_\nu(x, y)t\right]^{-\frac{1}{\nu-2}}\|y - x\|_x .\end{aligned}$$

Hence, we can further derive

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

Integrating $\frac{d\ln\psi(t)}{dt}$ with respect to t on $[0, 1]$ and using the last inequality and (17), we get

$$\begin{aligned}\left|\int_0^1 \frac{d\ln\psi(t)}{dt}dt\right| &\leq \int_0^1 \left|\frac{d\ln\psi(t)}{dt}\right|dt\\ &\leq \|y-x\|_x^{\nu-2}\,\|y-x\|_2^{3-\nu}\int_0^1 \frac{dt}{1-d_\nu(x,y)t}.\end{aligned}$$

Clearly, we can compute this integral explicitly as

$$\begin{aligned}\left|\ln\left[\frac{\|u\|_y^2}{\|u\|_x^2}\right]\right| = \left|\ln\left[\frac{\psi(1)}{\psi(0)}\right]\right| &\leq \frac{-2d_\nu(x,y)}{(\nu-2)d_\nu(x,y)}\ln\left[1-d_\nu(x,y)\right]\\ &= \ln\left[(1-d_\nu(x,y))^{\frac{-2}{\nu-2}}\right].\end{aligned}$$

Rearranging this inequality, we obtain

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

Since this inequality holds for any $0 \neq u \in \mathbb{R}^p$, it implies (15). If $u = 0$, then (15) obviously holds.

Now, we consider the case $\nu = 2$. It follows from (17) that

$$\begin{aligned}\left|\ln\left[\frac{\|u\|_y^2}{\|u\|_x^2}\right]\right| = \left|\int_0^1 \frac{d\ln\psi(t)}{dt}dt\right| &\leq \int_0^1 \left|\frac{d\ln\psi(t)}{dt}\right|dt\\ &\leq M_f\int_0^1 \|y-x\|_2\,dt = M_f\,\|y-x\|_2\,.\end{aligned}$$

Since this inequality holds for any $u \in \mathbb{R}^p$, it implies (16). □

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

Corollary 2 *For any $x, y \in \mathrm{dom}(f)$, let $d_\nu(x,y)$ be defined by (12). Then, we have*

$$\underline{\kappa}_\nu(d_\nu(x,y))\nabla^2 f(x) \preceq \int_0^1 \nabla^2 f(x+\tau(y-x))d\tau \preceq \overline{\kappa}_\nu(d_\nu(x,y))\nabla^2 f(x), \quad (18)$$

where

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

and

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

Here, if $\nu > 2$, then it requires $d_\nu(x, y) < 1$ for $x, y \in \mathrm{dom}(f)$ in (18).

We prove a bound on the gradient inner product of a generalized self-concordant function f.

Proposition 9 (Bounds of gradient map) *For any $x, y \in \mathrm{dom}(f)$, we have*

$$\bar{\omega}_\nu\left(-d_\nu(x, y)\right)\|y-x\|_x^2 \leq \langle \nabla f(y) - \nabla f(x), y-x\rangle \leq \bar{\omega}_\nu\left(d_\nu(x, y)\right)\|y-x\|_x^2, \quad (19)$$

where, if $\nu > 2$, then the right-hand side inequality of (19) *holds if $d_\nu(x, y) < 1$, and*

$$\bar{\omega}_\nu(\tau) := \begin{cases} \frac{e^{\tau}-1}{\tau} & \textit{if } \nu = 2 \\ \frac{\ln(1-\tau)}{-\tau} & \textit{if } \nu = 4 \\ \left(\frac{\nu-2}{\nu-4}\right)\frac{1-(1-\tau)^{\frac{\nu-4}{\nu-2}}}{\tau} & \textit{otherwise}. \end{cases} \quad (20)$$

Here, $\bar{\omega}_\nu(\tau) \geq 0$ for all $\tau \in \mathrm{dom}(\bar{\omega}_\nu)$.

Proof Let $y_t := x + t(y-x)$. By the mean-value theorem, we have

$$\langle \nabla f(y) - \nabla f(x), y-x\rangle = \int_0^1 \left\langle \nabla^2 f(y_t)(y-x), y-x\right\rangle \mathrm{d}t = \int_0^1 \frac{1}{t^2}\|y_t - x\|_{y_t}^2\,\mathrm{d}t. \quad (21)$$

We consider the function $\bar{\bar{\omega}}_\nu$ defined by (13). It follows from Proposition 7 that

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

Now, we note that $d_\nu(x, y_t) = t d_\nu(x, y)$ and $\|y_t - x\|_x = t\,\|y-x\|_x$, the last estimate leads to

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

Substituting this estimate into (21), we obtain

$$\begin{aligned}\|y-x\|_x^2\int_0^1 \bar{\bar{\omega}}_\nu\left(-td_\nu(x,y)\right)dt &\le \langle\nabla f(y)-\nabla f(x), y-x\rangle\\ &\le \|y-x\|_x^2\int_0^1 \bar{\bar{\omega}}_\nu\left(td_\nu(x,y)\right)dt.\end{aligned}$$

Using the function $\bar{\bar{\omega}}_\nu(\tau)$ from (13) to compute the left-hand side and the right-hand side integrals, we obtain (19). □

Finally, we prove a bound on the function values of an (M_f,ν)-*generalized self-concordant* function f in the following proposition.

Proposition 10 (Bounds of function values) *For any* $x, y\in\mathrm{dom}(f)$*, we have*

$$\omega_\nu\left(-d_\nu(x,y)\right)\|y-x\|_x^2 \le f(y)-f(x)-\langle\nabla f(x), y-x\rangle \le \omega_\nu\left(d_\nu(x,y)\right)\|y-x\|_x^2, \quad (22)$$

where, if $\nu>2$*, then the right-hand side inequality of* (22) *holds if* $d_\nu(x,y)<1$*. Here,* $d_\nu(x,y)$ *is defined by* (12) *and* ω_ν *is defined by*

$$\omega_\nu(\tau) := \begin{cases}\frac{e^\tau-\tau-1}{\tau^2} & \text{if } \nu=2\\ \frac{-\tau-\ln(1-\tau)}{\tau^2} & \text{if } \nu=3\\ \frac{(1-\tau)\ln(1-\tau)+\tau}{\tau^2} & \text{if } \nu=4\\ \left(\frac{\nu-2}{4-\nu}\right)\frac{1}{\tau}\left[\frac{\nu-2}{2(3-\nu)\tau}\left((1-\tau)^{\frac{2(3-\nu)}{2-\nu}}-1\right)-1\right] & \text{otherwise.}\end{cases} \quad (23)$$

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

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

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

Now, by Proposition 9, we have

$$\bar{\omega}_\nu\left(-d_\nu(x,y_t)\right)\|y_t-x\|_x^2 \le \langle\nabla f(y_t)-\nabla f(x), y_t-x\rangle \le \bar{\omega}_\nu\left(d_\nu(x,y_t)\right)\|y_t-x\|_x^2.$$

Clearly, by the definition (12), we have $d_\nu(x,y_t)=td_\nu(x,y)$ and $\|y_t-x\|_x = t\|y-x\|_x$. Combining these relations, and the above two inequalities, we can show that

$$\begin{aligned}\|y-x\|_x^2\int_0^1 t\bar{\omega}_\nu\left(-td_\nu(x,y)\right)dt &\le f(y)-f(x)-\langle\nabla f(x), y-x\rangle\\ &\le \|y-x\|_x^2\int_0^1 t\bar{\omega}_\nu\left(td_\nu(x,y)\right)dt.\end{aligned}$$

By integrating the left-hand side and the right-hand side of this estimate using the definition (20) of $\bar{\omega}_\nu(\tau)$, we obtain (22). □

3 Generalized self-concordant minimization

We apply the theory developed in the previous sections to design new Newton-type methods to minimize a generalized self-concordant function. More precisely, we consider the following non-composite convex problem:

$$f^\star := \min_{x \in \mathbb{R}^p} f(x), \tag{24}$$

where $f : \mathbb{R}^p \to \mathbb{R}$ is an (M_f, ν)-*generalized self-concordant* function in the sense of Definition 2 with $\nu \in [2, 3]$ and $M_f \geq 0$. Since f is smooth and convex, the optimality condition $\nabla f(x^\star_f) = 0$ is necessary and sufficient for $x^\star_f$ to be an optimal solution of (24).

The following theorem shows the existence and uniqueness of the solution $x^\star_f$ of (24). It can be considered as a special case of Theorem 4 below with $g \equiv 0$.

Theorem 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 $\sigma_{\min}(x) := \lambda_{\min}(\nabla^2 f(x))$ and $\lambda(x) := \|\nabla f(x)\|^*_x$ for $x \in \mathrm{dom}(f)$. Suppose further that there exists $x \in \mathrm{dom}(f)$ such that $\sigma_{\min}(x) > 0$ and*

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

Then, problem (24) has a unique solution $x^\star_f$ in $\mathrm{dom}(f)$.

We say that the unique solution $x^\star_f$ of (24) is *strongly regular* if $\nabla^2 f(x^\star_f) \succ 0$. The strong regularity of $x^\star_f$ for (24) is equivalent to the strong second order optimality condition. Theorem 1 covers [40, Theorem 4.1.11] for standard self-concordant functions as a special case.

We consider the following Newton-type scheme to solve (24). Starting from an arbitrary initial point $x^0 \in \mathrm{dom}(f)$, we generate a sequence $\{x^k\}_{k\geq 0}$ as follows:

$$x^{k+1} := x^k + \tau_k n^k_{\mathrm{nt}}, \quad \text{where } n^k_{\mathrm{nt}} := -\nabla^2 f(x^k)^{-1}\nabla f(x^k), \tag{25}$$

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

- If $\tau_k = 1$ for all $k \geq 0$, then we call (25) a *full-step* Newton scheme.
- Otherwise, i.e., $\tau_k \in (0, 1)$, we call (25) a *damped-step* Newton scheme.

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

$$\nabla^2 f(x^k) n^k_{\mathrm{nt}} = -\nabla f(x^k). \tag{26}$$

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

$$\lambda_k := \|n_{\mathrm{nt}}^k\|_{x^k} = \|\nabla f(x^k)\|_{x^k}^* \quad \text{and} \quad \beta_k := M_f \|n_{\mathrm{nt}}^k\|_2 = M_f \|\nabla^2 f(x^k)^{-1} \nabla f(x^k)\|_2. \tag{27}$$

With λ_k and β_k given by (27), we also define

$$d_k := \begin{cases} \beta_k & \text{if } \nu = 2 \\ \left(\frac{\nu}{2} - 1\right) M_f^{\nu-2} \lambda_k^{\nu-2} \beta_k^{3-\nu} & \text{if } \nu \in (2, 3]. \end{cases} \tag{28}$$

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.5.

Theorem 2 *Let $\{x^k\}$ be the sequence generated by the damped-step Newton scheme (25) with the following step-size:*

$$\tau_k := \begin{cases} \frac{1}{\beta_k} \ln(1 + \beta_k) & \textit{if } \nu = 2 \\ \frac{1}{d_k} \left[1 - \left(1 + \frac{4-\nu}{\nu-2} d_k\right)^{-\frac{\nu-2}{4-\nu}}\right] & \textit{if } \nu \in (2, 3], \end{cases} \tag{29}$$

where λ_k, β_k are defined by (27), and d_k is defined by (28). Then, $\tau_k \in (0, 1]$, $\{x^k\}$ in $\mathrm{dom}(f)$*, and this step-size guarantees the following descent property*

$$f(x^{k+1}) \le f(x^k) - \Delta_k, \tag{30}$$

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 (23).

Assume that the unique solution $x_f^\star$ of (24) exists. Then, there exists a neighborhood $\mathcal{N}(x_f^\star)$ such that if we initialize the Newton scheme (25) at $x^0 \in \mathcal{N}(x_f^\star) \cap \mathrm{dom}(f)$, then the whole sequence $\{x^k\}$ converges to $x_f^\star$ at a quadratic rate.

Example 4 (Better step-size for regularized logistic and exponential models) Consider the minimization problem (24) with the objective function $f(\cdot) := \phi(\cdot) + \frac{\gamma}{2}\|\cdot\|_2^2$, where ϕ is defined as in (8) with $\varphi_i(t) = \log(1 + e^{-t})$ being the logistic loss. That is

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

As we shown in Sect. 2 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 $R_A := \max\{\|a_i\|_2 \mid 1 \le i \le n\}$. Then, if we consider $\nu = 2$, then we have $M_f^{(2)} = R_A$ due to Corollary 1, while if we choose $\nu = 3$, then $M_f^{(3)} = \frac{1}{\sqrt{\gamma}} R_A$ due to Proposition 4. By the definition of f, we have $\nabla^2 f(x) \succeq \gamma \mathbb{I}$. Hence, using this inequality and the definition of λ_k and β_k from (27), we can show that

$$\beta_k = M_f^{(2)} \|\nabla^2 f(x^k)^{-1} \nabla f(x^k)\|_2 \le \frac{R_A}{\sqrt{\gamma}} \lambda_k = M_f^{(3)} \lambda_k. \tag{31}$$

For any $\tau > 0$, we have $\frac{\ln(1+\tau)}{\tau} > \frac{1}{1+0.5\tau}$. Using this elementary result and (31), we obtain

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

This inequality shows that the step-size τ_k given by Theorem 2 satisfies $\tau_k^{(2)} > \tau_k^{(3)}$, where $\tau_k^{(\nu)}$ is a given step-size computed by (29) 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 be empirically confirmed by our experiments in Sect. 6. □

Next, we study the full-step Newton scheme derived from (25) by setting the step-size $\tau_k = 1$ for all $k \geq 0$ as a full-step. Let

$$\underline{\sigma}_k := \lambda_{\min}\left(\nabla^2 f(x^k)\right)$$

be the smallest eigenvalue of $\nabla^2 f(x^k)$. Since $\nabla^2 f(x^k) \succ 0$, we have $\underline{\sigma}_k > 0$. The following theorem shows a local quadratic convergence of the full-step Newton scheme (25) for solving (24) whose proof can be found in Appendix A.6.

Theorem 3 *Let $\left\{x^k\right\}$ be the sequence generated by the full-step Newton scheme (25) by setting the step-size $\tau_k = 1$ for $k \geq 0$. Let $d_\nu^k := d_\nu(x^k, x^{k+1})$ be defined by (12) and λ_k be defined by (27). Then, the following statements hold:*

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

(b) *If $2 < \nu < 3$, and the starting point x^0 satisfies $\underline{\sigma}_0^{-\frac{3-\nu}{2}}\lambda_0 < \frac{1}{M_f}\min\left\{\frac{2d_\nu^\star}{\nu-2}, \frac{1}{2}\right\}$, then both sequences $\left\{\underline{\sigma}_k^{-\frac{3-\nu}{2}}\lambda_k\right\}$ and $\left\{d_\nu^k\right\}$ decrease and quadratically converge to zero, where $d_\nu^\star$ is the unique solution of the equation $(\nu-2)\,R_\nu(d_\nu) = 4(1-d_\nu)^{\frac{4-\nu}{\nu-2}}$ in d_ν with $R_\nu(\cdot)$ given by (56).*

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

As a consequence, if $\left\{d_\nu^k\right\}$ locally converges to zero at a quadratic rate, then $\left\{\|x^k - x_f^\star\|_{H_k}\right\}$ also locally converges to zero at a quadratic rate, where $H_k = \mathbb{I}$, the identity matrix, if $\nu = 2$; $H_k = \nabla^2 f(x^k)$ if $\nu = 3$; and $H_k = \nabla^2 f(x^k)^{\frac{\nu}{2}-1}$ if $2 < \nu < 3$. Hence, $\left\{x^k\right\}$ locally converges to $x_f^\star$, the unique solution of (24), at a quadratic rate.

If we combine the results of Theorem 2 and Theorem 3, then we can design a two-phase Newton algorithm for solving (24) as follows:

- *Phase 1* Starting from an arbitrary initial point $x^0 \in \mathrm{dom}(f)$, we perform the damped-step Newton scheme (25) until the condition in Theorem 3 is satisfied.

- *Phase 2* Using the output x^j of *Phase 1* as an initial point for the full-step Newton scheme (25) with $\tau_k = 1$, and perform this scheme until it achieves an ε-solution x^k to (24).

We also note that the damped-step Newton scheme (25) can also achieve a local quadratic convergence as shown in Theorem 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 $x^0 \in \mathrm{dom}(f)$ and a desired accuracy $\varepsilon > 0$.
**Output:** An $\varepsilon$-solution $x^k$ of (24).
**Initialization:** Compute $d_\nu^\star$ according to Theorem 3 if needed.
**For** $k = 0, \cdots, k_{\max}$, **perform:**
  Compute the Newton direction $n_{\mathrm{nt}}^k$ by solving $\nabla^2 f(x^k) n_{\mathrm{nt}}^k = -\nabla f(x^k)$.
  Compute $\lambda_k := \|n_{\mathrm{nt}}^k\|_{x^k}$, and compute $\beta_k := M_f \|n_{\mathrm{nt}}^k\|_2$ if $\nu \neq 3$.
  If $\lambda_k \leq \varepsilon$, then TERMINATE and return $x^k$.
  If *Phase 2 is used*, then compute $\underline{\sigma}_k = \lambda_{\min}(\nabla^2 f(x^k))$ if $2 \leq \nu < 3$.
  If *Phase 2 is used* and $(\lambda_k, \underline{\sigma}_k)$ satisfies Theorem 3, then set $\tau_k := 1$ (**full-step**).
  Otherwise, compute the step-size $\tau_k$ by (29) (**damped-step**)
  Update $x^{k+1} := x^k + \tau_k n_{\mathrm{nt}}^k$.
**End for**
```

Per-iteration complexity The main step of Algorithm 1 is the solution of the symmetric positive definite linear system (26). This system can be solved by using either Cholesky factorization or conjugate gradient methods which, in the worst-case, requires $\mathcal{O}(p^3)$ operations. Computing λ_k requires the inner product $\langle n_{\mathrm{nt}}^k, \nabla f(x^k)\rangle$ which needs $\mathcal{O}(p)$ operations.

Conceptually, the two-phase option of Algorithm 1 requires the smallest eigenvalue of $\nabla^2 f(x^k)$ 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 (29) 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 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 3 (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 (26). 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 [16], but defined via the local dual norm of f:

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

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

4 Composite generalized self-concordant minimization

Let $f \in \widetilde{\mathcal{F}}_{M_f,\nu}(\mathrm{dom}(f))$, and g be a proper, closed, and convex function. We consider the composite convex minimization problem (3) which we recall here for our convenience of references:

$$F^\star := \min_{x\in\mathbb{R}^p} \Big\{ F(x) := f(x) + g(x) \Big\}. \tag{32}$$

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 (32) can be written as follows:

$$0 \in \nabla f(x^\star) + \partial g(x^\star). \tag{33}$$

Under the qualification condition $0 \in \mathrm{ri}\,(\mathrm{dom}(g) - \mathrm{dom}(f))$, (33) is necessary and sufficient for $x^\star$ to be an optimal solution of (32), where $\mathrm{ri}\,(\mathcal{X})$ is the relative interior of $\mathcal{X}$.

4.1 Existence, uniqueness, and regularity of optimal solutions

Assume that $\nabla^2 f(x)$ is positive definite (i.e., nonsingular) at some point $x \in \mathrm{dom}(F)$. We prove in the following theorem that problem (32) has a unique solution $x^\star$. The proof can be found in Appendix A.4. This theorem can also be considered as a generalization of [40, Theorem 4.1.11] and [62, Lemma 4] in standard self-concordant settings in [40,62].

Theorem 4 *Suppose that the function f of (32) is (M_f, ν)-generalized self-concordant with $M_f > 0$ and $\nu \in [2, 3]$. Denote by $\sigma_{\min}(x) := \lambda_{\min}(\nabla^2 f(x))$ and $\lambda(x) := \|\nabla f(x) + v\|_x^*$ for $x \in \mathrm{dom}(F)$ and $v \in \partial g(x)$. Suppose further that there exists $x \in \mathrm{dom}(F)$ such that $\sigma_{\min}(x) > 0$ and*

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

Then, problem (32) has a unique solution $x^\star$ in $\mathrm{dom}(F)$.

Now, we recall a condition such that the solution $x^\star$ of (32) is strongly regular in the following Robinson's sense [56]. We say that the optimal solution $x^\star$ of (32) 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

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

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

If $\nabla^2 f(x^\star)\succ 0$, then $x^\star$ is strongly regular. While the strong regularity of the solution $x^\star$ requires a weaker condition than $\nabla^2 f(x^\star)\succ 0$. For further details of the regularity theory, we refer the reader to [56].

4.2 Scaled proximal operators

Given a matrix $H\in\mathcal{S}^p_{++}$, we define a scaled proximal operator of g in (32) as

$$\mathrm{prox}_{H^{-1}g}(x) := \mathrm{argmin}_{z}\left\{g(z)+\tfrac{1}{2}\,\|z-x\|_H^2\right\}. \tag{34}$$

Using the optimality condition of the minimization problem under (34), we can show that

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

Since g is proper, closed, and convex, $\mathrm{prox}_{H^{-1}g}$ is well-defined and single-valued. In particular, if we take $H=\mathbb{I}$, the identity matrix, then $\mathrm{prox}_{H^{-1}g}(\cdot)=\mathrm{prox}_g(\cdot)$, the standard proximal operator of g. If we can efficiently compute $\mathrm{prox}_{H^{-1}g}(\cdot)$ 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 [3,21,51].

4.3 Proximal Newton methods

The proximal Newton method can be considered as a special case of the variable metric proximal method in the literature [8]. This method has previously been studied by many authors, see, e.g., [8,34]. However, the convergence guarantee often requires certain assumptions as used in standard Newton-type methods. In this section, we develop a proximal Newton algorithm to solve the composite convex minimization problem (32) where f is a generalized self-concordant function. This problem covers [62,64] as special cases.

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

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

Next, the main step of the proximal Newton method requires to solve the following subproblem:

$$z^k := \operatorname*{argmin}_{x \in \mathrm{dom}(g)} \left\{ Q_f(x; x^k) + g(x) \right\} = \mathrm{prox}_{\nabla^2 f(x^k)^{-1} g}\left(x^k - \nabla^2 f(x^k)^{-1} \nabla f(x^k) \right). \quad (35)$$

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

$$0 \in \nabla f(x^k) + \nabla^2 f(x^k)(z^k - x^k) + \partial g(z^k). \quad (36)$$

Here, we note that $\mathrm{dom}(Q_f(\cdot; x^k)) = \mathbb{R}^p$. Hence, $\mathrm{dom}(Q_f(\cdot; x^k) + g(\cdot)) = \mathrm{dom}(g)$. In the setting (32), z^k may not be in $\mathrm{dom}(F)$. Our next step is to update the next iteration x^{k+1} as

$$x^{k+1} := x^k + \tau_k n^k_{\mathrm{pnt}} = (1 - \tau_k) x^k + \tau_k z^k, \quad (37)$$

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

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

$$\lambda_k := \|n^k_{\mathrm{pnt}}\|_{x^k} \quad \text{and} \quad \beta_k := M_f \|n^k_{\mathrm{pnt}}\|_2. \quad (38)$$

Our first goal is to show that we can explicitly compute the step-size τ_k in (37) 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.7.

Theorem 5 *Let $\left\{x^k\right\}$ be the sequence generated by the proximal Newton scheme (37) starting from $x^0 \in \mathrm{dom}(F)$. If we choose the step-size τ_k as in (29) of Theorem 2, then $\tau_k \in (0, 1]$, $\left\{x^k\right\}$ in $\mathrm{dom}(F)$, and*

$$F(x^{k+1}) \le F(x^k) - \Delta_k, \quad (39)$$

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 2.

There exists a neighborhood $\mathcal{N}(x^\star)$ of the unique solution $x^\star$ of (32) such that if we initialize the scheme (37) at $x^0 \in \mathcal{N}(x^\star) \cap \mathrm{dom}(F)$, then $\left\{x^k\right\}$ quadratically converges to $x^\star$.

Next, we prove a local quadratic convergence of the full-step proximal Newton method (37) with the unit step-size $\tau_k = 1$ for all $k \ge 0$. The proof is given in Appendix A.8.

Theorem 6 *Suppose that the sequence $\left\{x^k\right\}$ is generated by (37) with full-step, i.e., $\tau_k = 1$ for $k \geq 0$. Let $d_\nu^k := d_\nu(x^k, x^{k+1})$ be defined by (12) and λ_k be defined by (38). Then, the following statements hold:*

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

(b) *If $2 < \nu < 3$, and the starting point x^0 satisfies $\underline{\sigma}_0^{-\frac{3-\nu}{2}}\lambda_0 < \frac{1}{M_f}\min\left\{\frac{2d_\nu^\star}{\nu-2}, \frac{1}{2}\right\}$, then both sequences $\left\{\underline{\sigma}_k^{-\frac{3-\nu}{2}}\lambda_k\right\}$ and $\left\{d_\nu^k\right\}$ decrease and quadratically converge to zero, where $d_\nu^\star$ is the unique solution to the equation $(\nu - 2)\, R_\nu(d_\nu) = 4(1 - d_\nu)^{\frac{4-\nu}{\nu-2}}$ in d_ν with $R_\nu(\cdot)$ given in (56).*

(c) *If $\nu = 3$ and the starting point x^0 satisfies $\lambda_0 < \frac{2d_3^\star}{M_f}$, then the sequence $\{\lambda_k\}$ decreases and quadratically converges to zero, where $d_3^\star \approx 0.20943$.*

As a consequence, if $\left\{d_\nu^k\right\}$ locally converges to zero at a quadratic rate, then $\left\{\|x^k - x^\star\|_{H_k}\right\}$ also locally converges to zero at a quadratic rate, where $H_k = \mathbb{I}$, the identity matrix, if $\nu = 2$; $H_k = \nabla^2 f(x^k)$ if $\nu = 3$; and $H_k = \nabla^2 f(x^k)^{\frac{\nu}{2}-1}$ if $2 < \nu < 3$. Hence, $\left\{x^k\right\}$ locally converges to $x^\star$, the unique solution of (32), at a quadratic rate.

Similar to Algorithm 1, we can also combine the results of Theorems 5 and 6 to design a proximal Newton algorithm for solving (32). 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 $x^0 \in \mathrm{dom}(F)$ and a desired accuracy $\varepsilon > 0$.
Output: An $\varepsilon$-solution $x^k$ of (32).
Initialization: Compute $d_\nu^\star$ according to Theorem 6 if needed.
For $k = 0, \cdots, k_{\max}$, perform:
  Compute the proximal Newton direction $n_{\mathrm{pnt}}^k$ by solving (35).
  Compute $\lambda_k := \|n_{\mathrm{pnt}}^k\|_{x^k}$, and compute $\beta_k := M_f\|n_{\mathrm{pnt}}^k\|_2$ if $\nu \neq 3$.
  If $\lambda_k \leq \varepsilon$, then TERMINATE.
  If Phase 2 is used, then compute $\underline{\sigma}_k = \lambda_{\min}(\nabla^2 f(x^k))$ if $2 \leq \nu < 3$.
  If Phase 2 is used and $(\lambda_k, \underline{\sigma}_k)$ satisfies Theorem 6, then set $\tau_k := 1$ (full-step).
  Otherwise, compute the step-size $\tau_k$ by (29) (damped-step).
  Update $x^{k+1} := x^k + \tau_k n_{\mathrm{pnt}}^k$.
End for
```

Implementation remarks The main step of Algorithm 2 is the computation of the proximal Newton step n^k_{pnt}, or the trial point z^k in (35). This step requires to solve a composite quadratic-convex minimization problem (35) with strongly convex objective function. If g is proximally tractable, then we can apply proximal-gradient methods or splitting techniques [3,4,44] to solve this problem. We can also combine accelerated proximal-gradient methods with a restarting strategy [19,23,48] to accelerate the performance of these algorithms. These methods will be used in our numerical experiments in Sect. 6.

As noticed in Remark 3, we can also develop an inexact proximal Newton variant for Algorithm 2 by approximately solving the subproblem (35). We leave this extension to our forthcoming work.

5 Quasi-Newton methods for generalized self-concordant minimization

This section studies quasi-Newton variants of Algorithm 1 for solving (24). Extensions to the composite form (32) can be done by combining the result in this section and the approach in [62].

A quasi-Newton method for solving (24) updates the sequence $\{x^k\}$ using

$$x^{k+1} := x^k - \tau_k B_k \nabla f(x^k), \quad \text{where } B_k := H_k^{-1} \text{ and } H_k \approx \nabla^2 f(x^k), \tag{40}$$

where the step-size $\tau_k \in (0, 1]$ is appropriately chosen, and $x^0 \in \mathrm{dom}(f)$ is a given starting point.

Matrix H_k is symmetric and positive definite, and it approximates the Hessian matrix $\nabla^2 f(x^k)$ of f at the iteration x^k in some sense. The most common approximation sense is that H_k satisfies the well-known Dennis–Moré condition [15]. In the context of generalized self-concordant functions, we can modify this condition by imposing:

$$\lim_{k\to\infty} \frac{\|(H_k - \nabla^2 f(x^\star_f))(x^k - x^\star_f)\|^*_{\hat{x}}}{\|x^k - x^\star_f\|_{\hat{x}}} = 0, \quad \text{where } \hat{x} = x^\star_f \text{ or } \hat{x} = x^k. \tag{41}$$

Clearly, if we have $\lim_{k\to\infty} \|H_k - \nabla^2 f(x^k)\|_{\hat{x}} = 0$, then, with a simple argument, we can show that (41) automatically holds. In practice, we can update H_k to maintain the following *secant equation*:

$$H_{k+1}s^k = y^k, \quad \text{where } s^k := x^{k+1} - x^k, \text{ and } y^k := \nabla f(x^{k+1}) - \nabla f(x^k). \tag{42}$$

There are several candidates to update H_k to maintain this secant equation, see, e.g., [47]. Here, we propose to use a BFGS update as

$$H_{k+1} := H_k + \frac{y^k(y^k)^\top}{\langle y^k, s^k\rangle} - \frac{(H_k s^k)(H_k s^k)^\top}{(\langle H_k s^k, s^k\rangle}. \tag{43}$$

In practice, to avoid the inverse $B_k = H_k^{-1}$, we can update this inverse directly [47] in lieu of updating H_k as in (43). Note that the BFGS update (43) or its inverse version may not maintain the sparsity or block pattern structures of the sequence $\{H_k\}$ or $\{B_k\}$ even if $\nabla^2 f$ is sparse.

The following result shows that the quasi-Newton method (40) achieves a superlinear convergence whose proof can be found in Appendix A.9.

Theorem 7 *Assume that $x^\star_f \in \mathrm{dom}(f)$ is the unique solution of (24) and is strongly regular. Let $\left\{x^k\right\}$ be the sequence generated by (40). Then, the following statements hold:*

(a) *Assume, in addition, that the sequence of matrices $\{H_k\}$ satisfies the Dennis–Moré condtion (41) with $\hat{x} = x^\star_f$. Then, there exist $\bar{r} > 0$, and $\bar{k} \geq 0$ such that, for all $k \geq \bar{k}$, we have $\|x^k - x^\star_f\|_{x^\star_f} \leq \bar{r}$ and $\left\{x^k\right\}$ locally converges to $x^\star_f$ at a superlinear rate.*

(b) *Suppose that H_0 is chosen such that $H_0 \in \mathcal{S}^p_{++}$. Then, $\langle y^k, z^k\rangle > 0$ for all $k \geq 0$, and hence, the sequence $\{H_k\}$ generated by (43) is symmetric positive definite, and satisfies the secant equation (42). Moreover, if the sequence $\left\{x^k\right\}$ generated by (40) satisfies $\sum_{k=0}^{\infty} \|x^k - x^\star_f\|_{x^\star_f} < +\infty$, then $\left\{x^k\right\}$ locally converges to the unique solution $x^\star_f$ of (24) at a superlinear rate.*

Note that the condition $\sum_{k=0}^{\infty} \|x^k - x^\star_f\|_{x^\star_f} < +\infty$ in Theorem 7(b) can be guaranteed if $\|x^{k+1} - x^\star_f\|_{x^\star_f} \leq \rho \|x^k - x^\star_f\|_{x^\star_f}$ for some $\rho \in (0, 1)$ and $k \geq \bar{k} \geq 0$. Hence, if $\left\{x^k\right\}$ locally converges to $x^\star_f$ at a linear rate, then it also locally converges to $x^\star_f$ at a superlinear rate.

6 Numerical experiments

We provide five 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.

6.1 Comparison with [72] on regularized logistic regression

In this example, we empirically show that our theory provides a better step-size for logistic regression compared to [72] as theoretically shown in Example 4. 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:

$$f^\star := \min_{x \in \mathbb{R}^p} \left\{ f(x) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i(a_i^\top x + \mu)) + \frac{\gamma}{2}\|x\|_2^2 \right\}, \tag{44}$$

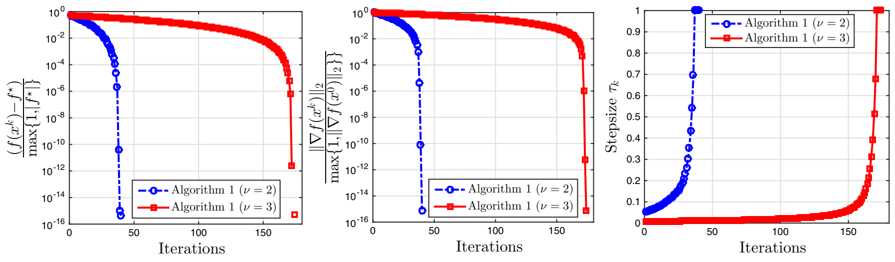

Fig. 1 The convergence of Algorithm 1 for `news20.binary` (left: relative objective residuals, middle: relative norms of gradient, and right: step-sizes)

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

As shown previously in Proposition 5, f can be cast into an $(M_f^{(3)}, 3)$-generalized self-concordant function with $M_f^{(3)} = \frac{1}{\sqrt{\gamma}} \max\{\|a_i\|_2 \mid 1 \le i \le n\}$. On the other hand, f can also be considered as an $(M_f^{(2)}, 2)$-generalized self-concordant with $M_f^{(2)} := \max\{\|a_i\|_2 \mid 1 \le i \le n\}$.

We implement Algorithm 1 using two different step-sizes $\tau_k^{(2)} = \frac{\ln(1+\beta_k)}{\beta_k}$ and $\tau_k^{(3)} := \frac{1}{1+0.5M_f^{(3)}\lambda_k}$ as suggested by Theorem 2 for $\nu = 2$ and $\nu = 3$, respectively. We terminate Algorithm 1 if $\|\nabla f(x^k)\|_2 \le 10^{-8} \max\left\{1, \|\nabla f(x^0)\|_2\right\}$, where $x^0 = \mathbf{0}$ is an initial point. To solve the linear system (26), we apply a conjugate gradient method to avoid computing the inverse $\nabla^2 f(x^k)^{-1}$ of the Hessian matrix $\nabla^2 f(x^k)$ in large-scale problems. We also compare our algorithms with the fast gradient method in [40] using an optimal step-size for strongly convex functions which has an optimal linear convergence rate.

We test all algorithms on a binary classification dataset downloaded from [12] at https://www.csie.ntu.edu.tw/~cjlin/libsvm/. As suggested in [72], we normalize the data such that each row a_i has $\|a_i\|_2 = 1$ for $i = 1, \ldots, n$. The parameter is set to $\gamma := 10^{-5}$ as in [72].

The convergence behavior of Algorithm 1 for $\nu = 2$ and $\nu = 3$ is plotted in Figure 1 for the `news20` problem.

As we can see from this figure that Algorithm 1 with $\nu = 2$ outperforms the case $\nu = 3$. The right-most plot reveals the relative objective residual $\frac{f(x^k)-f^\star}{\max\{1,|f^\star|\}}$, the middle one shows the relative gradient norm $\frac{\|\nabla f(x^k)\|_2}{\max\{1,\|\nabla f(x^0)\|_2\}}$, 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 3 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 [40]. Here, n is the number of data points,

Table 3 The performance and results of the three algorithms for solving the logistic regression problem (44)

Problem			Algorithm 1 ($\nu = 2$)				Algorithm 1 ($\nu = 3$)				Fast gradient method [40]			
Name	p	n	iter	time[s]	$f(x^k)$	error	iter	time[s]	$f(x^k)$	error	iter	time[s]	$f(x^k)$	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

p is the number of variables, `iter` is the number of iterations, `error` is the training error measured by $\frac{1}{2n}\sum_{i=1}^{n}(1-\mathrm{sign}(y_i(a_i^\top x+\mu)))$, and $f(x^k)$ is the objective value achieved by these three algorithms.

We observe that our step-size $\tau_k^{(2)}$ using $\nu=2$ works much better than $\tau_k^{(3)}$ using $\nu=3$ as in [72]. This confirms the theoretical analysis in Example 4. This step-size can be 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 [40] 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

$$f(x^k+\tau_k n_{\mathrm{nt}}^k)\le f(x^k)-c_1\tau_k\nabla f(x^k)^\top n_{\mathrm{nt}}^k, \tag{45}$$

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.

As shown in Table 4, 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 x^0 as well as the factor c_1 in (45). If 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 $c_1\tau_k\nabla f(x^k)^\top n_{\mathrm{nt}}^k$ may never be achieved, and the linesearch condition fails to hold. If we change the starting point x^0, the number of function evaluations can significantly be increased.

6.2 The case $\nu=2$: matrix balancing

We consider the following convex optimization problem originated from matrix balancing [14]:

$$f^\star:=\min_{x\in\mathbb{R}^p}\Big\{f(x):=\sum_{1\le i,j\le p}a_{ij}e^{x_i-x_j}\Big\}, \tag{46}$$

where $A=(a_{ij})_{p\times p}$ is a nonnegative square matrix in $\mathbb{R}^{p\times p}$. Although (46) is a unconstrained smooth convex 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, (46) is an important problem in scientific computing.

Table 4 The performance and results of the two linesearch variants of Algorithm 1 for solving (44)

Problem			Algorithm 1 (Standard linesearch)						Algorithm 1 (Linesearch with $\tau_k^{(2)}$)					
Name	p	n	iter	nfval	time[s]	$\frac{\Vert\nabla f(x^k)\Vert_2}{\Vert\nabla f(x^0)\Vert_2}$	$f(x^k)$	error	iter	nfval	time[s]	$\frac{\Vert\nabla f(x^k)\Vert_2}{\Vert\nabla f(x^0)\Vert_2}$	$f(x^k)$	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

By Proposition 1 and Corollary 1, f is generalized self-concordant with $M = \sqrt{2}$ and $\nu = 2$. We implement Algorithm 1 and the most recent method proposed in [14] (called Boxed-constrained Newton method (BCNM)) to solve (46). Note that [14] is not directly applicable to (46), but it solves a regularization of this problem. Since $\nabla^2 f(x)$ is not positive definite, we use a preconditioned conjugate gradient gradient (PCG) 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 [14]. We terminate PCG and FISTA using either a tolerance 10^{-9} or a maximum of 200 iterations. For the outer loop, we terminate Algorithm 1 and BCNM using the same stopping criterion: $\delta f_k' := \|\nabla f(x^k)\|_2 / \max\left\{1, \|\nabla f(x^0)\|_2\right\} \le 10^{-8}$. We choose $x^0 := \mathbf{0}^p$ as an initial point.

We test both algorithms on several synthetic and real datasets. The synthetic data is generated as in [52] with different structures. The basic matrix $H = (H_{ij})_{p\times p}$ is a $p \times p$ upper Hessenberg matrix defined as $H_{ij} = 0$ if $j < i - 1$, and $H_{ij} = 1$ otherwise. H_1 differs from H only in that H_{11} is replaced by p^2; H_2 differs from H only in that H_{12} is replaced by p^2; and $H_3 = H + (p^2 - 1)\mathbb{I}_p$. We use these matrices for A in (46). 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 is downloaded from https://math.nist.gov/MatrixMarket/searchtool.html with different structures from different application fields, suggested by [13]. Since we require the matrix A to be nonnegative, we take $A_0 := \max\{0, A\}$ (entry-wise). For the real data, if A is highly ill-conditioned, then we add a uniform noise $\mathcal{U}[0, \sigma]$ to A, where $\sigma = 10^{-5}\max\{a_{ij} | 1 \le i, j \le p\}$.

The final results of both algorithms are reported in Table 5, where p is the size of matrix A; iter/siter is the maximum number of Newton-type iterations / PCG 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 δx^k is the relative difference between 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 $f(x^k)$ and approximate solutions 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 $t_{\mathrm{rat}} = \frac{\mathrm{time}_{\mathrm{BCNM}}}{\mathrm{time}_{\mathrm{Alg.\ 1}}}$. 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.

6.3 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 [36]. In order to directly use Algorithm 1, we slightly modify the setting in [36] to obtain the following form:

Table 5 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]	$f(x^k)$	$\delta f_k'$	iter/siter	time[s]	$f(x^k)$	$\delta f_k'$	t_{rat}	δx^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

Table 5 continued

Datasets		Algorithm 1				BCNM				Comparison	
Name	p	iter/siter	time[s]	$f(x^k)$	$\delta f_k'$	iter/siter	time[s]	$f(x^k)$	$\delta f_k'$	$\mathrm{t_{rat}}$	δx^k
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

$$f^{\star} := \min_{x=[w,\xi,\mu]^{\top}\in\mathbb{R}^p} \left\{ f(x) := \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(a_i^{\top}w + \mu y_i + \xi_i)^q} + c^{\top}\xi \right.$$
$$\left. + \frac{1}{2}\left(\gamma_1\|w\|_2^2 + \gamma_2\mu^2 + \gamma_3\|\xi\|_2^2\right)\right\}, \tag{47}$$

where $q > 0$, a_i, y_i ($i = 1, \dots, n$) and c are given, and $\gamma_s > 0$ ($s = 1, 2, 3$) are three regularization parameters for w, μ and ξ, respectively. Here, the variable x consists of the support vector w, the intercept μ, and the slack variable ξ as used in [36]. Here, we penalize these variables by using least-squares terms instead of the ℓ_1-penalty term as in [36]. Note that the setting (47) 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 [69] if we choose an appropriate parameter q.

Since $\varphi(t) := \frac{1}{t^q}$ is (M_φ, ν)-generalized self-concordant 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 1, we can show that f is $(M_f, \frac{2(q+3)}{q+2})$-generalized self-concordant with $M_f := \frac{q+2}{\sqrt[(q+2)]{q(q+1)}} n^{\frac{1}{q+2}} \max\left\{ \|(a_i^{\top}, y_i, e_i^{\top})^{\top}\|_2^{q/(q+2)} \mid 1 \le i \le n\right\}$ (here, e_i is the i-th unit vector). Problem (47) can be transformed into a second-order cone program [25], 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 (47) 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 (47) 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 [60] or SDPA [70] 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 (47) 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 6 small and medium size problems using data from [12] (https://www.csie.ntu.edu.tw/~cjlin/libsvm/). 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 (47) becomes $p + n + 1$. Hence, we use a built-in Matlab conjugate gradient solver to compute the Newton direction n_{nt}^k. The initial point x^0 is chosen as $w^0 := \mathbf{0}$, $\mu^0 := 0$ and $\xi^0 := \mathbf{1}$. In our algorithms, we use $\|\nabla f(x^k)\|_2 \le 10^{-8}\max\left\{1, \|\nabla f(x^0)\|_2\right\}$ as a stopping criterion.

Note that, by the choice of γ_i for $i = 1, 2, 3$ as $\gamma_{\min} := \min\{\gamma_1, \gamma_2, \gamma_3\} = 10^{-7} > 0$, the objective function of (47) is strongly convex. By Proposition 4(a), we can cast this function into an $(\hat{M}_f, \hat{\nu})$-generalized self-concordant with $\hat{\nu} = 3$ 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 (47).

The results and performance of the four algorithms are reported in Table 6 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 $\frac{\|\nabla f(x^k)\|_2}{\|\nabla f(x^0)\|_2}$ 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 [72] for the logistic case.

6.4 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 (32) with $\nu = 3$. We illustrate this algorithm on the following portfolio optimization problem with logarithmic utility functions [59] (scaled by a factor of $\frac{1}{n}$):

$$f^\star = \min_{x\in\mathbb{R}^p}\left\{ f(x) := -\sum_{i=1}^{n}\log(w_i^\top x) \mid x \geq 0,\ \ \mathbf{1}^\top x = 1\right\}, \tag{48}$$

where $w_i \in \mathbb{R}^p_+$ 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 [9], 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 (48) can be cast into an online optimization model [27]. The authors in [27] 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 (48) 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 (48).

Let $\Delta := \left\{x \in \mathbb{R}^p \mid x \geq 0,\ \ \mathbf{1}^\top x = 1\right\}$ be the standard simplex, and $g(x) := \delta_\Delta(x)$ be the indicator function of Δ. Then, we can formulate (48) into (32). The function f defined in (48) is (M_f, ν)-generalized self-concordant with $\nu = 3$ and $M_f = 2$.

We implement Algorithm 2 using an accelerated projected gradient method [4,40] to compute the proximal Newton direction. We also implement the Frank–Wolfe algorithm and its linesearch variant in [20,30], and a projected gradient method using Barzilai and Borwein's step-size to solve (48). We name these algorithms by `FW`, `FW-LS`, and `PG-BB`, respectively.

Table 6 The performance and results of the four methods for solving the DWD problem (47)

Problem			Algorithm 1			Algorithm 1 (LS with τ_k)			Algorithm 1 ($\nu = 3$)			Mosek	
Name	n	p	iter	time[s]	$\frac{\Vert\nabla f(x^k)\Vert_2}{\Vert\nabla f(x^0)\Vert_2}$	iter	time[s]	$\frac{\Vert\nabla f(x^k)\Vert_2}{\Vert\nabla f(x^0)\Vert_2}$	iter	time[s]	$\frac{\Vert\nabla f(x^k)\Vert_2}{\Vert\nabla f(x^0)\Vert_2}$	time[s]	$\frac{\Vert\nabla f(x^k)\Vert_2}{\Vert\nabla f(x^0)\Vert_2}$
$q = 1$													
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

We emphasize that both PG-BB and FW-LS do not have a theoretical guarantee when solving (48). FW has a theoretical guarantee as recently proved in [49], but the complexity bound is rather pessimistic. We terminate all the algorithms using $\|x^{k+1} - x^k\|_2 \leq \varepsilon \max\left\{1, \|x^k\|_2\right\}$, 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 W with given price ratios as described above in Matlab. More precisely, we generate $W := \mathrm{ones}(n, p) + \mathcal{N}(0, 0.1)$ 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 http://www.excelclout.com/historical-stock-prices-in-excel/. 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 [9]. We pre-processed the data by removing stocks that are empty or lacking 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 7. Here, iter gives the number of iterations, time is the computational time in second, error measures the relative difference between the approximate solution x^k given by the algorithms and the interior-point solution provided by CVX [25] with the high precision configuration (up to 1.8×10^{-12}): $\left\|x^k - x^*_{\mathrm{cvx}}\right\| / \max\left\{1, \left\|x^*_{\mathrm{cvx}}\right\|\right\}$.

From Table 7 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 a theoretical guarantee. Note that Algorithm 2 and PG-BB are faster than the FW method and its linesearch variant although the optimal solution $x^\star$ 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 $x^\star$ but gives a high approximation to $x^\star$. In fact, the CVX solution is dense. Hence, it is not clear if PG-BB produces a better solution than other methods.

6.5 Proximal Quasi-Newton method for sparse multinomial logistic regression.

We apply our proximal Newton and proximal quasi-Newton methods to solve the following sparse multinomial logistic problem studied in various papers including [31]:

$$F^\star := \min_x \left\{ F(x) := \Big[\frac{1}{n} \sum_{j=1}^{n} \Big(\log \big(\sum_{i=1}^{m} e^{\langle w^{(j)}, x^{(i)} \rangle} \big) - \sum_{i=1}^{m} y_i^{(j)} \langle w^{(j)}, x^{(i)} \rangle \Big) \Big]_{f(x)} + \Big[\gamma \|\mathrm{vec}(x)\|_1 \Big]_{g(x)} \right\}, \tag{49}$$

Table 7 The performance and results of the four algorithms for solving the portfolio optimization problem (48)

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

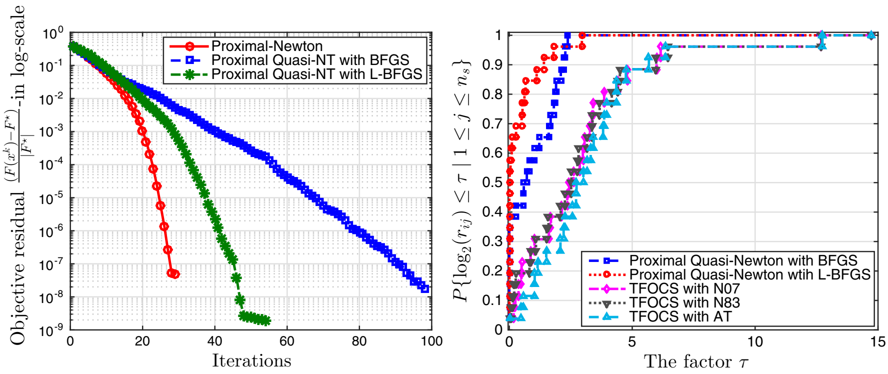

Fig. 2 Left: convergence behavior of three methods, right: performance profile in time [second] of 5 methods

where x can be considered as a matrix variable of size $m \times p$ formed from $x^{(1)}, \cdots, x^{(m)}$, $\mathrm{vec}(\cdot)$ is the vectorization operator, and $\gamma > 0$ is a regularization parameter. Both $y_i^{(j)} \in \{0, 1\}$ and $w^{(j)}$ are given as input data for $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

The function f defined in (49) has a closed form Hessian matrix. However, forming the full Hessian matrix $\nabla^2 f(x)$ requires an intensive computation in large-scale problems when $n \gg 1$. Hence, we apply our proximal-quasi-Newton methods in this case. As shown in [63, Lemma 4], the function f is (M_f, ν)-generalized self-concordant with $\nu = 2$ and $M_f := \frac{\sqrt{6}}{n} \max\left\{\|w^{(j)}\|_2 \mid 1 \leq j \leq n\right\}$.

We implement our proximal quasi-Newton methods to solve (49) and compare them with the accelerated first-order methods implemented in a well-established software package called TFOCS [5]. We use three different variants of TFOCS: TFOCS with N07 (using Nesterov's 2007 method with two proximal operations per iteration), TFOCS with N83 (using Nesterov's 1983 method with one proximal operation per iteration), and TFOCS with AT (using Auslender and Teboulle's accelerated method).

We test on a collection of 26 multi-class datasets downloaded from https://www.csie.ntu.edu.tw/~cjlin/libsvm/. We set the parameter γ in (49) at $\gamma := \frac{0.5}{\sqrt{N}}$ after performing a fine tuning. We terminate all the algorithms if $\|x^{k+1} - x^k\| \leq 10^{-6} \max\left\{1, \|x^k\|\right\}$.

We first plot the convergence behavior in terms of iterations of three proximal Newton-type algorithms we proposed in this paper in Figure 2 (left) for the `dna` problem with 3 classes, 2000 data points, and 180 features.

As we can see from this figure, the proximal Newton method takes fewer iterations than the other two methods. However, each iteration of this method is more expensive than the proximal-quasi-Newton methods due to the evaluation of the Hessian matrix. In our experiment, the quasi-Newton method with L-BFGS outperforms the one with BFGS.

Next, we build a performance profile in time [second] to compare five different algorithms: two proximal quasi-Newton methods proposed in this paper (BFGS and

L-BFGS), and three variants of the accelerated first-order methods implemented in TFOCS.

The performance profile was studied in [17] 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*. 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 $r_{ij} := \frac{T_{ij}}{\min\{T_{ik} \mid k \in \mathcal{S}\}}$. Now, let $\tilde{\rho}_j(\tilde{\tau}) := \frac{1}{n_p}\mathrm{size}\left\{i \in \mathcal{P} \mid r_{ij} \leq \tilde{\tau}\right\}$ for $\tilde{\tau} \in \mathbb{R}_+$. The function $\tilde{\rho}_j : \mathbb{R} \rightarrow [0, 1]$ 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 metric. In the following numerical examples, we plotted the performance profiles in $\log_2$-scale, i.e. $\rho_j(\tau) := \frac{1}{n_p}\mathrm{size}\left\{i \in \mathcal{P} \mid \log_2(r_{i,j}) \leq \tau := \log_2 \tilde{\tau}\right\}$.

Figure 2 (right) shows the performance profile of five algorithms on a collection of 26 problems indicated above. The proximal quasi-Newton method with L-BFGS achieves 13/26 (50%) with the best performance, while the BFGS obtains 10/26 (38%) with the best performance. In terms of computational time, both proximal quasi-Newton methods outperform the optimal proximal gradient methods in this experiment. It is also clear that our proximal quasi-Newton-type methods achieve a more accuracy solution in this experiment compared to the accelerated proximal gradient-type methods implemented in TFOCS.

7 Conclusion

We have generalized the self-concordance notion in [45] 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, just to name a few. We have developed a unified theory with several basic properties to reveal the smoothness structure of this functional class. We have provided several key bounds on local norms, Hessian mapping, gradient mapping, and function value of this functional class. Then, 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. Therefore, we limit our consideration to Newton and proximal Newton methods for $\nu \in [2, 3]$, but our key bounds in Sect. 2.7 remain valid for different ranges of ν with $\nu > 0$.

Our future research is to focus on several aspects. Firstly, we can exploit this theory to develop more practical inexact and quasi-Newton-type methods that can easily capture practical applications in large-scale settings. Secondly, we will combine our approach and stochastic, randomized, and coordinate descent methods to develop new variants of algorithms that can scale better in high-dimensional space. Thirdly, by exploiting both generalized self-concordant, Lipschitz gradient, and strong convexity, one can also develop first-order methods to solve convex optimization problems. Finally, we plan to generalize our theory to primal-dual settings and monotone operators to apply to other classes of convex problems such as convex–concave saddle points, constrained convex optimization, and monotone equations and inclusions.

Acknowledgements This work is partially supported by the NSF-grant No. DMS-1619884, USA.

A Appendix: The proof of technical results

This appendix provides the full proofs of technical results presented in this paper. We prove some technical results used in the paper, and missing proofs in the main text. We also provide a full convergence analysis of the Newton-type methods presented in the main text.

A.1 The proof of Proposition 6: Fenchel's conjugate

Let us consider the set $\mathcal{X} := \{x \in \mathbb{R}^p \mid f(u) - \langle x, u\rangle$ 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 $\mathrm{dom}(f^*) = \{x \in \mathbb{R}^p \mid f^*(x) < +\infty\}$. Take any $x \in \mathrm{dom}(f^*)$, one has $f^*(x) = \max_{u\in\mathrm{dom}(f)}\{\langle x, u\rangle - f(u)\} < +\infty$. Hence, $f(u) - \langle x, u\rangle \geq -f^*(x) > -\infty$ for all $u \in \mathrm{dom}(f)$ which implies $x \in \mathcal{X}$.

Conversely, assume that $x \in \mathcal{X}$. By the definition of $\mathcal{X}$, $f(u) - \langle x, u\rangle$ is bounded from below for all $u \in \mathrm{dom}(f)$. That is, there exists $M \in [0, +\infty)$, such that $f(u) - \langle x, u\rangle \geq -M$ for all $u \in \mathrm{dom}(f)$. By the definition of the conjugate, $f^*(x) = \max_{u\in\mathrm{dom}(f)}\{\langle x, u\rangle - f(u)\} \leq M < +\infty$. Hence, $x \in \mathrm{dom}(f^*)$.

For any $x \in \mathrm{dom}(f^*)$, the optimality condition of $\max_u\{\langle x, u\rangle - f(u)\}$ is $x = \nabla f(u)$. Let us denote by $x(u) = \nabla f(u)$. Then, we have $f^*(x(u)) = \langle x(u), u\rangle - f(u)$. Taking derivative of f^* with respect to x on both sides, and using $x(u) = \nabla f(u)$, we have

$$\nabla_x f^*(x(u)) = u + u'_x x(u) - u'_x \nabla f(u) = u.$$

We further take the second-order derivative of the above equation with respect to u to get

$$\nabla^2 f^*(x(u)) x'_u(u) = \mathbb{I}.$$

Using the two relations above and the fact that $x'_u(u) = \nabla^2 f(u)$, we can derive

$$\langle \nabla f^*(x(u)), x'_u(u)v\rangle = \langle u, x'_u(u)v\rangle = \langle \nabla^2 f(u)v, u\rangle \tag{50}$$

$$\langle \nabla^2 f^*(x(u)) x'_u(u)v, x'_u(u)w\rangle = \langle v, x'_u(u)w\rangle = \langle \nabla^2 f(u)v, w\rangle, \tag{51}$$

where $u \in \mathrm{dom}(f)$, and $v, w \in \mathbb{R}^p$. Using (50) and (51), we can compute the third-order derivative of f^* with respect to $x(u)$ as

$$
\begin{aligned}
\langle\nabla^3 f^*(x(u))[x_u'(u)w]x_u'(u)v, x_u'(u)v\rangle &= \langle\Big(\langle\nabla^2 f^*(x(u))x_u'(u)v, x_u'(u)v\rangle\Big)_u', w\rangle \\
&\quad - 2\langle\nabla^2 f^*(x(u))x_u'(u)v, (x_u'(u)v)_u'w\rangle \\
&\overset{(50)}{=} \langle(\langle x_u'(u)v, v\rangle)_u', w\rangle \\
&\quad - 2\langle\nabla^2 f^*(x(u))x_u'(u)v, (x_u'(u)v)_u'w\rangle \\
&\overset{(51)}{=} \langle\nabla^3 f(u)[w]v, v\rangle - 2\langle(x_u'(u)v)_u'w, v\rangle \\
&= -\langle\nabla^3 f(u)[w]v, v\rangle. \qquad (52)
\end{aligned}
$$

Denote $\xi := x_u'(u)w$ and $\eta := x_u'(u)v$. Since $x_u'(u) = \nabla^2 f(u)$, we have $\xi = \nabla^2 f(u)w$, $\eta = \nabla^2 f(u)v$, and $w = \nabla^2 f(u)^{-1}\xi$. Using these relations and $\nabla^2 f^*(x(u))x_u'(u) = \mathbb{I}$, we can derive

$$
\begin{aligned}
|\langle\nabla^3 f^*(x(u))[\xi]\eta, \eta\rangle| &\overset{(52)}{=} |\langle\nabla^3 f(u)[w]v, v\rangle \overset{(5)}{\le} M_f \left\|v\right\|_u^2 \left\|w\right\|_u^{\nu-2} \left\|w\right\|_2^{3-\nu} \\
&= M_f\langle\nabla^2 f(u)v, v\rangle\langle\nabla^2 f(u)w, w\rangle^{\frac{\nu-2}{2}} \left\|w\right\|_2^{3-\nu} \\
&= M_f\langle\eta, \nabla^2 f^*(x(u))x'(u)v\rangle \\
&\quad \langle\xi, \nabla^2 f^*(x(u))x'(u)w\rangle^{\frac{\nu-2}{2}} \left\|\nabla^2 f(u)^{-1}\xi\right\|^{3-\nu} \\
&= M_f\langle\nabla^2 f^*(x(u))\eta, \eta\rangle\langle\nabla^2 f^*(x(u))\xi, \xi\rangle^{\frac{\nu-2}{2}} \\
&\quad \langle\nabla^2 f^*(x(u))\xi, \nabla^2 f^*(x(u))\xi\rangle^{3-\nu}.
\end{aligned}
$$

For any $H \in \mathcal{S}_{++}^p$, we have $\langle H\xi, \xi\rangle \le \left\|H\xi\right\|_2 \left\|\xi\right\|_2$. For any $\nu \ge 3$, this inequality leads to

$$
\langle H\xi, \xi\rangle^{\frac{\nu-2}{2}} \left\|H\xi\right\|^{3-\nu} \le \langle H\xi, \xi\rangle^{\frac{4-\nu}{2}} \left\|\xi\right\|_2^{\nu-3}.
$$

Using this inequality with $H = \nabla^2 f^*(x(u))$ into the last expression, we obtain

$$
\begin{aligned}
|\langle\nabla^3 f^*(x(u))[\xi]\eta, \eta\rangle| &\le M_f\langle\nabla^2 f^*(x(u))\eta, \eta\rangle\langle\nabla^2 f^*(x(u))\xi, \xi\rangle^{\frac{4-\nu}{2}} \left\|\xi\right\|_2^{\nu-3} \\
&= M_f\left\|\eta\right\|_{x(u)}^2 \left\|\xi\right\|_{x(u)}^{4-\nu} \left\|\xi\right\|_2^{\nu-3}.
\end{aligned}
$$

By Definition 2, we need $\nu - 3 = 3 - \nu_*$ and $4 - \nu = \nu_* - 2$ which hold if $\nu_* = 6 - \nu$. Under the choice of ν_*, the above inequality shows that f^* is (M_{f^*}, ν_*)-generalized self-concordant with $M_{f^*} = M_f$ and $\nu_* = 6 - \nu$. However, to guarantee $\nu - 3 \ge 0$ and $6 - \nu > 0$, we require $3 \le \nu < 6$.

Finally, we prove the case of univariate functions, i.e., $p = 1$. Indeed, we have

$$
x(u) = f'(u), \quad (f^*)'(x(u)) = u, \quad \text{and} \quad (f^*)''(x(u))x'(u) = 1. \qquad (53)
$$

Here, f' is the derivative of f with respect to u. Taking the derivative of the last equation on both sides with respect to u, we obtain

$$
(f^*)'''(x(u))(x'(u))^2 + (f^*)''(x(u))x''(u) = 0.
$$

Solving this equation for $(f^*)'''(x(u))$ and then using (53) and $x''(u) = f'''(u)$, we get

$$\begin{aligned}\left|(f^*)'''(x(u))\right| &= \left|\tfrac{(f^*)''(x(u))x''(u)}{(x'(u))^2}\right| = \left|((f^*)''(x(u)))^3 f'''(u)\right| \\ &\le M_f \left|((f^*)''(x(u)))^3 (f''(u))^{\frac{\nu}{2}}\right| = M_f((f^*)''(x(u)))^{\frac{6-\nu}{2}}.\end{aligned}$$

This inequality shows that f^* is generalized self-concordant with $\nu_* = 6 - \nu$ for any $\nu \in (0, 6)$. □

A.2 The proof of Corollary 2: bound on the mean of Hessian operator

Let $y_\tau := x + \tau(y - x)$. Then $d_\nu(x, y_\tau) = \tau d_\nu(x, y)$. By (15), we have $\nabla^2 f(x + \tau(y - x)) \preceq (1 - \tau d_\nu(x, y))^{\frac{-2}{\nu-2}} \nabla^2 f(x)$ and $\nabla^2 f(x + \tau(y - x)) \succeq (1 - \tau d_\nu(x, y))^{\frac{2}{\nu-2}} \nabla^2 f(x)$. Hence, we have

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

where $\underline{I}_\nu(x, y) := \int_0^1 (1 - \tau d_\nu(x, y))^{\frac{2}{\nu-2}} d\tau$ and $\overline{I}_\nu(x, y) := \int_0^1 (1 - \tau d_\nu(x, y))^{\frac{-2}{\nu-2}} d\tau$ are the two integrals in the above inequality. Computing these integrals explicitly, we can show that

- If $\nu = 4$, then $\underline{I}_\nu(x, y) = \frac{1-(1-d_4(x,y))^2}{2d_4(x,y)}$ and $\overline{I}_\nu(x, y) = \frac{-\ln(1-d_4(x,y))}{d_4(x,y)}$.
- If $\nu \neq 4$, then we can easily compute $\underline{I}_\nu(x, y) = \frac{(\nu-2)}{\nu d_\nu(x,y)}\left(1 - (1 - d_\nu(x, y))^{\frac{\nu}{\nu-2}}\right)$, and $\overline{I}_\nu(x, y) = \frac{(\nu-2)}{(\nu-4)d_\nu(x,y)}\left(1 - (1 - d_\nu(x, y))^{\frac{\nu-4}{\nu-2}}\right)$.

Hence, we obtain (18).

Finally, we prove for the case $\nu = 2$. Indeed, by (16), we have $e^{-d_2(x,y_\tau)}\nabla^2 f(x) \preceq \nabla^2 f(y_\tau) \preceq e^{d_2(x,y_\tau)}\nabla^2 f(x)$. Since $d_2(x, y_\tau) = \tau d_2(x, y)$, the last estimate leads to

$$\left(\int_0^1 e^{-d_2(x,y)\tau}d\tau\right)\nabla^2 f(x) \preceq \int_0^1 \nabla^2 f(y_\tau)d\tau \preceq \left(\int_0^1 e^{d_2(x,y)\tau}d\tau\right)\nabla^2 f(x),$$

which is exactly (18). □

A.3 Techical lemmas

The following lemmas will be used in our analysis. Lemma 1 is elementary, but we provide its proof for completeness.

Lemma 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*

$$\lim_{t\to 0^+} \psi_r(t) = \tfrac{r+1}{2}, \quad \lim_{t\to 1^-} \psi_r(t) = +\infty, \quad \textit{and} \quad \sup_{0\leq t\leq \bar{t}} |\psi_r(t)| \leq \bar{C}_r(\bar{t}) < +\infty,$$

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

$$\begin{aligned}\lim_{t\to 0^+} \psi_r(t) &= \frac{\lim_{t\to 0^+} r(r+1)t(1-t)^{r-1}}{\lim_{t\to 0^+} rt(2-(2+r)t)(1-t)^{r-1}} \\ &= \frac{\lim_{t\to 0^+}(r+1)}{\lim_{t\to 0^+}(2-(2+r)t)} = \frac{r+1}{2}.\end{aligned}$$

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 2, we can prove the following lemma on the bound of the Hessian difference.

Lemma 2 *Given* $x, y \in \mathrm{dom}(f)$*, the matrix* $H(x, y)$ *defined by*

$$H(x, y) := \nabla^2 f(x)^{-1/2}\left[\int_0^1 \left(\nabla^2 f(x + \tau(y-x)) - \nabla^2 f(x)\right)d\tau\right]\nabla^2 f(x)^{-1/2}, \quad (54)$$

satisfies

$$\|H(x, y)\| \leq R_\nu\left(d_\nu(x, y)\right) d_\nu(x, y), \quad (55)$$

where $R_\nu(t)$ *is defined as follows for* $t \in [0, 1)$*:*

$$R_\nu(t) := \begin{cases} \left(\frac{3}{2} + \frac{t}{3}\right) e^t & \textit{if } \nu = 2 \\ \dfrac{1-(1-t)^{\frac{4-\nu}{\nu-2}} - \left(\frac{4-\nu}{\nu-2}\right)t(1-t)^{\frac{4-\nu}{\nu-2}}}{\left(\frac{4-\nu}{\nu-2}\right)t^2(1-t)^{\frac{4-\nu}{\nu-2}}} & \textit{if} 2 < \nu \leq 3. \end{cases} \quad (56)$$

Moreover, for a fixed $\bar{t} \in (0, 1)$*, we have* $\sup_{0 \le t \le \bar{t}} |R_\nu(t)| \le \bar{M}_\nu(\bar{t})$*, where*

$$\bar{M}_\nu(\bar{t}) := \max\left\{ \frac{1 - (1-\bar{t})^{\frac{4-\nu}{\nu-2}} - \left(\frac{4-\nu}{\nu-2}\right)\bar{t}(1-\bar{t})^{\frac{4-\nu}{\nu-2}}}{\left(\frac{4-\nu}{\nu-2}\right)\bar{t}^2(1-\bar{t})^{\frac{4-\nu}{\nu-2}}}, \left(\frac{3}{2} + \frac{\bar{t}}{2}\right)e^{\bar{t}} \right\} \in (0, +\infty).$$

Proof By Corollary 2, if we define $G(x, y) := \int_0^1 \left[\nabla^2 f(x + \tau(y - x)) - \nabla^2 f(x)\right] d\tau$, then

$$\left[\underline{\kappa}_\nu(d_\nu(x, y)) - 1\right]\nabla^2 f(x) \preceq G(x, y) \preceq [\overline{\kappa}_\nu(d_\nu(x, y)) - 1]\nabla^2 f(x). \tag{57}$$

Since $H(x, y) = \nabla^2 f(x)^{-1/2} G(x, y) \nabla^2 f(x)^{-1/2}$, the last inequality implies

$$\|H(x, y)\| \le \max\left\{1 - \underline{\kappa}_\nu(d_\nu(x, y)), \overline{\kappa}_\nu(d_\nu(x, y)) - 1\right\}.$$

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 three cases.

(a) For $\nu = 2$, since $e^{-t} + e^t \ge 2$, we have $\frac{1-e^{-t}}{t} + \frac{e^t-1}{t} \ge 2$ which implies $C_{\max}(t) = \overline{\kappa}_\nu(t) - 1 = \frac{e^t-1-t}{t}$. Hence, by Lemma 1, we have $C_{\max}(t) \le \left(\frac{3}{2} + \frac{t}{3}\right)te^t$ which leads to $R_\nu(t) := \left(\frac{3}{2} + \frac{t}{3}\right)e^t$.

(b) For $\nu \in (2, 3]$, we have

$$\begin{aligned} C_{\max}(t) &= \max\left\{1 - \frac{(\nu-2)}{\nu t}\left[1 - (1-t)^{\frac{\nu}{\nu-2}}\right], \frac{(\nu-2)}{(4-\nu)t}\left[\frac{1}{(1-t)^{\frac{4-\nu}{\nu-2}}} - 1\right] - 1\right\} \\ &= \frac{(\nu-2)}{(4-\nu)t}\left[\frac{1}{(1-t)^{\frac{4-\nu}{\nu-2}}} - 1\right] - 1. \end{aligned}$$

Indeed, we show that $\frac{(\nu-2)}{(4-\nu)t}\left[\frac{1}{(1-t)^{\frac{4-\nu}{\nu-2}}} - 1\right] + \frac{(\nu-2)}{\nu t}\left[1 - (1-t)^{\frac{\nu}{\nu-2}}\right] \ge 2$. Let $u := \frac{4-\nu}{\nu-2} > 0$ and $v := \frac{\nu}{\nu-2} > 0$. The last inequality is equivalent to $\frac{1}{u}\left[\frac{1}{(1-t)^u} - 1\right] + \frac{1}{v}\left[1 - (1-t)^v\right] \ge 2t$ which can be reformulated as $\frac{1}{v} - \frac{1}{u} + \frac{1}{u(1-t)^u} - \frac{(1-t)^v}{v} - 2t \ge 0$. Consider $s(t) := \frac{1}{v} - \frac{1}{u} + \frac{1}{u(1-t)^u} - \frac{(1-t)^v}{v} - 2t$. It is clear that $s'(t) = \frac{1}{(1-t)^{u+1}} + (1-t)^{v-1} - 2 = (1-t)^{-\frac{2}{\nu-2}} + (1-t)^{\frac{2}{\nu-2}} - 2 \ge 0$ for all $t \in [0, 1)$. We obtain $s(t) \ge s(0) = 0$. Hence, $C_{\max}(t) = \frac{(\nu-2)}{(4-\nu)t}\left[\frac{1}{(1-t)^{\frac{4-\nu}{\nu-2}}} - 1\right] - 1$.

Let us define $r := \frac{4-\nu}{\nu-2} = \frac{2}{\nu-2} - 1$. Then, it is clear that $\nu = 2 + \frac{2}{1+r}$, and $\nu \in (2, 3]$ is equivalent to $r \ge 1$. Now, using Lemma 1 with $r = \frac{2}{\nu-2} - 1 \ge 1$, we obtain $R_\nu(t) := \frac{1-(1-t)^{\frac{4-\nu}{\nu-2}} - \left(\frac{4-\nu}{\nu-2}\right)t(1-t)^{\frac{4-\nu}{\nu-2}}}{\left(\frac{4-\nu}{\nu-2}\right)t^2(1-t)^{\frac{4-\nu}{\nu-2}}}$. Put (a) and (b) together, we obtain (55) with R_ν defined by (56). The boundedness of R_ν follows from Lemma 1. □

A.4 The proof of Theorem 4: solution existence and uniqueness

Consider a sublevel set $\mathcal{L}_F(x) := \{y \in \mathrm{dom}(F) \mid F(y) \leq F(x)\}$ of F in (32). For any $y \in \mathcal{L}_F(x)$ and $v \in \partial g(x)$, by (22) and the convexity of g, we have

$$F(x) \geq F(y) \geq F(x) + \langle \nabla f(x) + v, y - x\rangle + \omega_\nu\left(-d_\nu(x, y)\right)\|y - x\|_x^2 .$$

By the Cauchy-Schwarz inequality, we have

$$\omega_\nu\left(-d_\nu(x, y)\right)\|y - x\|_x \leq \|\nabla f(x) + v\|_x^* . \tag{58}$$

Now, using the assumption $\nabla^2 f(x) \succ 0$ for some $x \in \mathrm{dom}(F)$, we have $\sigma_{\min}(x) := \lambda_{\min}(\nabla^2 f(x)) > 0$, the smallest eigenvalue of $\nabla^2 f(x)$.

(a) If $\nu = 2$, then $d_2(x, y) = M_f\|y - x\|_2 \leq \frac{M_f}{\sqrt{\sigma_{\min}(x)}}\|y - x\|_x$. This estimate together with (58) imply

$$\omega_2\left(-d_2(x, y)\right)d_2(x, y) \leq \frac{M_f}{\sqrt{\sigma_{\min}(x)}}\|\nabla f(x) + v\|_x^* = \frac{M_f}{\sqrt{\sigma_{\min}(x)}}\lambda(x). \tag{59}$$

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 $\lim_{t\to+\infty} s_2(t) = 1$. Therefore, if $\frac{M_f}{\sqrt{\sigma_{\min}(x)}}\lambda(x) < 1$, then the equation $s_2(t) - \frac{M_f}{\sqrt{\sigma_{\min}(x)}}\lambda(x) = 0$ has a unique solution $t^* \in (0, +\infty)$. In this case, for $0 \leq d_2(x, y) \leq t^*$, (59) holds. This condition leads to $M_f\|y - x\|_2 \leq t^* < +\infty$ which implies that the sublevel set $\mathcal{L}_F(x)$ is bounded. Consequently, solution $x^\star$ of (32) exists.

(b) If $2 < \nu < 3$, then

$$d_\nu(x, y) \leq \left(\frac{\nu}{2} - 1\right)\frac{M_f}{\sigma_{\min}(x)^{\frac{3-\nu}{2}}}\|y - x\|_x .$$

This inequality together with (58) imply

$$\begin{aligned}\omega_\nu\left(-d_\nu(x, y)\right)d_\nu(x, y) &\leq \left(\frac{\nu}{2} - 1\right)\frac{M_f}{\sigma_{\min}(x)^{\frac{3-\nu}{2}}}\|\nabla f(x) + v\|_x^* \\ &= \left(\frac{\nu}{2} - 1\right)\frac{M_f}{\sigma_{\min}(x)^{\frac{3-\nu}{2}}}\lambda(x).\end{aligned}$$

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 $s_\nu(t) < \frac{\nu-2}{4-\nu}$ for all $t > 0$, and $\lim_{t\to+\infty} s_\nu(t) = \frac{\nu-2}{4-\nu}$. Hence, if $\left(\frac{\nu}{2} - 1\right)\frac{M_f}{\sigma_{\min}(x)^{\frac{3-\nu}{2}}}\lambda(x) < \frac{\nu-2}{4-\nu}$, then, similar to

Case (a), we can show that solution $x^\star$ of (32) exists. This condition implies that $\lambda(x) < \frac{2\sigma_{\min}(x)^{\frac{3-\nu}{2}}}{(4-\nu)M_f}$.

(c) If $\nu = 3$, then $d_3(x, y) = \frac{M_f}{2}\|y - x\|_x$. Combining this estimate and (58) we get

$$\omega_3\left(-d_3(x, y)\right) d_3(x, y) \leq \frac{M_f}{2}\|\nabla f(x) + v\|_x^* .$$

With the same proof as in [40, Theorem 4.1.11], if $\frac{M_f}{2}\|\nabla f(x) + v\|_x^* < 1$ which is equivalent to $\lambda(x) < \frac{2}{M_f}$, then solution $x^\star$ of (32) exists.

Note that the condition on $\lambda(x)$ in three cases (a), (b), and (c) can be unified. The uniqueness of the solution $x^\star$ in these three cases follows from the strict convexity of F. □

A.5 The proof of Theorem 2: convergence of the damped-step Newton method

The proof of this theorem is divided into two parts: computing the step-size, and proving the local quadratic convergence.

Computing the step-size τ_k: From Proposition 10, for any $x^k, x^{k+1} \in \mathrm{dom}(f)$, if $d_\nu(x^k, x^{k+1}) < 1$, then we have

$$f(x^{k+1}) \leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \omega_\nu\left(d_\nu(x^k, x^{k+1})\right)\left\|x^{k+1} - x^k\right\|_{x^k}^2 .$$

Now, using (25), we have $\langle \nabla f(x^k), x^{k+1} - x^k\rangle = -\tau_k\left(\|\nabla f(x^k)\|_{x^k}^*\right)^2 = -\tau_k\lambda_k^2$. On the other hand, we have

$$\begin{aligned}
\|x^{k+1} - x^k\|_{x^k}^2 &\overset{(25)}{=} \tau_k^2\langle \nabla^2 f(x^k)^{-1}\nabla f(x^k), \nabla f(x^k)\rangle \overset{(27)}{=} \tau_k^2\lambda_k^2,\\
\|x^{k+1} - x^k\|_2^2 &\overset{(25)}{=} \tau_k^2\langle \nabla^2 f(x^k)^{-1}\nabla f(x^k), \nabla^2 f(x^k)^{-1}\nabla f(x^k)\rangle \overset{(27)}{=} \frac{\tau_k^2\beta_k^2}{M_f^2}.
\end{aligned}$$

Using the definition of $d_\nu(\cdot)$ in (12), the two last equalities, and (28), we can easily show that $d_\nu(x^k, x^{k+1}) = \tau_k d_k$. Substituting these relations into the first estimate, we obtain

$$f(x^{k+1}) \leq f(x^k) - \left(\tau_k\lambda_k^2 - \omega_\nu\left(\tau_k d_k\right)\tau_k^2\lambda_k^2\right).$$

We consider the following cases:

(a) If $\nu = 2$, then, by (23), we have $\eta_k(\tau) := \lambda_k^2\tau - \left(\frac{\lambda_k}{d_k}\right)^2\left(e^{\tau d_k} - \tau d_k - 1\right)$ with $d_k = \beta_k$. This function attains the maximum at $\tau_k := \frac{\ln(1+d_k)}{d_k} = \frac{\ln(1+\beta_k)}{\beta_k} \in (0, 1)$ with

$$\eta_k(\tau_k) = \left(\frac{\lambda_k}{d_k}\right)^2\Big[(1+d_k)\ln(1+d_k) - d_k\Big] = \left(\frac{\lambda_k}{\beta_k}\right)^2\Big[(1+\beta_k)\ln(1+\beta_k) - \beta_k\Big].$$

It is easy to check from the right-most term of the last expression that $\varDelta_k := \eta_k(\tau_k) > 0$ for $\tau_k > 0$.

(b) If $\nu = 3$, then, by (23), we have $\eta_k(\tau) := \lambda_k^2\tau + \left(\frac{\lambda_k}{d_k}\right)^2 [\tau d_k + \ln(1 - \tau d_k)]$ with $d_k = 0.5M_f\lambda_k$. We can show that $\eta_k(\tau)$ achieves the maximum at $\tau_k = \frac{1}{1+d_k} = \frac{1}{1+0.5M_f\lambda_k} \in (0, 1)$ with

$$\eta_k(\tau_k) = \frac{\lambda_k^2}{1 + 0.5M_f\lambda_k} + \left(\frac{2}{M_f}\right)^2 \left[\frac{0.5M_f\lambda_k}{1 + 0.5M_f\lambda_k} + \ln\left(1 - \frac{0.5M_f\lambda_k}{1 + 0.5M_f\lambda_k}\right)\right].$$

We can also easily check that the last term $\varDelta_k := \eta_k(\tau_k)$ of this expression is positive for $\lambda_k > 0$.

(c) If $2 < \nu < 3$, then we have $d_k = M_f^{\nu-2}\left(\frac{\nu}{2} - 1\right)\lambda_k^{\nu-2}\beta_k^{3-\nu}$. By (23), we have

$$\eta_k(\tau) = \left(\lambda_k^2 + \frac{\lambda_k^2}{d_k}\frac{\nu - 2}{4 - \nu}\right)\tau - \left(\frac{\lambda_k}{d_k}\right)^2 \frac{(\nu - 2)^2}{2(4 - \nu)(3 - \nu)}\left((1 - \tau d_k)^{\frac{2(3-\nu)}{2-\nu}} - 1\right).$$

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:

$$\eta_k'(\tau) = \lambda_k^2\left[1 - \frac{1}{d_k}\frac{\nu - 2}{\nu - 4}\left(1 - (1 - \tau d_k)^{\frac{\nu-4}{\nu-2}}\right)\right] \text{ and } \eta_k''(\tau) = -\lambda_k^2(1 - \tau d_k)^{\frac{-2}{\nu-2}}.$$

Let us set $\eta_k'(\tau_k) = 0$. Then, we get

$$\tau_k = \frac{1}{d_k}\left[1 - \left(1 + \frac{4 - \nu}{\nu - 2}d_k\right)^{-\frac{\nu-2}{4-\nu}}\right] \in (0, 1) \quad \text{(by the Bernoulli inequality),}$$

with

$$\eta_k(\tau_k) = \frac{\lambda_k^2}{d_k}\left[1 - \frac{4 - \nu}{2(3 - \nu)}\left(1 + \frac{4 - \nu}{\nu - 2}d_k\right)^{2-\nu}\right] + \left(\frac{\lambda_k}{d_k}\right)^2 \frac{\nu - 2}{2(3 - \nu)}\left[1 - \left(1 + \frac{4 - \nu}{\nu - 2}d_k\right)^{2-\nu}\right].$$

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 $\varDelta_k := \eta_k(\tau_k) > \eta_k(0) = 0$.

The proof of local quadratic convergence Let $x_f^\star$ be the optimal solution of (24). We have

$$\begin{aligned}\|x^{k+1} - x_f^\star\|_{x^k} &= \|x^k - \tau_k\nabla^2 f(x^k)^{-1}\nabla f(x^k) - x_f^\star\|_{x^k} \\ &= (1 - \tau_k)\|x^k - x_f^\star\|_{x^k} + \tau_k\|x^k - x_f^\star - \nabla^2 f(x^k)^{-1}\nabla f(x^k)\|_{x^k}.\end{aligned}$$

Hence, we can write

$$\begin{aligned}\|x^{k+1}- x^{\star}_f\|_{x^k} = (1-\tau_k)\|x^k - x^{\star}_f\|_{x^k}+\tau_k\|\nabla^2 f(x^k)^{-1}\\ \times\Big[\nabla f(x^{\star}_f) - \nabla f(x^k)-\nabla^2 f(x^k)(x^{\star}_f-x^k)\Big]\|_{x^k}.\end{aligned}\tag{60}$$

Let us define $T_k := \Big\|\nabla^2 f(x^k)^{-1}\Big[\nabla f(x^{\star}_f) - \nabla f(x^k)-\nabla^2 f(x^k)(x^{\star}_f-x^k)\Big]\Big\|_{x^k}$ and consider three cases as follows:

(a) For $\nu = 2$, using Corollary 2, we have $\left(\frac{1-e^{-\bar\beta_k}}{\bar\beta_k}\right)\nabla^2 f(x^k) \preceq \int_0^1 \nabla^2 f(x^k + t(x^{\star}_f - x^k))dt \preceq \left(\frac{e^{\bar\beta_k}-1}{\bar\beta_k}\right)\nabla^2 f(x^k)$, where $\bar\beta_k := M_f\|x^k - x^{\star}_f\|_2$. Using the above inequality, we can show that

$$\begin{aligned}T_k &\le \max\left\{1-\frac{1-e^{-\bar\beta_k}}{\bar\beta_k}, \frac{e^{\bar\beta_k}-1}{\bar\beta_k}-1\right\}\|x^k - x^{\star}_f\|_{x^k}\\ &= \left(\frac{e^{\bar\beta_k}-1-\bar\beta_k}{\bar\beta_k^2}\right)\bar\beta_k\|x^k - x^{\star}_f\|_{x^k}.\end{aligned}$$

Let $\underline{\sigma}_k := \lambda_{\min}(\nabla^2 f(x^k))$. We first derive

$$\begin{aligned}\|\nabla^2 f(x^k)^{-1}\nabla f(x^k)\|_2 &= \|\nabla^2 f(x^k)^{-1}(\nabla f(x^k) - \nabla f(x^{\star}_f))\|_2\\ &= \|\textstyle\int_0^1 \nabla^2 f(x^k)^{-1}\nabla^2 f(x^k + t(x^{\star}_f - x^k))(x^k - x^{\star}_f)dt\|_2\\ &= \|\nabla^2 f(x^k)^{-1/2}K(x^k, x^{\star}_f)\nabla^2 f(x^k)^{1/2}(x^k - x^{\star}_f)\|_2\\ &\le \tfrac{1}{\sqrt{\underline{\sigma}_k}}\|K(x^k, x^{\star}_f)\|\|x^k - x^{\star}_f\|_{x^k}.\end{aligned}$$

where $K(x^k, x^{\star}_f) := \int_0^1 \nabla^2 f(x^k)^{-1/2}\nabla^2 f(x^k + t(x^{\star}_f - x^k)\nabla^2 f(x^k)^{-1/2}dt$. Using Corollary 2 and noting that $\bar\beta_k := M_f\|x^k - x^{\star}_f\|_2$, we can estimate $\|K(x^k, x^{\star}_f)\| \le \frac{e^{\bar\beta_k}-1}{\bar\beta_k}$. Using the two last estimates, and the definition of β_k, we can derive

$$\beta_k = M_f\|\nabla^2 f(x^k)^{-1}\nabla f(x^k)\|_2 \le \frac{M_f e^{\bar\beta_k-1}}{\bar\beta_k\sqrt{\underline{\sigma}_k}}\|x^k - x^{\star}_f\|_{x^k} \le M_f e\frac{\|x^k-x^{\star}_f\|_{x^k}}{\sqrt{\underline{\sigma}_k}},$$

provided that $\bar\beta_k \le 1$. Since, the step-size $\tau_k = \frac{1}{\beta_k}\ln(1+\beta_k)$, we have $1-\tau_k \le \frac{\beta_k}{2} \le \frac{M_f e\|x^k-x^{\star}_f\|_{x^k}}{2\sqrt{\underline{\sigma}_k}}$. On the other hand, $\frac{e^{\bar\beta_k}-1-\bar\beta_k}{\bar\beta_k^2} \le \frac{e}{2}$ for all $0 \le \bar\beta_k \le 1$. Substituting T_k into (60) and using these relations, we have

$$\|x^{k+1} - x^{\star}_f\|_{x^k} \le \tfrac{e}{2}\bar\beta_k\|x^k - x^{\star}_f\|_{x^k} + \frac{M_f e}{2}\frac{\|x^k-x^{\star}_f\|^2_{x^k}}{\sqrt{\underline{\sigma}_k}},$$

provided that $\bar\beta_k \le 1$. On the other hand, by Proposition 8, we have $\|x^{k+1} - x^{\star}_f\|_{x^{k+1}} \le e^{\frac{\bar\beta_{k+1}+\bar\beta_k}{2}}\|x^{k+1} - x^{\star}_f\|_{x^k}$ and $\underline{\sigma}_{k+1}^{-1} \le e^{\bar\beta_k+\bar\beta_{k+1}}\underline{\sigma}_k^{-1}$. In addition, $\bar\beta_k \le \frac{M_f}{\sqrt{\underline{\sigma}_k}}\|x^k - x^{\star}_f\|_{x^k}$

Combining the above inequalities, we finally get

$$\frac{\|x^{k+1} - x^{\star}_f\|_{x^{k+1}}}{\sqrt{\underline{\sigma}_{k+1}}} \leq M_f e^{1+\bar{\beta}_{k+1}+\bar{\beta}_k} \left(\frac{\|x^k - x^{\star}_f\|_{x^k}}{\sqrt{\underline{\sigma}_k}}\right)^2.$$

Under the fact that $\beta_k \leq 1$, and $\beta_{k+1} \leq 1$, this estimate shows that $\left\{\frac{\|x^k - x^{\star}_f\|_{x^k}}{\sqrt{\underline{\sigma}_k}}\right\}$ quadratically converges to zero. Since $\|x^k - x^{\star}_f\|_2 \leq \frac{\|x^k - x^{\star}_f\|_{x^k}}{\sqrt{\underline{\sigma}_k}}$, we can also conclude that $\left\{\|x^k - x^{\star}_f\|_2\right\}$ quadratically converges to zero.

(b) For $\nu = 3$, we can follow [40]. However, for completeness, we give a short proof here. Using Corollary 2, we have $\left(1 - r_k + \frac{r_k^2}{3}\right)\nabla^2 f(x^k) \preceq \int_0^1 \nabla^2 f(x^k + t(x^{\star}_f - x^k))dt \preceq \frac{1}{1-r_k}\nabla^2 f(x^k)$, where $r_k := 0.5M_f\|x^k - x^{\star}_f\|_{x^k} < 1$. Using the above inequality, we can show that

$$T_k \leq \max\left\{r_k - \frac{r_k^2}{3}, \frac{r_k}{1-r_k}\right\}\|x^k - x^{\star}_f\|_{x^k} = \frac{0.5M_f\|x^k - x^{\star}_f\|^2_{x^k}}{1 - 0.5M_f\|x^k - x^{\star}_f\|_{x^k}}.$$

Substituting T_k into (60) and using $\tau_k = \frac{1}{1+0.5M_f\lambda_k}$, we have

$$\begin{aligned}\|x^{k+1} - x^{\star}_f\|_{x^k} \leq{}& \frac{0.5M_f\lambda_k}{1+0.5M_f\lambda_k}\|x^k - x^{\star}_f\|_{x^k} \\ &+ \frac{1}{1+0.5M_f\lambda_k}\left(\frac{0.5M_f\|x^k - x^{\star}_f\|^2_{x^k}}{1 - 0.5M_f\|x^k - x^{\star}_f\|_{x^k}}\right).\end{aligned}$$

Next, we need to upper bound λ_k. Since $\nabla f(x^{\star}_f) = 0$. Using Corollary 2, we can bound λ_k as

$$\begin{aligned}\lambda_k &= \|\nabla f(x^k)\|^*_{x^k} = \|\nabla^2 f(x^k)^{-1/2}(\nabla f(x^k) - \nabla f(x^{\star}_f))\|_2 \\ &= \|\textstyle\int_0^1 \nabla^2 f(x^k)^{-1/2}\nabla^2 f(x^k + t(x^{\star}_f - x^k))(x^{\star}_f - x^k)dt\|_2 \\ &\leq \|x^k - x^{\star}_f\|_{x^k}\|\textstyle\int_0^1 \nabla^2 f(x^k)^{-1/2}\nabla^2 f(x^k + t(x^{\star}_f - x^k))\nabla^2 f(x^k)^{-1/2}dt\|_2 \\ &\overset{\text{Corollary 2}}{\leq} \frac{\|x^k - x^{\star}_f\|_{x^k}}{1 - 0.5M_f\|x^k - x^{\star}_f\|_{x^k}} \leq 2\|x^k - x^{\star}_f\|_{x^k},\end{aligned}$$

provided that $M_f\|x^k - x^{\star}_f\|_{x^k} < 1$. Overestimating the above inequality using this bound, we get

$$\begin{aligned}\|x^{k+1} - x^{\star}_f\|_{x^k} &\leq 0.5M_f\lambda_k\|x^k - x^{\star}_f\|_{x^k} + \frac{0.5M_f\|x^k - x^{\star}_f\|^2_{x^k}}{1 - 0.5M_f\|x^k - x^{\star}_f\|_{x^k}} \\ &\leq M_f\|x^k - x^{\star}_f\|^2_{x^k} + M_f\|x^k - x^{\star}_f\|^2_{x^k} = 2M_f\|x^k - x^{\star}_f\|^2_{x^k},\end{aligned}$$

provided that $M_f\|x^k - x^\star_f\|_{x^k} < 1$. On the other hand, we can also estimate $\|x^{k+1} - x^\star_f\|_{x^{k+1}} \le \frac{\|x^{k+1}-x^\star_f\|_{x^k}}{1-0.5M_f\left(\|x^{k+1}-x^\star_f\|_{x^k}+\|x^k-x^\star_f\|_{x^k}\right)}$. Combining the last two inequalities, we get

$$\|x^{k+1} - x^\star_f\|_{x^{k+1}} \le \frac{2M_f\|x^k - x^\star_f\|^2_{x^k}}{1-2M_f\|x^k-x^\star_f\|^2_{x^k} - 0.5M_f\|x^k-x^\star_f\|_{x^k}}$$

The right-hand side function $\psi(t) = \frac{2M_f}{1-2M_ft^2-0.5M_ft} \le 4M_f$ on $t\in\left[0,\frac{1}{2M_f}\right]$. Hence, if $\|x^k-x^\star_f\|_{x^k}\le\frac{1}{2M_f}$, then $\|x^{k+1}-x^\star_f\|_{x^{k+1}} \le 4M_f\|x^k-x^\star_f\|^2_{x^k}$. This shows that if $x^0\in\mathrm{dom}(f)$ is chosen such that $\|x^0-x^\star_f\|_{x^0}\le\frac{1}{4M_f}$, then $\left\{\|x^k-x^\star_f\|_{x^k}\right\}$ quadratically converges to zero.

(c) For $\nu\in(2,3)$, with the same argument as in the proof of Theorem 3, we can show that

$$\|x^{k+1}-x^\star_f\|_{x^k} \le R_\nu(d^k_\nu)d^k_\nu\|x^k-x^\star_f\|_{x^k},$$

where R_ν is defined by (56) and $d^k_\nu := M_f^{\nu-2}\left(\frac{\nu}{2}-1\right)\|x^k-x^\star_f\|_2^{3-\nu}\|x^k-x^\star_f\|^{\nu-2}_{x^k}$. Using again the argument as in the proof of Theorem 3, we have

$$\frac{\|x^{k+1}-x^\star_f\|_{x^{k+1}}}{\underline{\sigma}_{k+1}^{\frac{3-\nu}{2}}} \le C_\nu(d^k_\nu,\|x^k-x^\star_f\|_{x^k})\left(\frac{\|x^k-x^\star_f\|_{x^k}}{\underline{\sigma}_k^{\frac{3-\nu}{2}}}\right)^2.$$

Here, $C_\nu(\cdot,\cdot)$ is a given function deriving from R_ν. Under the condition that d^k_ν and $\|x^k-x^\star_f\|_{x^k}$ are sufficiently small, we can show that $C_\nu(d^k_\nu,\|x^k-x^\star_f\|_{x^k})\le\bar{C}_\nu$. Hence, the last inequality shows that $\left\{\frac{\|x^k-x^\star_f\|_{x^k}}{\underline{\sigma}_k^{\frac{3-\nu}{2}}}\right\}$ quadratically converges to zero. Since $\underline{\sigma}_k^{\frac{3-\nu}{2}}\|x^k-x^\star_f\|_{H_k}\le\|x^k-x^\star_f\|_{x^k}$, where $H_k:=\nabla^2f(x^k)^{\frac{\nu-2}{2}}$, we have $\|x^k-x^\star_f\|_{H_k}\le\frac{\|x^k-x^\star_f\|_{x^k}}{\underline{\sigma}_k^{\frac{3-\nu}{2}}}$. Hence, we can conclude that $\left\{\|x^k-x^\star_f\|_{H_k}\right\}$ also locally converges to zero at a quadratic rate. □

A.6 The proof of Theorem 3: the convergence of the full-step Newton method

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 $\left\{\|x^k-x^\star_f\|_{H_k}\right\}$.

The quadratic convergence of $\left\{\frac{\lambda_k}{\underline{\sigma}_k^{\frac{3-\nu}{2}}}\right\}$: Since the full-step Newton scheme updates $x^{k+1}:=x^k-\nabla^2f(x^k)^{-1}\nabla f(x^k)$, if we denote by $n^k_{\mathrm{nt}}=x^{k+1}-x^k=-\nabla^2f(x^k)^{-1}\nabla f(x^k)$, then the last expression leads to $\nabla f(x^k)+\nabla^2f(x^k)n^k_{\mathrm{nt}}=0$. In addition, $\|n^k_{\mathrm{nt}}\|_{x^k}=\|\nabla f(x^k)\|^*_{x^k}=\lambda_k$. Using the definition of $d_\nu(\cdot,\cdot)$ in (12), we denote $d^k_\nu:=d_\nu(x^k,x^{k+1})$.

First, by $\nabla f(x^k) + \nabla^2 f(x^k) n_{\mathrm{nt}}^k = 0$ and the mean-value theorem, we can show that

$$\begin{aligned}\nabla f(x^{k+1}) &= \nabla f(x^{k+1}) - \nabla f(x^k) - \nabla^2 f(x^k) n_{\mathrm{nt}}^k \\ &= \int_0^1 \left[\nabla^2 f(x^k + t n_{\mathrm{nt}}^k) - \nabla^2 f(x^k)\right] n_{\mathrm{nt}}^k dt.\end{aligned}$$

Let us define $G_k := \int_0^1 \left[\nabla^2 f(x^k + t n_{\mathrm{nt}}^k) - \nabla^2 f(x^k)\right] dt$ and $H_k := \nabla^2 f(x^k)^{-1/2} G_k \nabla^2 f(x^k)^{-1/2}$. Then, the above estimate implies $\nabla f(x^{k+1}) = G_k n_{\mathrm{nt}}^k$. Hence, we can show that

$$\begin{aligned}\left[\|\nabla f(x^{k+1})\|_{x^k}^*\right]^2 &= \langle \nabla^2 f(x^k)^{-1} G_k n_{\mathrm{nt}}^k, G_k n_{\mathrm{nt}}^k\rangle \\ &= \langle H_k \nabla^2 f(x^k)^{1/2} n_{\mathrm{nt}}^k, H_k \nabla^2 f(x^k)^{1/2} n_{\mathrm{nt}}^k\rangle \\ &\leq \|H_k\|^2 \|n_{\mathrm{nt}}^k\|_{x^k}^2 = \|H_k\|^2 \lambda_k^2.\end{aligned}$$

By Lemma 2, we can estimate

$$\|H_k\| \leq R_\nu(d_\nu^k) d_\nu^k,$$

where R_ν is defined by (56). Combining the two last inequalities and using Proposition 8, we consider the following cases:

(a) If $\nu = 2$, then we have $\lambda_{k+1}^2 \leq e^{d_2^k}\left[\|\nabla f(x^{k+1})\|_{x^k}^*\right]^2$ 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 $d_2^k < d_2^\star \approx 0.12964$, 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 $\frac{\lambda_0}{\sqrt{\underline{\sigma}_0}} < \frac{1}{M_f}\min\left\{d_2^\star, 0.5\right\} = \frac{d_2^\star}{M_f}$, 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 $\left\{\frac{\lambda_k}{\sqrt{\underline{\sigma}_k}}\right\}$ converges to zero at a quadratic rate.

Now, if $\nu > 2$, then we consider different cases. Note that

$$\lambda_{k+1}^2 \le (1-d_\nu^k)^{\frac{-2}{\nu-2}} \left[\left\|\nabla f(x^{k+1})\right\|_{x^k}^*\right]^2,$$

which follows that

$$\lambda_{k+1} \le (1-d_\nu^k)^{\frac{-1}{\nu-2}} R_\nu(d_\nu^k) d_\nu^k \lambda_k. \tag{61}$$

Note that $d_\nu^k = \left(\frac{\nu}{2}-1\right) M_f \left\|d^k\right\|_2^{3-\nu} \lambda_k^{\nu-2}$ and $\underline{\sigma}_{k+1}^{-1} \le (1-d_\nu^k)^{\frac{-2}{\nu-2}} \underline{\sigma}_k^{-1}$. Based on these relations and (61) we can argue as follows:

(b) If $2 < \nu < 3$, then $\lambda_k \ge \left\|d^k\right\|_2 \sqrt{\underline{\sigma}_k}$ which follows that $d_\nu^k \le \left(\frac{\nu}{2}-1\right) M_f \underline{\sigma}_k^{-\frac{3-\nu}{2}} \lambda_k$. Hence,

$$\frac{\lambda_{k+1}}{\underline{\sigma}_{k+1}^{\frac{3-\nu}{2}}} \le (1-d_\nu^k)^{-\frac{4-\nu}{\nu-2}} R_\nu(d_\nu^k) \left(\frac{\nu}{2}-1\right) M_f \left(\frac{\lambda_k}{\underline{\sigma}_k^{\frac{3-\nu}{2}}}\right)^2.$$

If $d_\nu^k < d_\nu^\star$, where $d_\nu^\star$ is the unique solution to the equation

$$\left(\frac{\nu}{2}-1\right) \frac{R_\nu(d_\nu^k)}{(1-d_\nu^k)^{\frac{4-\nu}{\nu-2}}} = 2,$$

then $\underline{\sigma}_{k+1}^{-\frac{3-\nu}{2}} \lambda_{k+1} \le 2M_f \left(\underline{\sigma}_k^{-\frac{3-\nu}{2}} \lambda_k\right)^2$. Note that it is straightforward to check that this equation always admits a positive solution. Hence, if we choose $x^0 \in \mathrm{dom}(f)$ such that $\underline{\sigma}_0^{-\frac{3-\nu}{2}} \lambda_0 < \frac{1}{M_f} \min\left\{\frac{2d_\nu^\star}{\nu-2}, \frac{1}{2}\right\}$, then we can prove the following two inequalities together by induction:

$$d_\nu^k \le d_\nu^{k+1} \text{ and } \underline{\sigma}_{k+1}^{-\frac{3-\nu}{2}} \lambda_{k+1} \le \underline{\sigma}_k^{-\frac{3-\nu}{2}} \lambda_k.$$

In addition, the above inequality also shows that $\left\{\underline{\sigma}_k^{-\frac{3-\nu}{2}} \lambda_k\right\}$ quadratically converges to zero.

(c) If $\nu = 3$, then $d_3^k = \frac{M_f}{2}\lambda_k$, and

$$\lambda_{k+1} \le (1-d_3^k)^{-1} R_3(d_3^k) d_3^k \lambda_k = M_f \frac{R_3(d_3^k)}{2(1-d_3^k)} \lambda_k^2.$$

Directly checking the right-hand side of the above estimate, one can show that if $d_3^k < d_3^\star = 0.5$, then $\lambda_{k+1} \le 2M_f\lambda_k^2$. Hence, if $\lambda_0 < \frac{1}{M_f}\min\left\{2d_3^\star, 0.5\right\} = \frac{1}{2M_f}$, then we can prove the following two inequalities together by induction:

$$d_3^{k+1} \le d_3^k \text{ and } \lambda_{k+1} \le \lambda_k.$$

Moreover, the first inequality above also shows that $\{\lambda_k\}$ converges to zero at a quadratic rate.

The quadratic convergence of $\left\{\|x^k - x^\star_f\|_{H_k}\right\}$: First, using Proposition 9 with $x := x^k$ and $y = x^\star_f$, and noting that $\nabla f(x^\star_f) = 0$, we have

$$\bar{\omega}_\nu(-d_\nu(x^k, x^\star_f))\|x^k - x^\star_f\|^2_{x^k} \le \langle \nabla f(x^k), x^k - x^\star_f\rangle \le \|\nabla f(x^k)\|^*_{x^k}\|x^k - x^\star_f\|_{x^k},$$

where the last inequality follows from the Cauchy-Schwarz inequality. Hence, we obtain

$$\bar{\omega}_\nu(-d_\nu(x^k, x^\star_f))\|x^k - x^\star_f\|_{x^k} \le \|\nabla f(x^k)\|^*_{x^k} = \lambda_k. \tag{62}$$

We consider three cases:

(1) When $\nu = 2$, we have $\bar{\omega}_\nu(\tau) = \frac{e^\tau - 1}{\tau}$. Hence, $\bar{\omega}_\nu(-d_\nu(x^k, x^\star_f)) = \frac{1 - e^{-d_\nu(x^k, x^\star_f)}}{d_\nu(x^k, x^\star_f)} \ge 1 - \frac{d_\nu(x^k, x^\star_f)}{2} \ge \frac{1}{2}$ whenever $d_\nu(x^k, x^\star_f) \le 1$. Using this inequality in (62), we have $\|x^k - x^\star_f\|_{x^k} \le 2\|\nabla f(x^k)\|^*_{x^k} = 2\lambda_k$ provided that $d_\nu(x^k, x^\star_f) \le 1$. One the other hand, by the definition of $\underline{\sigma}_k$, we have $\sqrt{\underline{\sigma}_k}\|x^k - x^\star_f\|_2 \le \|x^k - x^\star_f\|_{x^k}$. Combining the two last inequalities, we obtain $\|x^k - x^\star_f\|_2 \le \frac{2\lambda_k}{\sqrt{\underline{\sigma}_k}}$ provided that $d_\nu(x^k, x^\star_f) \le 1$. Since $\left\{\frac{\lambda_k}{\sqrt{\underline{\sigma}_k}}\right\}$ locally converges to zero at a quadratic rate, the last relation also shows that $\left\{\|x^k - x^\star_f\|_2\right\}$ also locally converges to zero at a quadratic rate.

(2) For $\nu = 3$, we have $\bar{\omega}_\nu(-d_\nu(x^k, x^\star_f)) = \frac{1}{1 + d_\nu(x^k, x^\star_f)}$ and $d_\nu(x^k, x^\star_f) = \frac{M_f}{2}\|x^k - x^\star_f\|_{x^k}$. Hence, from (62), we obtain $\frac{\|x^k - x^\star_f\|_{x^k}}{1 + 0.5M_f\|x^k - x^\star_f\|_{x^k}} \le \lambda_k$. This implies $\|x^k - x^\star_f\|_{x^k} \le \frac{\lambda_k}{1 - 0.5M_f\lambda_k}$ as long as $0.5M_f\lambda_k < 1$. Clearly, since λ_k locally converges to zero at a quadratic rate, $\|x^k - x^\star_f\|_{x^k}$ also locally converges to zero at a quadratic rate.

(3) For $2 < \nu < 3$, we have $\bar{\omega}_\nu(-d_\nu(x^k, x^\star_f)) = \left(\frac{\nu-2}{\nu-4}\right)\frac{\left(1 + d_\nu(x^k, x^\star_f)\right)^{\frac{\nu-4}{\nu-2}} - 1}{d_\nu(x^k, x^\star_f)} \ge 1 - \frac{1}{\nu-2}d_\nu(x^k, x^\star_f) \ge \frac{1}{2}$ provided that $d_\nu(x^k, x^\star_f) < \frac{\nu}{2} - 1$. Similar to the case $\nu = 2$, we have $\underline{\sigma}_k^{\frac{3-\nu}{2}}\|x^k - x^\star_f\|_{H_k} \le \|x^k - x^\star_f\|_{x^k} \le 2\lambda_k$, where $H_k := \nabla^2 f(x^k)^{\frac{\nu-2}{2}}$. Hence, $\|x^k - x^\star_f\|_{H_k} \le \frac{2\lambda_k}{\underline{\sigma}_k^{\frac{3-\nu}{2}}}$. Since $\left\{\frac{\lambda_k}{\underline{\sigma}_k^{\frac{3-\nu}{2}}}\right\}$ locally converges to zero at a quadratic rate, $\left\{\|x^k - x^\star_f\|_{H_k}\right\}$ also locally converges to zero at a quadratic rate. □

A.7 The proof of Theorem 5: convergence of the damped-step PN method

Given $H \in \mathcal{S}^p_{++}$ and a proper, closed, and convex function $g : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$, we define

$$\mathcal{P}^g_H(u) := (H + \partial g)^{-1}(u) = \operatorname*{argmin}_x \left\{g(x) + \tfrac{1}{2}\langle Hx, x\rangle - \langle u, x\rangle\right\}.$$

If $H = \nabla^2 f(x)$ is the Hessian mapping of a strictly convex function f, then we can also write $\mathcal{P}_{\nabla^2 f(x)}(u)$ shortly as $\mathcal{P}_x(u)$ for our notational convenience. The following lemma will be used in the sequel whose proof can be found in [62].

Lemma 3 *Let $g : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ be a proper, closed, and convex function, and $H \in \mathcal{S}^p_{++}$. Then, the mapping $\mathcal{P}^g_H$ defined above is non-expansive with respect to the weighted norm defined by H, i.e., for any $u, v \in \mathbb{R}^p$, we have*

$$\left\| \mathcal{P}^g_H(u) - \mathcal{P}^g_H(v) \right\|_H \leq \|u - v\|^*_H . \tag{63}$$

Let us define

$$S_x(u) := \nabla^2 f(x)u - \nabla f(u) \quad \text{and} \quad e_x(u, v) := [\nabla^2 f(x) - \nabla^2 f(u)](v - u), \tag{64}$$

for any vectors $x, u \in \mathrm{dom}(f)$ and $v \in \mathbb{R}^p$. We now prove Theorem 5 in the main text.

The proof of Theorem 5 ***Computing the step-size*** τ_k: Since z^k satisfies the optimality condition (36), we have

$$-\nabla f(x^k) - \nabla^2 f(x^k) n^k_{\text{pnt}} \in \partial g(z^k).$$

Using Proposition 10 we obtain

$$f(x^{k+1}) \leq f(x^k) + \tau_k \left\langle \nabla f(x^k), n^k_{\text{pnt}} \right\rangle + \omega_\nu(\tau_k d_k)\tau_k^2 \lambda_k^2.$$

Since $x^{k+1} = (1 - \tau_k)x^k + \tau_k z^k$, using this relation and the convexity of g, we have

$$g(x^{k+1}) \leq g(x^k) - \tau_k \left\langle \nabla f(x^k) + \nabla^2 f(x^k) n^k_{\text{pnt}}, n^k_{\text{pnt}} \right\rangle.$$

Summing up the last two inequalities, we obtain the following estimate

$$F(x^{k+1}) \leq F(x^k) - \eta_k(\tau_k).$$

With the same argument as in the proof of Theorem 2, we obtain the conclusion of Theorem 5.

The proof of local quadratic convergence We consider the distance between x^{k+1} and $x^\star$ measured by $\|x^{k+1} - x^\star\|_{x^\star}$. By the definition of x^{k+1}, we have

$$\|x^{k+1} - x^\star\|_{x^\star} \leq (1 - \tau_k)\|x^k - x^\star\|_{x^\star} + \tau_k \|z^k - x^\star\|_{x^\star}. \tag{65}$$

Using the new notations in (64), it follows from the optimality condition (33) and (36) that $z^k = \mathcal{P}^g_{x^\star}(S_{x^\star}(x^k) + e_{x^\star}(x^k, z^k))$ and $x^\star = \mathcal{P}^g_{x^\star}(S_{x^\star}(x^\star))$. By Lemma 3 and the triangle inequality, we can show that

$$\|z^k - x^\star\|_{x^\star} \leq \|S_{x^\star}(x^k) - S_{x^\star}(x^\star)\|^*_{x^\star} + \|e_{x^\star}(x^k, z^k)\|^*_{x^\star}. \tag{66}$$

By following the same argument as in [62], if we apply Lemma 2, then we can derive

$$\|S_{x^\star}(x^k) - S_{x^\star}(x^\star)\|_{x^\star}^* \le R_\nu(d_\nu(x^\star, x^k))d_\nu(x^\star, x^k)\|x^k - x^\star\|_{x^\star}, \tag{67}$$

where $R_\nu(\cdot)$ is defined by (56).

Next, using the same argument as the proof of (72) in Theorem 6 below, we can bound the second term $\|e_{x^\star}(x^k, z^k)\|_{x^\star}^*$ of (66) as

$$\|e_{x^\star}(x^k, z^k)\|_{x^\star}^* \le \begin{cases} \left[(1 - d_\nu(x^\star, x^k))^{\frac{-2}{\nu-2}} - 1\right]\|z^k - x^k\|_{x^\star}, & \text{if } \nu > 2 \\ \left(e^{d_\nu(x^\star, x^k)} - 1\right)\|z^k - x^k\|_{x^\star} & \text{if } \nu = 2. \end{cases}$$

Combining this inequality, (66), (67), and the triangle inequality, we obtain

$$\begin{cases} \|z^k - x^k\|_{x^\star} \le \hat{R}_\nu(d_\nu(x^\star, x^k))\|x^k - x^\star\|_{x^\star}, \\ \|z^k - x^\star\|_{x^\star} \le \tilde{R}_\nu(d_\nu(x^\star, x^k))d_\nu(x^\star, x^k)\|x^k - x^\star\|_{x^\star}, \end{cases} \tag{68}$$

where $\hat{R}_\nu$ and $\tilde{R}_\nu$ are defined as

$$\hat{R}_\nu(t) := \begin{cases} \frac{tR_\nu(t)+1}{2-(1-t)^{\frac{-2}{\nu-2}}}, & \text{if } \ \nu > 2 \\ \frac{tR_\nu(t)+1}{2-e^t} & \text{if } \nu = 2 \end{cases} \quad \text{and} \quad \tilde{R}_\nu(t) := \begin{cases} \frac{tR_\nu(t)+(1-t)^{\frac{-2}{\nu-2}}-1}{t\left(2-(1-t)^{\frac{-2}{\nu-2}}\right)}, & \text{if } \ \nu > 2 \\ \frac{tR_\nu(t)+e^t-1}{t(2-e^t)} & \text{if } \nu = 2, \end{cases}$$

respectively. After a few 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 $\tilde{R}_\nu(t) \in [0, c_\nu]$ (when $t \to 0+$, consider the limit), where $\bar{d}_2 := \frac{3}{5}$ and $\bar{d}_\nu := 1 - \left(\frac{2}{3}\right)^{\frac{\nu-2}{2}}$ for $\nu > 2$, respectively. Using this bound, (65), (68), and the fact that $\tau_k \le 1$, we can bound

$$\|x^{k+1} - x^\star\|_{x^\star} \le \left[(1 - \tau_k) + c_\nu d_\nu(x^\star, x^k)\right]\|x^k - x^\star\|_{x^\star}. \tag{69}$$

Let $\underline{\sigma}^\star := \sigma_{\min}(\nabla^2 f(x^\star))$ be the smallest eigenvalue of $\nabla^2 f(x^\star)$. We consider the following cases:

(a) If $\nu = 2$, then, for $0 \le d_\nu(x^\star, x^k) \le \bar{d}_\nu$, we can bound $1 - \tau_k$ as

$$\begin{aligned} 1 - \tau_k &= 1 - \tfrac{\ln(1+\beta_k)}{\beta_k} \le \tfrac{\beta_k}{2} = \tfrac{M_f}{2}\|z^k - x^k\|_2 \\ &\le \tfrac{M_f}{2}\tfrac{\|z^k - x^k\|_{x^\star}}{\sqrt{\underline{\sigma}^\star}} \overset{(68)}{\le} \tfrac{c_\nu M_f}{2\sqrt{\underline{\sigma}^\star}}\|x^k - x^\star\|_{x^\star}. \end{aligned}$$

On the other hand, we have $d_\nu(x^\star, x^k) = M_f\|x^k - x^\star\|_2 \le \frac{M_f}{\sqrt{\underline{\sigma}^\star}}\|x^k - x^\star\|_{x^\star}$. Using these estimates into (69), we get

$$\begin{aligned}\|x^{k+1}-x^\star\|_{x^\star} &\leq \left(\frac{c_\nu M_f}{2\sqrt{\underline{\sigma}^\star}}\|x^k-x^\star\|_{x^\star}+\frac{c_\nu M_f}{\sqrt{\underline{\sigma}^\star}}\|x^k-x^\star\|_{x^\star}\right)\|x^k-x^\star\|_{x^\star}\\ &= \frac{3c_\nu M_f}{2\sqrt{\underline{\sigma}^\star}}\|x^k-x^\star\|_{x^\star}^2.\end{aligned}$$

Let $c_\nu^\star := \frac{3c_\nu M_f}{2\sqrt{\underline{\sigma}^\star}}$. The last estimate shows that if $\|x^0-x^\star\|_{x^\star} \leq \min\left\{\frac{\bar{d}_\nu\sqrt{\underline{\sigma}^\star}}{M_f}, \frac{1}{c_\nu^\star}\right\}$, then $\left\{\|x^k-x^\star\|_{x^\star}\right\}$ quadratically converges to zero.

(b) If $2 < \nu \leq 3$, then we first show that

$$d_\nu(x^\star, x^k) = \left(\tfrac{\nu}{2}-1\right)M_f\|x^k-x^\star\|_2^{3-\nu}\|x^k-x^\star\|_{x^\star}^{\nu-2} \leq \left(\tfrac{\nu}{2}-1\right)\frac{M_f}{\left(\underline{\sigma}^\star\right)^{\frac{3-\nu}{2}}}\|x^k-x^\star\|_{x^\star}.$$

Hence, if $\|x^k-x^\star\|_{x^\star} \leq m_\nu\bar{d}_\nu$, where $m_\nu := \frac{2}{\nu-2}\frac{(\underline{\sigma}^\star)^{\frac{3-\nu}{2}}}{M_f}$, then $d_\nu(x^\star, x^k) \leq \bar{d}_\nu$. Next, using the definition of d_k in (28), we can bound it as

$$\begin{aligned}d_k &= M_f\left(\tfrac{\nu}{2}-1\right)\|z^k-x^k\|_{x^k}^{\nu-2}\|z^k-x^k\|_2^{3-\nu}\\ &\overset{(15)}{\leq} M_f\left(\tfrac{\nu}{2}-1\right)\left[\frac{\|z^k-x^k\|_{x^\star}}{(1-d_\nu(x^\star,x^k))^{\frac{1}{\nu-2}}}\right]^{\nu-2}\frac{\|z^k-x^k\|_{x^\star}^{3-\nu}}{(\underline{\sigma}^\star)^{\frac{3-\nu}{2}}}\\ &\leq \frac{M_f}{(1-\bar{d}_\nu)(\underline{\sigma}^\star)^{\frac{3-\nu}{2}}}\left(\tfrac{\nu}{2}-1\right)\|z^k-x^k\|_{x^\star} \overset{(68)}{\leq} \frac{M_f(\nu-2)}{2(1-\bar{d}_\nu)(\underline{\sigma}^\star)^{\frac{3-\nu}{2}}}c_\nu\|x^k-x^\star\|_{x^\star}.\end{aligned}$$

Using this estimate, we can bound $1-\tau_k$ as follows:

$$\begin{aligned}1-\tau_k &= 1-\tfrac{1}{d_k}+\tfrac{1}{d_k}\left(1-\frac{\frac{4-\nu}{\nu-2}d_k}{1+\frac{4-\nu}{\nu-2}d_k}\right)^{\frac{\nu-2}{4-\nu}}\\ &\overset{\text{Bernoulli's inequality}}{\leq} 1-\tfrac{1}{d_k}+\tfrac{1}{d_k}\left(1-\frac{\nu-2}{4-\nu}\frac{\frac{4-\nu}{\nu-2}d_k}{1+\frac{4-\nu}{\nu-2}d_k}\right)\\ &= \frac{\frac{4-\nu}{\nu-2}d_k}{1+\frac{4-\nu}{\nu-2}d_k} \leq \frac{4-\nu}{\nu-2}d_k \leq \frac{M_f(4-\nu)}{2(1-\bar{d}_\nu)(\underline{\sigma}^\star)^{\frac{3-\nu}{2}}}c_\nu\|x^k-x^\star\|_{x^\star} = n_\nu\|x^k-x^\star\|_{x^\star},\end{aligned}$$

where $n_\nu := \frac{(4-\nu)M_f}{2(1-\bar{d}_\nu)(\underline{\sigma}^\star)^{\frac{3-\nu}{2}}}c_\nu > 0$. Substituting this estimate into (69) and noting that $d_\nu(x^\star, x^k) \leq \frac{1}{m_\nu}\|x^k-x^\star\|_{x^\star}$, we get

$$\|x^{k+1}-x^\star\|_{x^\star} \leq \left(n_\nu+\frac{c_\nu}{m_\nu}\right)\|x^k-x^\star\|_{x^\star}^2 := c_\nu^*\|x^k-x^\star\|_{x^\star}^2.$$

Hence, if $\|x^0-x^\star\|_{x^\star} \leq \min\left\{m_\nu\bar{d}_\nu, \frac{1}{c_\nu^\star}\right\}$, then the last estimate shows that the sequence $\left\{\|x^k-x^\star\|_{x^\star}\right\}$ quadratically converges to zero.

In summary, there exists a neighborhood $\mathcal{N}(x^\star)$ of $x^\star$, such that if $x^0 \in \mathcal{N}(x^\star) \cap \mathrm{dom}(F)$, then the whole sequence $\left\{\|x^k-x^\star\|_{x^\star}\right\}$ quadratically converges to zero. □

A.8 The proof of Theorem 6: locally quadratic convergence of the PN method

Since z^k is the optimal solution to (35) which satisfies (36), we have $\nabla^2 f(x^k)x^k - \nabla f(x^k) \in (\nabla^2 f(x^k) + \partial g)(z^k)$. Using this optimality condition, we get

$$\begin{aligned} x^{k+1} = z^k &= \mathcal{P}^g_{x^k}(S_{x^k}(x^k) + e_{x^k}(x^k, z^k)) \quad \text{and} \\ x^{k+2} = z^{k+1} &= \mathcal{P}^g_{x^k}(S_{x^k}(x^{k+1}) + e_{x^k}(x^{k+1}, z^{k+1})). \end{aligned}$$

Let us define $\tilde{\lambda}_{k+1} := \|n^{k+1}_{\text{pnt}}\|_{x^k}$. Then, by Lemma (3) and the triangular inequality, we have

$$\begin{aligned} \tilde{\lambda}_{k+1} &\le \left\|S_{x^k}(x^{k+1}) - S_{x^k}(x^k)\right\|^*_{x^k} + \left\|e_{x^k}(x^{k+1}, z^{k+1}) - e_{x^k}(x^k, z^k)\right\|^*_{x^k} \\ &= \left\|S_{x^k}(x^{k+1}) - S_{x^k}(x^k)\right\|^*_{x^k} + \left\|e_{x^k}(x^{k+1}, z^{k+1})\right\|^*_{x^k}. \end{aligned} \tag{70}$$

Let us first bound the term $\left\|S_{x^k}(x^{k+1}) - S_{x^k}(x^k)\right\|^*_{x^k}$ as follows:

$$\left\|S_{x^k}(x^{k+1}) - S_{x^k}(x^k)\right\|^*_{x^k} \le R_\nu(d^k_\nu)d^k_\nu\lambda_k, \tag{71}$$

where $R_\nu(t)$ is defined as (56). Indeed, from the mean-value theorem, we have

$$\begin{aligned} \left\|S_{x^k}(x^{k+1}) - S_{x^k}(x^k)\right\|^*_{x^k} &= \left\|\int_0^1 [\nabla^2 f(x^k + tn^k_{\text{pnt}}) - \nabla^2 f(x^k)]n^k_{\text{pnt}}\mathrm{d}t\right\|_{x^k} \\ &\le \left\|H(x^k, x^{k+1})\right\| \lambda_k, \end{aligned}$$

where H is defined as (54). Combining the above inequality and (56) in Lemma 2, we get (71).

Next we bound the term $\left\|e_{x^k}(x^{k+1}, z^{k+1})\right\|^*_{x^k}$ as follows:

$$\left\|e_{x^k}(x^{k+1}, z^{k+1})\right\|_{x^k} \le \begin{cases} \left[(1 - d^k_\nu)^{\frac{-2}{\nu-2}} - 1\right]\tilde{\lambda}_{k+1}, & \text{if } \nu > 2 \\ (e^{d^k_\nu} - 1)\tilde{\lambda}_{k+1} & \text{if } \nu = 2. \end{cases} \tag{72}$$

Note that

$$\begin{aligned} \left\|e_{x^k}(x^{k+1}, z^{k+1})\right\|^*_{x^k} &= \left\|[\nabla^2 f(x^k) - \nabla^2 f(x^{k+1})](z^{k+1} - x^{k+1})\right\|^*_{x^k} \\ &\le \|\widetilde{H}(x^k, x^{k+1})\|\tilde{\lambda}_{k+1}, \end{aligned}$$

where

$$\begin{aligned} \widetilde{H}(x, y) &:= \nabla^2 f(x)^{-1/2}\left(\nabla^2 f(x) - \nabla^2 f(y)\right)\nabla^2 f(x)^{-1/2} \\ &= \mathbb{I} - \nabla^2 f(x)^{-1/2}\nabla^2 f(y)\nabla^2 f(x)^{-1/2}. \end{aligned}$$

By Proposition 8, we have

$$\|\widetilde{H}(x,y)\| \leq \begin{cases} \max\left\{1-(1-d_\nu(x,y))^{\frac{2}{\nu-2}}, (1-d_\nu(x,y))^{\frac{-2}{\nu-2}}-1\right\}, & \text{if } \nu > 2 \\ \max\left\{1-e^{-d_\nu(x,y)}, e^{d_\nu(x,y)}-1\right\} & \text{if } \nu = 2. \end{cases}$$

This inequality can be simplified as

$$\|\widetilde{H}(x,y)\| \leq \begin{cases} (1-d_\nu(x,y))^{\frac{-2}{\nu-2}}-1, & \text{if } \nu > 2 \\ e^{d_\nu(x,y)}-1 & \text{if } \nu = 2. \end{cases} \tag{73}$$

Hence, the inequality (72) holds.

Now, we combine (70), (71), and (72), if $\nu = 2$, and assuming that $d_2^k < \ln 2$, then we get

$$\tilde{\lambda}_{k+1} \leq \frac{R_2(d_2^k)d_2^k}{2-e^{d_2^k}}\lambda_k.$$

By Proposition 8, we have $\lambda_{k+1}^2 \leq e^{d_\nu^k}\tilde{\lambda}_{k+1}^2$. Combining this estimate and the last inequality, we get

$$\lambda_{k+1} \leq \frac{R_2(d_2^k)d_2^k e^{\frac{d_2^k}{2}}}{2-e^{d_2^k}}\lambda_k. \tag{74}$$

Note that $\lambda_k \geq \frac{\sqrt{\underline{\sigma}_k}d_2^k}{M_f}$ and $\underline{\sigma}_{k+1}^{-1} \leq e^{d_2^k}\underline{\sigma}_k^{-1}$. It follows from (74) that

$$\frac{\lambda_{k+1}}{\sqrt{\underline{\sigma}_{k+1}}} \leq M_f\frac{R_2(d_2^k)e^{d_2^k}}{2-e^{d_2^k}}\left(\frac{\lambda_k}{\sqrt{\underline{\sigma}_k}}\right)^2.$$

By a numerical calculation, we can check that if $d_2^k \leq d_2^\star \approx 0.35482$, then

$$\frac{\lambda_{k+1}}{\sqrt{\underline{\sigma}_{k+1}}} \leq 2M_f\left(\frac{\lambda_k}{\sqrt{\underline{\sigma}_k}}\right)^2.$$

Hence, if we choose $x^0 \in \mathrm{dom}(F)$ such that $\frac{\lambda_0}{\sqrt{\underline{\sigma}_0}} \leq \frac{1}{M_f}\min\left\{d_2^\star, 0.5\right\} = \frac{d_2^\star}{M_f}$, then we can prove the following two inequalities together by induction:

$$d_2^{k+1} \leq d_2^k \quad \text{and} \quad \frac{\lambda_{k+1}}{\sqrt{\underline{\sigma}_{k+1}}} \leq \frac{\lambda_k}{\sqrt{\underline{\sigma}_k}}.$$

These inequalities show the nonincreasing monotonicity of $\left\{d_2^k\right\}$ and $\{\lambda_k\}$. The above inequality also shows the local quadratic convergence of the sequence $\left\{\frac{\lambda_k}{\sqrt{\underline{\sigma}_k}}\right\}$.

Now, if $\nu > 2$ and assume that $d_\nu^k < 1 - \left(\frac{1}{2}\right)^{\frac{\nu-2}{2}}$, then

$$\tilde{\lambda}_{k+1} \le \frac{R_\nu(d_\nu^k)d_\nu^k}{2 - (1-d_\nu^k)^{\frac{-2}{\nu-2}}}\lambda_k.$$

By Proposition 8, we have $\lambda_{k+1}^2 \le (1-d_\nu^k)^{\frac{-2}{\nu-2}}\tilde{\lambda}_{k+1}^2$. Hence, combining these inequalities, we get

$$\lambda_{k+1} \le \frac{R_\nu(d_\nu^k)d_\nu^k(1-d_\nu^k)^{\frac{-1}{\nu-2}}}{2-(1-d_\nu^k)^{\frac{-2}{\nu-2}}}\lambda_k. \tag{75}$$

Note that $d_\nu^k = \left(\frac{\nu}{2}-1\right) M_f \left\|p^k\right\|_2^{3-\nu}\lambda_k^{\nu-2}$, $\underline{\sigma}_{k+1}^{-1} \le (1-d_\nu^k)^{\frac{-2}{\nu-2}}\underline{\sigma}_k^{-1}$ and $\sigma_{k+1}^{-1} \le (1-d_\nu^k)^{\frac{-2}{\nu-2}}\sigma_k^{-1}$. Using these relations and (75), we consider two cases:

(a) If $\nu = 3$, then $d_3^k = \frac{M_f}{2}\lambda_k$, and

$$\lambda_{k+1} \le \frac{R_3(d_3^k)(1-d_3^k)^{-1}}{2-(1-d_3^k)^{-2}}d_3^k\lambda_k = M_f\frac{R_3(d_3^k)(1-d_3^k)^{-1}}{2\left(2-(1-d_3^k)^{-2}\right)}\lambda_k^2.$$

By a simple numerical calculation, we can show that if $d_3^k \le d_3^\star \approx 0.20943$, then $\lambda_{k+1} \le 2M_f\lambda_k^2$. Hence, if $\lambda_0 < \frac{1}{M_f}\min\left\{2d_3^\star, 0.5\right\} = \frac{2}{M_f}d_3^\star$, then we can prove the following two inequalities together by induction

$$d_3^{k+1} \le d_3^k \text{ and } \lambda_{k+1} \le \lambda_k.$$

These inequalities show the non-increasing monotonicity of $\left\{d_2^k\right\}$ and $\{\lambda_k\}$. The above inequality also shows the quadratic convergence of the sequence $\{\lambda_k\}$.

(b) If $2 < \nu < 3$, then $\lambda_k \ge \|p^k\|_2\sqrt{\underline{\sigma}_k}$ which implies that $d_\nu^k \le \left(\frac{\nu}{2}-1\right)M_f\underline{\sigma}_k^{-\frac{3-\nu}{2}}\lambda_k$. Hence, we have

$$\frac{\lambda_{k+1}}{\underline{\sigma}_{k+1}^{\frac{3-\nu}{2}}} \le \frac{R_\nu(d_\nu^k)(1-d_\nu^k)^{-\frac{4-\nu}{\nu-2}}}{2-(1-d_\nu^k)^{\frac{-2}{\nu-2}}}\left(\frac{\nu}{2}-1\right)M_f\left(\frac{\lambda_k}{\underline{\sigma}_k^{\frac{3-\nu}{2}}}\right)^2.$$

If $d_\nu^k < d_\nu^\star$, then $\underline{\sigma}_{k+1}^{-\frac{3-\nu}{2}}\lambda_{k+1} \le 2M_f\left(\underline{\sigma}_k^{-\frac{3-\nu}{2}}\lambda_k\right)^2$, where $d_\nu^\star$ is the unique solution to the equation

$$\frac{R_\nu(d_\nu^k)(1-d_\nu^k)^{-\frac{4-\nu}{\nu-2}}}{2-(1-d_\nu^k)^{\frac{-2}{\nu-2}}}\left(\frac{\nu}{2}-1\right) = 2.$$

Note that it is straightforward to check that this equation always admits a positive solution. Therefore, if $\underline{\sigma}_0^{-\frac{3-\nu}{2}}\lambda_0 \le \frac{1}{M_f}\min\left\{\frac{2d_\nu^\star}{\nu-2}, \frac{1}{2}\right\}$, then we can prove the following two inequalities together by induction:

$$d_\nu^k \le d_\nu^{k+1} \text{ and } \underline{\sigma}_{k+1}^{-\frac{3-\nu}{2}}\lambda_{k+1} \le \underline{\sigma}_k^{-\frac{3-\nu}{2}}\lambda_k.$$

These inequalities show the non-increasing monotonicity of $\left\{d_2^k\right\}$ 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 $\left\{x^k\right\}$ to $x^\star$, we use the same argument as in the proof of Theorem 3 and Theorem 5, where we omit the details here. □

A.9 The proof of Theorem 7: convergence of the quasi-Newton method

The full-step quasi-Newton method for solving (24) can be written as $x^{k+1} = x^k - B_k\nabla f(x^k)$. This is equivalent to $H_k(x^{k+1} - x^k) + \nabla f(x^k) = 0$. Using this relation and $\nabla f(x_f^\star) = 0$, we can write

$$\begin{aligned} x^{k+1}-x_f^\star = \nabla^2 f(x_f^\star)^{-1}\Big[&\nabla^2 f(x_f^\star)(x^k-x_f^\star)+\left(\nabla^2 f(x_f^\star)-H_k\right)(x^{k+1}-x^k)\\ &-\nabla f(x^k)+\nabla f(x_f^\star)\Big]. \end{aligned} \tag{76}$$

We first consider $T_k := \|\nabla^2 f(x_f^\star)^{-1}\left[\nabla f(x^k) - \nabla f(x_f^\star) - \nabla^2 f(x_f^\star)(x^k - x_f^\star)\right]\|_{x_f^\star}$. Similar to the proof of Theorem 3, we can show that

$$\begin{aligned} T_k &= \left\| \int_0^1 \nabla^2 f(x_f^\star)^{-1}\left[\nabla^2 f(x_f^\star+t(x^k-x_f^\star))-\nabla^2 f(x_f^\star)\right](x^k-x_f^\star)\right\|_{x_f^\star}\\ &\le R_\nu(d_\nu^k)d_\nu^k\|x^k-x_f^\star\|_{x_f^\star} \end{aligned} \tag{77}$$

where R_ν is defined by (56) and $d_\nu^k := M_f^{\nu-2}\left(\frac{\nu}{2}-1\right)\|x^k - x_f^\star\|_2^{3-\nu}\|x^k - x_f^\star\|_{x_f^\star}^{\nu-2}$. Moreover, we note that

$$\begin{aligned} S_k &:= \|\nabla^2 f(x_f^\star)^{-1}\left(H_k - \nabla^2 f(x_f^\star)\right)(x^{k+1}-x^k)\|_{x_f^\star}\\ &= \|\left(H_k - \nabla^2 f(x^\star)\right)(x^{k+1}-x^k)\|_{x_f^\star}^* \end{aligned}$$

Combining this estimate, (76), and (77), we can derive

$$\|x^{k+1} - x_f^\star\|_{x_f^\star} \le R_\nu(d_\nu^k)d_\nu^k\|x^k - x_f^\star\|_{x_f^\star} + \|\left(H_k - \nabla^2 f(x_f^\star)\right)(x^{k+1}-x^k)\|_{x_f^\star}^*. \tag{78}$$

First, we prove statement (a). Indeed, from the Dennis–Moré condition (41), we have

$$\begin{aligned} \|\left(H_k - \nabla^2 f(x_f^\star)\right)(x^{k+1}-x^k)\|_{x_f^\star}^* &\le \gamma_k\|x^{k+1} - x_k\|_{x_f^\star}\\ &\le \gamma_k\left(\|x^{k+1} - x_f^\star\|_{x_f^\star} + \|x^k - x_f^\star\|_{x_f^\star}\right), \end{aligned}$$

where $\lim_{k\to\infty}\gamma_k = 0$. Substituting this estimate into (78), and noting that $\|x^k - x^\star_f\|_2 \leq \frac{1}{\underline{\sigma}^\star}\|x^k - x^\star_f\|_{x^\star_f}$, where $\underline{\sigma}^\star := \lambda_{\min}(\nabla^2 f(x^\star_f)) > 0$, we can show that

$$\|x^{k+1} - x^\star_f\|_{x^\star_f} \leq \frac{1}{1-\gamma_k}\left(R^\star_\nu\|x^k - x^\star_f\|^2_{x^\star_f} + \gamma_k\|x^k - x^\star_f\|_{x^\star_f}\right), \tag{79}$$

provided that $\|x^k - x^\star_f\|_{x^\star_f} \leq \bar{r}$ and $R^\star_\nu := \max\left\{R_\nu(d^k_\nu) \mid \|x^k - x^\star_f\|_{x^\star_f} \leq \bar{r}\right\} < +\infty$. Here, $\bar{r} > 0$ is a given value such that $R^\star_\nu$ is finite. The estimate (79) shows that if $\bar{r}$ is sufficiently small, $\left\{\|x^k - x^\star_f\|_{x^\star_f}\right\}$ superlinearly converges to zero. Finally, the statement (b) is proved similarly by combining statement (a) and [62, Theorem 11]. □

References

1. Bach, F.: Self-concordant analysis for logistic regression. Electron. J. Stat. **4**, 384–414 (2010)
2. Bach, F.: Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. J. Mach. Learn. Res. **15**(1), 595–627 (2014)
3. Bauschke, H.H., Combettes, P.: Convex Analysis and Monotone Operators Theory in Hilbert Spaces, 2nd edn. Springer, Berlin (2017)
4. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding agorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
5. Becker, S., Candès, E.J., Grant, M.: Templates for convex cone problems with applications to sparse signal recovery. Math. Program. Comput. **3**(3), 165–218 (2011)
6. Becker, S., Fadili, M.J.: A quasi-Newton proximal splitting method. In: Proceedings of Neutral Information Processing Systems Foundation (NIPS) (2012)
7. Bollapragada, R., Byrd, R., Nocedal, J.: Exact and inexact subsampled Newton methods for optimization (2016). arXiv preprint arXiv:1609.08502
8. Bonnans, J.F.: Local analysis of Newton-type methods for variational inequalities and nonlinear programming. Appl. Math. Optim. **29**, 161–186 (1994)
9. Borodin, A., El-Yaniv, R., Gogan, V.: Can we learn to beat the best stock. J. Artif. Intell. Res. (JAIR) **21**, 579–594 (2004)
10. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
11. Byrd, R.H., Hansen, S.L., Nocedal, J., Singer, Y.: A stochastic quasi-Newton method for large-scale optimization. SIAM J. Optim. **26**(2), 1008–1031 (2016)
12. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 27:1–27:27 (2011). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
13. Chen, T.-Y., Demmel, J.W.: Balancing sparse matrices for computing eigenvalues. Linear Algebra Appl. **309**(1–3), 261–287 (2000)
14. Cohen, M., Madry, A., Tsipras, D., Vladu, A.: Matrix scaling and balancing via box constrained Newton's method and interior-point methods. The 58th Annual IEEE Symposium on Foundations of Computer Science, pp. 902–913 (2017)
15. Dennis, J.E., Moré, J.J.: A characterisation of superlinear convergence and its application to quasi-Newton methods. Math. Comput. **28**, 549–560 (1974)
16. Deuflhard, P.: Newton Methods for Nonlinear Problems—Affine Invariance and Adaptative Algorithms, volume 35 of Springer Series in Computational Mathematics, 2nd edn. Springer, Berlin (2006)
17. Dolan, E.D., Moré, J.J.: Benchmarking optimization software with performance profiles. Math. Program. **91**, 201–213 (2002)
18. Erdogdu, M.A., Montanari, A.: Convergence rates of sub-sampled Newton methods. In: Advances in Neural Information Processing Systems, pp. 3052–3060 (2015)
19. Fercoq, O., Qu, Z.: Restarting accelerated gradient methods with a rough strong convexity estimate (2016). arXiv preprint arXiv:1609.07358
20. Frank, M., Wolfe, P.: An algorithm for quadratic programming. Naval Res. Logist. Q. **3**, 95–110 (1956)

21. Friedlander, M., Goh, G.: Efficient evaluation of scaled proximal operators. Electron. Trans. Numer. Anal. **46**, 1–22 (2017)
22. Gao, W., Goldfarb, D.: Quasi-Newton methods: superlinear convergence without line search for self-concordant functions (2016). arXiv preprint arXiv:1612.06965
23. Giselsson, P., Boyd, S.: Monotonicity and restart in fast gradient methods. In: IEEE Conference on Decision and Control, Los Angeles, USA, December 2014, pp. 5058–5063. CDC
24. Goel, V., Grossmann, I.E.: A class of stochastic programs with decision dependent uncertainty. Math. Program. **108**, 355–394 (2006)
25. Grant, M., Boyd, S., Ye, Y.: Disciplined convex programming. In: Liberti, L., Maculan, N. (eds.) Global Optimization From Theory to Implementation, Nonconvex Optimization and its Applications, pp. 155–210. Springer, Berlin (2006)
26. Halko, N., Martinsson, P.-G., Tropp, J.A.: Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions (2009)
27. Hazan, E., Arora, S.: Efficient Algorithms for Online Convex Optimization and their Applications. Princeton University, Princeton (2006)
28. He, N., Harchaoui, Z., Wang, Y., Song, L.: Fast and simple optimization for Poisson likelihood models (2016). arXiv preprint arXiv:1608.01264
29. Hosmer, D.W., Lemeshow, S.: Applied Logistic Regression. Wiley, New York (2005)
30. Jaggi, M.: Revisiting Frank–Wolfe: projection-free sparse convex optimization. JMLR W&CP **28**(1), 427–435 (2013)
31. Krishnapuram, B., Figueiredo, M., Carin, L., Hartemink, H.: Sparse multinomial logistic regression: fast algorithms and generalization bounds. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **27**, 957–968 (2005)
32. Kyrillidis, A., Karimi, R., Tran-Dinh, Q., Cevher, V.: Scalable sparse covariance estimation via self-concordance. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence, pp. 1946–1952 (2014)
33. Lebanon, G., Lafferty, J.: Boosting and maximum likelihood for exponential models. Adv. Neural Inf. Process. Syst. (NIPS) **14**, 447 (2002)
34. Lee, J.D., Sun, Y., Saunders, M.A.: Proximal Newton-type methods for convex optimization. SIAM J. Optim. **24**(3), 1420–1443 (2014)
35. Lu, Z.: Randomized block proximal damped Newton method for composite self-concordant minimization. SIAM J. Optim. **27**(3), 1910–1942 (2016)
36. Marron, J.S., Todd, M.J., Ahn, J.: Distance-weighted discrimination. J. Am. Stat. Assoc. **102**(480), 1267–1271 (2007)
37. McCullagh, P., Nelder, J.A.: Generalized Linear Models, vol. 37. CRC Press, Boca Raton (1989)
38. Monteiro, R.D.C., Sicre, M.R., Svaiter, B.F.: A hybrid proximal extragradient self-concordant primal barrier method for monotone variational inequalities. SIAM J. Optim. **25**(4), 1965–1996 (2015)
39. Nelder, J.A., Baker, R.J.: Generalized Linear Models. Encyclopedia of Statistical Sciences. Wiley, New York (1972)
40. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course, volume 87 of Applied Optimization. Kluwer Academic Publishers, Dordrecht (2004)
41. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. **103**(1), 127–152 (2005)
42. Nesterov, Y.: Cubic regularization of Newton's method for convex problems with constraints. CORE Discussion Paper 2006/39, Catholic University of Louvain (UCL) - Center for Operations Research and Econometrics (CORE) (2006)
43. Nesterov, Y.: Accelerating the cubic regularization of Newtons method on convex problems. Math. Program. **112**, 159–181 (2008)
44. Nesterov, Y.: Gradient methods for minimizing composite objective function. Math. Program. **140**(1), 125–161 (2013)
45. Nesterov, Y., Nemirovski, A.: Interior-point Polynomial Algorithms in Convex Programming. Society for Industrial Mathematics, Philadelphia (1994)
46. Nesterov, Y., Polyak, B.T.: Cubic regularization of Newton's method and its global performance. Math. Program. **112**(1), 177–205 (2006)
47. Nocedal, J., Wright, S.J.: Numerical Optimization, Springer Series in Operations Research and Financial Engineering, 2nd edn. Springer, Berlin (2006)
48. O'Donoghue, B., Candes, E.: Adaptive restart for accelerated gradient schemes. Found. Comput. Math. **15**, 715–732 (2015)

49. Odor, G., Li, Y.-H., Yurtsever, A., Hsieh, Y.-P., Tran-Dinh, Q., El-Halabi, M., Cevher, V.: Frank–Wolfe works for non-Lipschitz continuous gradient objectives: scalable Poisson phase retrieval. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6230–6234. IEEE (2016)
50. Ortega, J.M., Rheinboldt, W.C.: Iterative Solution of Nonlinear Equations in Several Variables. Society for Industrial and Applied Mathematics, Philadelphia (2000)
51. Parikh, N., Boyd, S.: Proximal algorithms. Found. Trends Optim. **1**(3), 123–231 (2013)
52. Parlett, B.N., Landis, T.L.: Methods for scaling to doubly stochastic form. Linear Algebra Appl. **48**, 53–79 (1982)
53. Peng, J., Roos, C., Terlaky, T.: Self-Regularity. A New Paradigm for Primal-Dual Interior-Point Algorithms. Princeton University Press, Princeton (2009)
54. Pilanci, M., Wainwright, M.J.: Newton sketch: a linear-time optimization algorithm with linear-quadratic convergence (2015). Arxiv preprint arXiv:1505.02250
55. Polyak, R.A.: Regularized Newton method for unconstrained convex optimization. Math. Program. **120**(1), 125–145 (2009)
56. Robinson, S.M.: Strongly regular generalized equations. Math. Oper. Res. **5**(1), 43–62, **5**:43–62 (1980)
57. Roosta-Khorasani, F., Mahoney, M.W.: Sub-sampled Newton methods I: globally convergent algorithms (2016). arXiv preprint arXiv:1601.04737
58. Roosta-Khorasani, F., Mahoney, M.W.: Sub-sampled Newton methods II: local convergence rates (2016). arXiv preprint arXiv:1601.04738
59. Ryu, E.K., Boyd, S.: Stochastic proximal iteration: a non-asymptotic improvement upon stochastic gradient descent. Author website, early draft (2014)
60. Toh, K.-Ch., Todd, M.J., Tütüncü, R.H.: On the implementation and usage of SDPT3—a Matlab software package for semidefinite-quadratic-linear programming. Tech. Report 4, NUS Singapore (2010)
61. Tran-Dinh, Q., Kyrillidis, A., Cevher, V.: A proximal Newton framework for composite minimization: graph learning without Cholesky decompositions and matrix inversions. JMLR W&CP **28**(2), 271–279 (2013)
62. Tran-Dinh, Q., Kyrillidis, A., Cevher, V.: Composite self-concordant minimization. J. Mach. Learn. Res. **15**, 374–416 (2015)
63. Tran-Dinh, Q., Li, Y.-H., Cevher, V.: Composite convex minimization involving self-concordant-like cost functions. In: Pham Dinh, T., Le-Thi, H., Nguyen, N. (eds.) Modelling, Computation and Optimization in Information Systems and Management Sciences, pp. 155–168. Springer, New York (2015)
64. Tran-Dinh, Q., Necoara, I., Diehl, M.: A dual decomposition algorithm for separable nonconvex optimization using the penalty function framework. In: Proceedings of the Conference on Decision and Control (CDC), Florence, Italy, December, pp. 2372–2377 (2013)
65. Tran-Dinh, Q., Necoara, I., Diehl, M.: Path-following gradient-based decomposition algorithms for separable convex optimization. J. Global Optim. **59**(1), 59–80 (2014)
66. Tran-Dinh, Q., Necoara, I., Savorgnan, C., Diehl, M.: An inexact perturbed path-following method for Lagrangian decomposition in large-scale separable convex optimization. SIAM J. Optim. **23**(1), 95–125 (2013)
67. Tran-Dinh, Q., Sun, T., Lu, S.: Self-concordant inclusions: a unified framework for path-following generalized Newton-type algorithms. Technical Report (submitted) (2016)
68. Vapnik, V.N., Vapnik, V.: Statistical Learning Theory, vol. 1. Wiley, New York (1998)
69. Verscheure, D., Demeulenaere, B., Swevers, J., De Schutter, J., Diehl, M.: Time-optimal path tracking for robots: a convex optimization approach. IEEE Trans. Autom. Control **54**, 2318–2327 (2009)
70. Yamashita, M., Fujisawa, K., Kojima, M.: Implementation and evaluation of SDPA 6.0 (SemiDefinite Programming Algorithm 6.0). Optim. Method Softw. **18**, 491–505 (2003)
71. Yang, T., Lin, Q.: RSG: beating SGD without smoothness and/or strong convexity. CoRR abs/1512.03107 (2016)
72. Zhang, Y., Lin, X.: DiSCO: Distributed optimization for self-concordant empirical loss. In: Proceedings of the 32th International Conference on Machine Learning, pp. 362–370 (2015)