Some Theoretical Aspects on Newton's Method for Constrained Best Interpolation[1]

Houduo Qi
School of Mathematics
The University of Southampton
Highfield, Southampton SO17 1BJ
Great Britain
E-mail: hdqi@soton.ac.uk

September 2, 2004

Abstract. The paper contains new results as well as surveys on recent developments on the constrained best interpolation problem, and in particular on the convex best interpolation problem. Issues addressed include theoretical reduction of the problem to a system of nonsmooth equations, nonsmooth analysis of those equations and development of Newton's method, convergence analysis and globalization. We frequently use the convex best interpolation to illustrate the seemingly complex theory. Important techniques such as splitting are introduced and interesting links between approaches from approximation and optimization are also established. Open problems related to polyhedral constraints and strips may be tackled by the tools introduced and developed in this paper.

AMS subject classifications. 49M45, 90C25, 90C33

[1]The work was done while the author was with School of Mathematics, The University of New South Wales, Australia, and was supported by Australian Research Council.

1 Introduction

The convex best interpolation problem is defined as follows:

$$\text{minimize} \quad \|f''\|_2 \tag{1}$$

$$\text{subject to} \qquad f(t_i) = y_i, \quad i = 1, 2, \cdots, n+2,$$
$$f \text{ is convex on } [a,b], \quad f \in W^{2,2}[a,b],$$

where $a = t_1 < t_2 < \ldots < t_{n+2} = b$ and y_i, $i = 1, \ldots, n+2$ are given numbers, $\|\cdot\|_2$ is the Lebesgue $L^2[a,b]$ norm, and $W^{2,2}[a,b]$ denotes the Sobolev space of functions with absolutely continuous first derivatives and second derivatives in $L^2[a,b]$, and equipped with the norm being the sum of the $L^2[a,b]$ norms of the function, its first, and its second derivatives.

Using an integration by parts technique, Favard [22] and, more generally, de Boor [9] showed that this problem has an equivalent reformulation as follows:

$$\min\left\{\|u\| \mid u \in L^2[a,b],\ u \geq 0,\ \langle u, x^i\rangle = d_i,\ i = 1, \ldots, n\right\}, \tag{2}$$

where the functions $x^i \in L^2[a,b]$ and the numbers d_i can be expressed in terms of the original data $\{t_i, y_i\}$ (in fact, $x^i = B_i(t)$, the B-spline of order 2 defined by the given data and $\{d_i\}$ are the second divided differences of $\{(t_i, y_i)\}_{i=1}^{n+2}$). Under the assumption $d_i > 0$, $i = 1, \ldots, n$ the optimal solution u^* of (2) has the form

$$u^*(t) = \left(\sum_{i=1}^{n} \lambda_i^* B_i(t)\right)_+ \tag{3}$$

where $\tau_+ := \max\{0, \tau\}$ and $\{\lambda_i^*\}$ satisfy the following interpolation condition:

$$\int_a^b \left(\sum_{i=1}^{n} \lambda_i B_i(t)\right)_+ B_i(t)dt = d_i, \qquad i = 1, \ldots, n. \tag{4}$$

Once we have the solution u^*, the function required by (1) can be obtained by $f'' = u$. This representation result was obtained first by Hornung [25] and subsequently extended to a much broader circle of problems in [1, 16, 26, 27, 37, 38]. We briefly discuss below both theoretically and numerically important progresses on those problems.

Theoretically, prior to [38] by Micchelli and Utreras, most of research is mainly centered on the problem (1) and its slight relaxations such as f'' is bounded below or above, see [26, 37, 27, 1, 16]. After [38] the main focus is on to what degree the solution characterization like (3) and (4) can be extended to a more general problem proposed in Hilbert spaces:

$$\min\left\{\frac{1}{2}\|x - x^0\|^2 \mid x \in C \text{ and } Ax = b\right\} \tag{5}$$

where $C \subset X$ is a closed convex set in a Hilbert space X, $A : X \mapsto \mathbb{R}^n$ is a bounded linear operator, $b \in \mathbb{R}^n$. It is easy to see that if we let

$$X = L^2[a,b],\ C = \{x \in X | x \geq 0\},\ Ax = (\langle B_1, x\rangle, \ldots, \langle B_n, x\rangle),\ x^0 = 0,\ b = d \tag{6}$$

then (5) becomes (2). The abstract interpolation problem (5), initially studied in [38], was extensively studied in a series of papers by Chui, Deutsch, and Ward [7, 8], Deutsch, Ubhaya, Ward, and Xu [14], and Deutsch, Li, and Ward [12]. For the complete treatment on this problem in the spirit of those papers, see the recent book by Deutsch [11].

Among the major developments in those papers is an important concept called the *strong CHIP* [12], which is the refinement of the property CHIP [7] (Conical Hull Intersection Property). More studies on the strong CHIP, CHIP and other properties can be found in the two recent papers [4, 3]. Roughly speaking, the importance of the strong CHIP is with the following characterization result: The strong CHIP holds for the constraints in (5) if and only the unique solution x^* has the following representation:

$$x^* = P_C(x^0 + A^*\lambda^*), \tag{7}$$

where P_C denotes the projection to the closed convex set C (the closeness and convexity guarantees the the existence of P_C), and A^* is the adjoint of A, and $\lambda^* \in \mathbb{R}^n$ satisfies the following nonlinear nonsmooth equation:

$$AP_C(x^0 + A^*\lambda) = b. \tag{8}$$

To see (7) and (8) recover (3) and (4) it is enough to use the fact:

$$P_C = x_+ \ \text{ where } C = \{x \in L^2[a,b] | x \geq 0\}.$$

If the strong CHIP does not hold we still have similar characterization in which P_C is replaced by P_{C_b}, where C_b is an extremal face of C satisfying some properties [11]. However, it is often hard to get enough information to make the calculation of P_{C_b} possible, unless in some particular cases. Hence, we mainly focus on the case where the strong CHIP holds. We will see that the assumption $d_i > 0$, $i = 1, \ldots, n$ for problem (1) is a sufficient condition for the strong CHIP, and much more than that, it ensures the quadratic convergence of Newton's method.

Numerically, problem (1) has been well studied [26, 27, 1, 38, 16, 18, 19]. As demonstrated in [27] and verified in several other occasions [1, 16], the Newton method is the most efficient compared to many other global methods for solving the equation (4). We delay the description of the Newton method to the end of Section 3, instead we list some difficulties in designing algorithms for (4) and (8). First of all, the equation (4) is generally nonsmooth. The nonsmoothness was a major barrier for Andersson and Elfving [1] to establish the convergence of Newton's method (they have to assume that the equation is smooth near the solution (the simple case) in order that the classical convergence result of Newton's method applies). Second, as having been both noticed in [27, 1], in the simple (i.e., smooth) case, the method presented in [27, 1] becomes the classical Newton method. More justification is needed to consolidate the name and the use of Newton's method when the equation is nonsmooth. To do this, we appeal to the theory of the generalized Newton method developed by Kummer [33] and Qi and Sun [40] for nonsmooth equations. This was done in [18, 19]. We will review this theory in Section 3. Third, Newton's method is only developed for the conical case, i.e., C is a cone. It is yet to know in what form the Newton method appear even for the polyhedral case (i.e., C is intersection of finitely many halfspaces). We will tackle those difficulties against the problem (5).

The problem (5) can also be studied via a very different approach developed by Borwein and Lewis [5] for partially finite convex programming problems:

$$\inf \{ f(x) |\ Ax \in b + Q, \quad x \in C \}, \tag{9}$$

where $C \in X$ is a closed convex set, X is a topological vector space, $A : X \mapsto \mathbb{R}^n$ is a bounded linear operator, $b \in \mathbb{R}^n$, Q is a polyhedral set in $\mathbb{R}^n$, and $f : X \mapsto (-\infty, \infty]$ is convex. If $f(x) = \|x^0 - x\|^2$ and $Q = \{0\}$, then (9) becomes (5). Under the constraint qualification that there is a feasible point which is in the quasi-relative interior of C, the problem (9) can be solve by its Fenchel-Rockafellar dual problem. We will see in the next section that this approach also leads to the solution characterization (7) and (8). See, e.g., [24, 28, 30] for further development of Borwein-Lewis approach.

An interesting aspect of (9) is when $Q = \mathbb{R}^n_+$, the nonnegative orthant of $\mathbb{R}^n$. This yields the following approximation problem:

$$\min \left\{ \frac{1}{2} \|x^0 - x\|^2 |\ Ax \geq b,\ x \in C \right\}. \tag{10}$$

This problem was systematically studied by Deutsch, Li and Ward in [12], proving that the strong CHIP again plays an important role but the sufficient condition ensuring the strong CHIP takes a very different form from that (i.e., $b \in \mathrm{ri}\ AC$) for (5). We will prove in Section 2 that the constraint qualification of Borwein and Lewis also implies the strong CHIP. Nonlinear convex and nonconvex extension of (10) can be found in [34, 35, 36].

The paper is organized as follows: The next section contains some necessary background materials. In particular, we review the approach initiated by Micchelli and Utreras [38] and all the way to the advent of the strong CHIP and its consequences. We then review the approach of Borwein and Lewis [5] and state its implications by establishing the fact that the nonemptiness of the quasi-relative interior of the feasible set implies the strong CHIP. In section 3, we review the theory of Newton's method for nonsmooth equations, laying down the basis for the analysis of the Newton method for (5), which is conducted in Section 4. In the last section, we discuss some extensions to other problems such as interpolation in a strip. Throughout the paper we use the convex best interpolation problem (1) and (4) as an example to illustrate the seemingly complex theory.

2 Constrained Interpolation in Hilbert Space

Since X is a Hilbert space, the bounded linear operator $A : X \mapsto \mathbb{R}^n$ has the following representation: there exist $x_1, \ldots, x_n \in X$ such that

$$Ax = (\langle x_1, x \rangle, \ldots, \langle x_n, x \rangle), \quad \forall x \in X.$$

Defining

$$H_i := \{ x \in X |\ \langle x_i, x \rangle = b_i \}, \quad i = 1, \ldots, n$$

the interpolation problem (5) has the following appearance

$$\min\left\{\frac{1}{2}\|x^0 - x\|^2 \mid x \in K := C \cap \left(\cap_{j=1}^n H_j\right)\right\}. \tag{11}$$

Recall that for any convex set $D \subset X$, the (negative) polar of D, denoted by D°, is defined by

$$D^\circ := \{y \in X \mid \langle y, x\rangle \le 0,\ \forall x \in D\}.$$

The well-know strong CHIP is now defined as follows.

Definition 2.1 *[11, Definition 10.2] A collection of closed convex sets $\{C_1, C_2, \ldots, C_m\}$ in X, which has a nonempty intersection, is said to have the strong conical hull intersection property, or the strong CHIP, if*

$$(\cap_1^m C_i - x)^\circ = \sum_1^m (C_i - x)^\circ \qquad \forall x \in \cap_1^m C_i.$$

The concept of the strong CHIP is a refinement of CHIP [12], which requires

$$(\cap_1^m C_i - x)^\circ = \overline{\sum_1^m (C_i - x)^\circ} \qquad \forall x \in \cap_1^m C_i, \tag{12}$$

where $\bar{C}$ denotes the closure of C. It is worth of mentioning that one direction of (12) is automatic, that is

$$(\cap_1^m C_i - x)^\circ \supseteq \overline{\sum_1^m (C_i - x)^\circ} \qquad \forall x \in \cap_1^m C_i.$$

Hence, the strong CHIP is actually assuming the other direction. The importance of the strong CHIP is with the following solution characterization of the problem (11).

Theorem 2.2 *[12, Theorem 3.2] and [11, Theorem 10.13] The set $\{C, \cap_1^n H_j\}$ has the strong CHIP if and only if for every $x^0 \in X$ there exists $\lambda^* \in \mathbb{R}^n$ such that the optimal solution $x^* = P_K(x^0)$ has the representation:*

$$x^* = P_C(x^0 + A^*\lambda^*)$$

and λ^* satisfies the interpolation equation

$$AP_C(x^0 + A^*\lambda) = b.$$

We remark that in general the strong CHIP of the sets $\{C, H_1, \ldots, H_n\}$ implies the strong CHIP of the sets $\{C, \cap_1^n H_j\}$. The following lemma gives a condition that ensures their equivalence.

Lemma 2.3 *[11, Lemma 10.11] Suppose that X is a Hilbert space and $\{C_0, C_1, \ldots, C_m\}$ is a collection of closed convex subsets such that $\{C_1, \ldots, C_m\}$ has the strong CHIP. Then the following statements are equivalent:*

(i) $\{C_0, C_1, \dots, C_m\}$ *has the strong CHIP.*

(ii) $\{C_0, \cap_1^m C_j\}$ *has the strong CHIP.*

Since each H_j is a hyperplane, $\{H_1, \dots, H_n\}$ has the strong CHIP [11, Example 10.9]. It follows from Lemma 2.3 that the strong CHIP of $\{C, H_1, \dots, H_n\}$ is equivalent to that of $\{C, A^{-1}(b)\}$. However, it is often difficult to know if $\{C, A^{-1}(b)\}$ has the strong CHIP. Fortunately, there are available easy-to-be-verified sufficient conditions for this property. Given a convex subset $D \subset \mathbb{R}^n$, let ri D denote the relative of D. Note that ri $D \neq \emptyset$ if $D \neq \emptyset$.

Theorem 2.4 *[11, Theorem 10.32] and [12, Theorem 3.12] If* $b \in$ *ri* AC*, then* $\{C, A^{-1}(b)\}$ *has the strong CHIP.*

Theorem 2.4 also follows from the approach of Borwein and Lewis [5]. The concept of quasi-relative interior of convex sets plays an important role in this approach. We assume temporarily that X be a locally convex topological vector space. Let X^* denote the dual space of X (if $X$is a Hilbert space then $X^* = X$) and $N_C(\hat{x}) \subset X^*$ denote the normal cone to C at $\hat{x} \in C$, i.e.,

$$N_C(\hat{x}) := \{y \in X^* |\ \langle y, x - \hat{x} \rangle \leq 0, \quad \forall x \in C\}.$$

The most useful properties of the quasi-relative interiors are contained in the following

Proposition 2.5 *[5] Suppose* $C \subset X$ *is convex, then*

(i) If X *is finite-dimensional then qri* $C =$ *ri* C.

(ii) Let $\hat{x} \in C$ *then* $\hat{x} \in$ *qri* C *if and only if* $N_C(\hat{x})$ *is a subspace of* X^*.

(iii) Let $A : X \mapsto \mathbb{R}^n$ *be a bounded linear map. If qri* $C \neq \emptyset$ *then* $A($*qri* $C) =$ *ri* AC.

We note that (ii) serves a definition for the quasi-relative interior of convex sets. One can find several other interesting properties of the quasi-relative interior in [5]. Although in finite-dimensional case quasi-relative interior becomes classical relative interior, it is a genuine new concept in infinite-dimensional cases. To see this, let $X = L^p[0,1]$, $(p \geq 1)$, $C := \{x \in X | x \geq 0 \ a.e.\}$. Since C reproduces X (i.e., $X = C - C$), ri $C = \emptyset$, however, qri $C = \{x \in X | x > 0 \ a.e.\}$. One of the basic results in [5] is

Theorem 2.6 *[5, Corollary 4.8] Let the assumptions on problem (9) hold. Consider its dual problem*

$$\max\left\{-(f + \delta(\cdot|C))^*(A^*\lambda) + b^T\lambda |\ \lambda \in Q^+\right\}. \tag{13}$$

If the following constraint qualification is satisfied

$$\text{there exists an } \hat{x} \in \text{qri } C \text{ which is feasible for (9)}, \tag{14}$$

then the value of (9) and (13) are equal with attainment in (13). Suppose further that $(f + \delta(\cdot|C))$ *is closed. If* λ^* *is optimal for the dual and* $(f + \delta(\cdot|C))^*$ *is differentiable at* $A^*\lambda^*$ *with Gateaux derivative* $x^* \in X$*, then* x^* *is optimal for (9) and furthermore the unique optimal solution.*

In (13), $Q^+ := \{y \in X^* | \ \langle y, x\rangle \geq 0, \ \forall\, x \in Q\}$. We now apply Theorem 2.6 to problem (11), i.e., we let

$$f(x) = \frac{1}{2}\|x^0 - x\|^2, \ Q = \{0\} \text{ so that } Q^+ = \mathbb{R}^n.$$

Obviously, in this case (9) has a unique solution since $f(x)$ is strongly convex. For $y \in X^*$ we calculate

$$\begin{aligned}
(f + \delta(\cdot|C))^*(y) &= \sup_{x\in X}\{\langle y, x\rangle - f(x) - \delta(x|C)\} \\
&= \sup_{x\in C}\left\{\langle y, x\rangle - \frac{1}{2}\|x - x^0\|^2\right\} \\
&= \sup_{x\in C}\left\{\langle x, y + x^0\rangle - \frac{1}{2}\|x\|^2 - \frac{1}{2}\|x^0\|^2\right\} \\
&= \sup_{x\in C}\left\{\frac{1}{2}\|y + x^0\|^2 - \frac{1}{2}\|y + x^0 - x\|^2 - \frac{1}{2}\|x^0\|^2\right\} \\
&= \frac{1}{2}\|y + x^0\|^2 - \frac{1}{2}\|y + x^0 - P_C(y + x^0)\|^2 - \frac{1}{2}\|x^0\|^2. \qquad (15)
\end{aligned}$$

It is well known (see, e.g., [38, Theorem 3.2]) that the right side of (15) is Gateaux differentiable with

$$\nabla(f + \delta(\cdot|C))^*(y) = P_C(y + x^0).$$

Returning to (13), which is an unconstrained convex optimization problem, we know that the optimal solution λ^* to (13) satisfies

$$AP_C(x^0 + A^*\lambda) = b$$

and the optimal solution to (9) is

$$x^* = P_C(x^0 + A^*\lambda^*).$$

Following Theorem 2.2 we see that the sets $\{C, A^{-1}(b)\}$ has the strong CHIP. In fact, the qualification (14) is exactly the condition $b \in \text{ri}\ (AC)$ by Proposition 2.5, except that (14) needs a priori assumption qri $C \neq \emptyset$.

However, for the problem (10), where

$$K = C \cap \{x | Ax \geq b\},$$

the condition $b \in \text{ri}\ AC$ is not suitable as it might happen that $b \notin AC$. It turns out that the strong CHIP again plays an essential role in this case. Let

$$\mathcal{H}_j := \{x | \langle a_j, x\rangle \geq b_j\}.$$

Theorem 2.7 *[13, Theorem 3.2] The sets $\{C, \cap_1^n \mathcal{H}_j\}$ has the strong CHIP if and only if the optimal solution of (10) $x^* = P_K(x^0)$ has the following representation:*

$$x^* = P_C(x^* + A^*\lambda^*), \qquad (16)$$

where λ^* is any solution of the nonlinear complementarity problem:

$$\lambda \geq 0, \ w := AP_C(x^0 + A^*\lambda) - b \geq 0, \ \lambda^T w = 0. \qquad (17)$$

The following question was raised in [13] that if the constraint qualification (14) is a sufficient condition for the strong CHIP of $\{C, \cap_1^n \mathcal{H}_j\}$. We give an affirmative answer in the next result.

Theorem 2.8 *If it holds*

$$qri\ C \cap (\cap_1^n \mathcal{H}_j) \neq \emptyset, \tag{18}$$

then the sets $\{C, \cap_1^n \mathcal{H}_j\}$ *has the strong CHIP.*

Proof. Suppose (18) is in place, it follows from Theorem 2.6 with $f(x) = \frac{1}{2}\|x - x^0\|^2$ that there exists an optimal solution λ^* to the problem (13). (15) says that

$$(f + \delta(\cdot|C))^*(y) = \frac{1}{2}\|y + x^0\|^2 - \frac{1}{2}\|y + x^0 - P_C(y + x^0)\|^2 - \frac{1}{2}\|x^0\|^2$$

and it is Gateaux differentiable and convex [38, Lemma 3.1]. Then (13) becomes

$$\min \left\{ \frac{1}{2}\|A^*\lambda + x^0\|^2 - \frac{1}{2}\|A^*\lambda + x^0 - P_C(A^*\lambda + x^0)\|^2 - b^T\lambda |\ \lambda \geq 0 \right\}.$$

It is a finite-dimensional convex optimization problem and the optimal solution is attained. Hence, the optimal solution λ^* is exactly a solution of (17) and the optimal solution of (10) is $x^* = P_C(x^0 + A^*\lambda)$. It then follows from the characterization in Theorem 2.7 that the sets $\{C, \cap_1^n \mathcal{H}_j\}$ has the strong CHIP. $\square$

Illustration to problem (2). We recall the problem (2) and the setting in (6). From the fact [5, Lemma 7.17]

$$\left\{ \left(\int_a^b B_i x dt \right)_1^n |\ x > 0 \text{ a.e. } x \in L^2[a, b] \right\} = \{ r \in \mathbb{R}^n |\ r_i > 0, i = 1, \ldots, n \}$$

and the fact qri $C = \{x \in L^2[a, b] |\ x > 0 \text{ a.e.}\}$, we have

$$A\text{qri } C = \text{ri } AC = \text{int } AC = \{ r \in \mathbb{R}^n |\ r_i > 0, i = 1, \ldots, n \}.$$

It follows from Theorem 2.4 or Theorem 2.6 that the solution to (2) is given by (3) and (4), under the assumption that $d_i > 0$ for all i. Moreover, we will see that this assumption implies the uniqueness of the solution λ^*, and eventually guarantees the quadratic convergence of the Newton method.

3 Nonsmooth Functions and Equations

As is well known, if $F : \mathbb{R}^n \mapsto \mathbb{R}^n$ is smooth the classical Newton method for finding solution x^* of the equation $F(x) = 0$ takes the following form:

$$x^{k+1} = x^k - \left(F'(x^k) \right)^{-1} F(x^k) \tag{19}$$

where F' is the Jacobian of F. If $F'(x^*)$ is nonsingular then (19) is well defined near the solution x^* and is quadratically convergent. However, as we see from the previous sections we are encountered with nonsmooth equations. There is need to develop Newton's method for nonsmooth equation, which is presented below.

Now we suppose that $F : \mathbb{R}^n \mapsto \mathbb{R}^n$ is only locally Lipschitz and we want to find a solution of the equation

$$F(x) = 0. \tag{20}$$

Since F is differentiable almost everywhere according to Redemacher's theorem, the Bouligand differential of F at x, denoted by $\partial_B F(x)$, is defined by

$$\partial_B F(x) := \left\{ V | \; V = \lim_{x^i \to x} F'(x^i), \; F \text{ is differentiable at } x^i \right\}.$$

In other words, $\partial_B F(x)$ is the set of all limits of any sequence $\{F'(x^i)\}$ where F' exists at x^i and $x^i \to x$. The generalized Jacobian of Clark [6] is then the convex hull of $\partial_B F(x)$, i.e.,

$$\partial F(x) = \text{co } \partial_B F(x).$$

The basic properties of ∂F are included in the following result.

Proposition 3.1 *[6, Proposition 2.6.2]*

(a) ∂F is a nonempty convex compact subset of $\mathbb{R}^{n \times n}$.

(b) ∂F is closed at x; that is, if $x^i \to x$, $M_i \in \partial F(x^i)$, $M_i \to M$, then $M \in \partial F(x)$.

(c) ∂F is upper semicontinuous at x.

Having the object of ∂F, the nonsmooth version of Newton's method for the solution of (20) can be described as follows (see, e.g., [33, 40]).

$$x^{k+1} = x^k - V_k^{-1} F(x^k), \quad V_k \in \partial F(x^k). \tag{21}$$

We note that different choice of V_k results in different sequence of $\{x^k\}$. Hence, it is more accurate to say that (21) defines a class of Newton-type methods rather than a single method. It is always arguable which element in $\partial F(x^k)$ is the most suitable in defining (21). We will say more about the choice with regard to the convex best interpolation problem. We also note that there are other ways in defining nonsmooth Newton's method, essentially using different definitions $\partial F(x)$, but servicing the same objective as ∂F, see, e.g., [29, 43, 32].

Definition 3.2 *We say that F is regular at x if each element in $\partial F(x)$ is nonsingular.*

If F is regular at x^* it follows from the upper semicontinuity of F at x^* (Prop. 3.1) that F is regular near x^*, and consequently, (21) is well defined near x^*. Contrasted to the smooth case, the regularity at x^* only is no long a sufficient condition for the convergence of the method (21). It turns out that its convergence also relies on another important property of F, named the *semismoothness*.

Definition 3.3 *[40] We say that F is semismooth at x^* if the following conditions hold:*

(i) F is directionally differentiable at x, and

(ii) it holds

$$F(x+h) - F(x) - Vh = o(\|h\|) \qquad \forall V \in \partial F(x+h) \text{ and } h \in \mathbb{R}^n. \tag{22}$$

Furthermore, if

$$F(x+h) - F(x) - Vh = O(\|h\|^2) \qquad \forall V \in \partial F(x+h) \text{ and } h \in \mathbb{R}^n, \tag{23}$$

F is said strongly semismooth at x. If F is (strongly) semismooth everywhere, we simply say that F is (strongly) semismooth.

The property of semismoothness, as introduced by Mifflin [39] for functionals and scalar-valued functions and further extended by Qi and Sun [40] for vector-valued functions, is of particular interest due to the key role it plays in the superlinear convergence of the nonsmooth Newton method (21). It is worth of mentioning that in a largely ignored paper [33] by Kummer, the relation (22), being put in a very general form in [33], has been revealed to be essential for the convergence of a class of Newton type methods, which is essentially the same as (21). Nevertheless, Qi and Sun's work [40] makes it more accessible to and much easier to use by many researchers (see, e.g., the book [21] by Facchinei and Pang). The importance of the semismoothness can be seen from the following convergence result for (21).

Theorem 3.4 *[40, Theorem 3.2] Let x^* be a solution of the equation $F(x) = 0$ and let F be a locally Lipschitz function which is semismooth at x^*. Assume that F is regular at x^*. Then every sequence generated by the method (21) is superlinearly convergent to x^* provided that the starting point x^0 is sufficiently close to x^*. Furthermore, if F is strongly semismooth at x^*, then the convergence rate is quadratic.*

The use of Theorem 3.4 relies on the availability of the following three elements: (a) availability of an element in $\partial F(x)$ near the solution x^*, (b) regularity of F at x^* and, (c) (strong) semismoothness of F at x^*. We illustrate how the first can be easily calculated below for the convex best interpolation problem and leave the other two tasks to the next section.

Illustration to the convex best interpolation problem. It follow from (3) and (4) that the solution of the convex best interpolation problem can be obtained by solving the following equation:

$$F(\lambda) = d, \tag{24}$$

where $d = (d_1, \ldots, d_n)^T$ and each component of F is given by

$$F_j(\lambda) = \int_a^b \left(\sum_{\ell=1}^n \lambda_\ell B_\ell \right)_+ B_j(t)dt, \quad j = 1, \ldots, n. \tag{25}$$

Irvine, Marin, and Smith [27] developed Newton's method for (24):

$$\lambda^+ = \lambda - (M(\lambda))^{-1}(F(\lambda) - d), \tag{26}$$

where λ and λ^+ denote respectively the old and the new iterate, and $M(\lambda) \in \mathbb{R}^{n\times n}$ is given by

$$(M(\lambda))_{ij} = \int_a^b \left(\sum_{\ell=1}^n \lambda_\ell B_\ell\right)_+^0 B_i(t)B_j(t)dt,$$

and

$$(\tau)_+^0 = \begin{cases} 1 & \text{if } \tau > 0 \\ 0 & \text{if } \tau \le 0. \end{cases}$$

Let e denote the element of all ones in $\mathbb{R}^n$, then it is easy to see that the directional derivative of F at λ along the direction e is

$$F'(\lambda, e) = M(\lambda)e.$$

Moreover, if F is differentiable at λ then $F'(\lambda) = M(\lambda)$. Due to those reasons, the iteration (26) was then called Newton's method, and based on extensive numerical experiments, was observed quadratically convergent in [27]. Independent of [27], partial theoretical results on the convergence of (26) was established by Andersson and Elfving [1]. Complete convergence analysis was established by Dontchev, Qi, and Qi [18, 19] by casting (26) as a particular instance of (21). The convergence analysis procedure verifies exactly the availability of the three elements discussed above, in particular, $M(\lambda) \in \partial F(\lambda)$. We will present in the next section the procedure on the constrained interpolation problem in Hilbert space.

4 Newton's Method and Convergence Analysis

4.1 Newton's Method

We first note that all results in Section 2 assume no other requirements for the set C except being convex and closed. Consequently, we are able to develop (conceptual, at least) Newton's method for the nonsmooth equation (8). However, efficient implementation of Newton's method relies on the assumption that there is an efficient way to calculate the generalized Jacobian of $AP_C(x)$. The most interesting case due to this consideration is when C is a closed convex cone (i.e., the conical case [5]), which covers many problems including (1). We recall our setting below

$$X = L^2[a,b],\ C = \{x \in X | x \ge 0\}, Ax = (\langle a_1, x\rangle, \ldots, \langle a_n, x\rangle),\ b \in \mathbb{R}^n$$

where $a_\ell \in X$, $\ell = 1, \ldots, n$ (in fact we may assume that $X = L^p[a,b]$, in this case $a_\ell \in L^q[a,b]$ where $1/p + 1/q = 1$). This setting simplifies our description.

We want to develop Newton's method for the equation:

$$AP_C(x^0 + A^*\lambda) = b.$$

Taking into account of the fact $P_C(x) = x_+$, we let

$$F(\lambda) - b = 0 \tag{27}$$

where each component of $F : \mathbb{R}^n \mapsto \mathbb{R}^n$ is given by

$$F_j(\lambda) := \langle a_j, (x^0 + \sum_{\ell=1}^{n} a_\ell \lambda_\ell)_+ \rangle. \tag{28}$$

We propose a nonsmooth Newton method (in the spirit of Section 3) for nonsmooth equation (27) as follows:

$$V(\lambda)(\lambda^+ - \lambda) = b - F(\lambda), \qquad V(\lambda) \in \partial F(\lambda). \tag{29}$$

One of several difficulties with the Newton method (29) is to select an appropriate matrix $V(\lambda)$ from $\partial F(\lambda)$, which is well defined as F is Lipschitz continuous under Assumption 4.1 stated later. We will also see the following choice satisfies all the requirements.

$$(V(\lambda))_{ij} := \int_a^b \left(x^0 + \sum_{\ell=1}^{n} \lambda_\ell a_\ell \right)_+^0 a_i a_j dt. \tag{30}$$

We note that for $\beta \in \mathbb{R}^n$

$$\beta^T V(\lambda) \beta = \int_a^b \left(x^0 + \sum_{\ell=1}^{n} \lambda_\ell a_\ell \right)_+^0 \left(\sum_{\ell=1}^{n} \beta_\ell a_\ell \right)^2 dt \geq 0. \tag{31}$$

That is, $V(\lambda)$ is positive semidefinite for arbitrary choice $\lambda \in \mathbb{R}^n$. We need an assumption to make it positive definite. Let the support of a_ℓ be

$$\text{supp}(a_\ell) := \{t \in [a,b] | a_\ell(t) \neq 0\}.$$

Assumption 4.1 *Each a_ℓ is continuous, and any subset of functions*

$$\{a_\ell, \ell \in \mathcal{I} \subseteq \{1, \ldots, n\} | supp(a_i) \cap supp(a_j) \neq \emptyset \text{ for any pair } i, j \in \mathcal{I}\},$$

are linearly independent on $\cup_{\ell \in \mathcal{I}} supp(a_\ell)$. Moreover,

$$\cup_{\ell=1}^{n} supp(a_\ell) = [a,b].$$

This assumption is not restrictive. Typical choices of a_ℓ are $\{a_i = t^i\}$ or $\{a_i = B_i\}$. With Assumption 4.1 we have the following result.

Lemma 4.2 *Suppose Assumption 4.1 holds. $V(\lambda)$ is positive definite if and only if $(x^0 + \sum_{\ell=1}^{n} \lambda_\ell a_\ell)_+$ does not vanish identically on the supporting set of each a_ℓ, $\ell = 1, \ldots, n$.*

Proof. Suppose that $(x^0 + \sum_{\ell=1}^n \lambda_\ell a_\ell)_+$ is nonzero on each $\mathrm{supp}(a_\ell)$. Due to the continuity of $(x^0 + \sum_{\ell=1}^n \lambda_\ell a_\ell)$ and a_ℓ, there exists a Borel set $\Omega_\ell \subseteq \mathrm{supp}(a_\ell)$ such that $(x^0 + \sum_{\ell=1}^n \lambda_\ell a_\ell)_+^0 = 1$ for all $t \in \Omega_\ell$ and the measure of Ω_ℓ is not zero. Let

$$\mathcal{I}(\Omega_\ell) := \{j |\ \mathrm{supp}(a_j) \cap \Omega_\ell \neq \emptyset\}.$$

Since $\{a_j | j \in \mathcal{I}(\Omega_\ell)\}$ are linearly independent, $\beta^T V(\lambda)\beta = 0$ implies $\beta_j = 0$ for all $j \in \mathcal{I}(\Omega_\ell)$. We also note that

$$\cup_{\ell=1}^n \mathcal{I}(\Omega_\ell) = \{1, \dots, n\}.$$

We see that $\beta_j = 0$ for all $j = 1, \dots, n$ if $\beta^T V(\lambda)\beta = 0$. Hence, (31) yields the positive definiteness of $V(\lambda)$. The converse follows from the observation that if $(x^0 + \sum_{\ell=1}^n \lambda_\ell a_\ell)_+ \equiv 0$ on $\mathrm{supp}(a_\ell)$ for some ℓ then $\beta^T V(\lambda)\beta = 0$ for $\beta \in \mathbb{R}^n$ with $\beta_\ell = 1$ and $\beta_j = 0$ for $j \neq \ell$. $\square$

Due to the special structure of $V(\lambda)$, Newton's method (29) can be simplified by noticing that

$$\begin{aligned} F_j(\lambda) &= \int_a^b \left(x^0 + \sum_{\ell=1}^n \lambda_\ell a_\ell\right)_+ a_j dt \\ &= \int_a^b \left(x^0 + \sum_{\ell=1}^n \lambda_\ell a_\ell\right)_+^0 \left(x^0 + \sum_{\ell=1}^n \lambda_\ell a_\ell\right)_+ a_j dt \\ &= \sum_{\ell=1}^n \lambda_\ell (V(\lambda))_{j\ell} + \int_a^b \left(x^0 + \sum_{\ell=1}^n \lambda_\ell a_\ell\right)_+^0 a_j x^0 d(t). \end{aligned}$$

Thus we have

$$F(\lambda) = V(\lambda)\lambda + A\left(\left(x^0 + \sum_{\ell=1}^n \lambda_\ell a_\ell\right)_+^0 x^0\right).$$

Recalling (29) we have

$$V(\lambda)\lambda^+ = b - A\left(\left(x^0 + \sum_{\ell=1}^n \lambda_\ell a_\ell\right)_+^0 x^0\right). \tag{32}$$

A very interesting case is when $x^0 = 0$, which implies that no function evaluations are required to implement Newton's method, i.e, (32) takes the form $V(\lambda)\lambda^+ = b$.

Other choices of $V(\lambda)$ are also possible as $\partial F(\lambda)$ usually contains infinitely many elements. For example,

$$\left(\tilde{V}(\lambda)\right)_{ij} := \int_a^b \left((x^0 + \sum_{\ell=1}^n \lambda_\ell a_\ell\right)_-^0 a_i a_j dt, \quad \text{and} \quad (\tau)_-^0 := \begin{cases} 1 & \text{if } \tau \geq 0 \\ 0 & \text{if } \tau < 0. \end{cases}$$

It is easy to see that $\beta^T \tilde{V}(\lambda)\beta \geq \beta^T V(\lambda)\beta$ for any $\beta \in \mathbb{R}^n$. This means that $\tilde{V}(\lambda)$ "increases the positivity" of $V(\lambda)$ in the sense that $\tilde{V}(\lambda) - V(\lambda)$ is positive semidefinite. The argument leading to (32) also applies to $\tilde{V}(\lambda)$. We will show below that both $V(\lambda)$ and $\tilde{V}(\lambda)$ are contained in $\partial F(\lambda)$.

4.2 Splitting and Regularity

We now introduce a splitting technique that decomposes the (nonsmooth) function F into two parts, namely F^+ and F^-, satisfying that F^+ is continuously differentiable at the given point and F^- is necessarily nonsmooth nearby. This technique facilitates our arguments that lead to the conclusion that $V(\lambda)$ belongs to $\partial F(\lambda)$ and pave the ways to study the regularity of F at the solution. For the moment, we let $\bar{\lambda}$ be our reference point. Let

$$T(\bar{\lambda}) := \{t \in [a,b] | \; x^0 + \sum_{\ell=1}^{n} \bar{\lambda}_\ell a_\ell = 0\}, \qquad \bar{T}(\bar{\lambda}) := [a,b] \setminus T(\bar{\lambda}).$$

Due to Assumption (4.1), $T(\bar{\lambda})$ contains closed intervals in $[a,b]$, possibly isolated points. For $j = 1, \ldots, n$, define

$$F_j^+(\lambda) := \int_{\bar{T}(\bar{\lambda})} \left(x^0 + \sum_{\ell=1}^{n} \lambda_\ell a_\ell \right)_+ a_j dt, \quad F_j^-(\lambda) := \int_{T(\bar{\lambda})} \left(x^0 + \sum_{\ell=1}^{n} \lambda_\ell a_\ell \right)_+ a_j dt,$$

and

$$F^+(\lambda) := (F_1^+(\lambda), \ldots, F_n^+(\lambda))^T, \quad F^-(\lambda) := (F_1^-(\lambda), \ldots, F_n^-(\lambda))^T.$$

It is easy to see that

$$F(\lambda) = F^+(\lambda) + F^-(\lambda).$$

It is elementary to see that the vector-valued function F^+ is continuous differentiable in a neighborhood $\mathcal{N}(\bar{\lambda})$ of $\bar{\lambda}$. Then from the definition of the generalized Jacobian we obtain that for any $\lambda \in \mathcal{N}(\bar{\lambda})$,

$$\partial F(\lambda) = \nabla F^+(\lambda) + \partial F^-(\lambda), \tag{33}$$

where $\nabla F^+(\lambda)$ denotes the usual Jacobian of F^+ at λ. More precisely,

$$\left(\nabla F^+(\bar{\lambda}) \right)_{ij} = \int_{\bar{T}(\bar{\lambda})} \left(x^0 + \sum_{\ell=1}^{n} \bar{\lambda} a_\ell \right)_+^0 a_i a_j dt. \tag{34}$$

Since

$$x^0 + \sum_{\ell=1}^{n} \bar{\lambda} a_\ell = 0 \quad \text{ for all } \; t \in T(\bar{\lambda}),$$

(34) can be written as

$$\left(\nabla F^+(\bar{\lambda}) \right)_{ij} = \int_a^b \left(x^0 + \sum_{\ell=1}^{n} \bar{\lambda} a_\ell \right)_+^0 a_i a_j dt = V(\bar{\lambda}). \tag{35}$$

Regarding to F^- we need following assumption:

Assumption 4.3 *There exists a sequence of* $\{\lambda^k\}$ *in* $\mathbb{R}^n$ *converging to* zero *such that the sum* $\sum_{\ell=1}^{n} \lambda_\ell^k a_\ell$ *is negative on* $[a,b]$ *for all* λ^k.

This assumption also holds if each of a_ℓ is nonnegative or nonpositive.

Lemma 4.4 *For any $\lambda \in \mathbb{R}^n$ every element in $\partial F^-(\lambda)$ is positive semidefinite. Moreover, if Assumption 4.3 holds then the zero matrix belongs to $\partial F^-(\bar\lambda)$.*

Proof. We denote

$$y := \left(x^0 + \sum_{\ell=1}^{n} \lambda_\ell a_\ell\right)_+ \chi_{T(\bar\lambda)},$$

where $\chi_{T(\bar\lambda)}$ is the characteristic function of the set $T(\bar\lambda)$. In terms of y, F^- can be written as $F^-(\lambda) = Ay$. Since $T(\bar\lambda)$ consists of only closed intervals, without loss of generality we assume $T(\bar\lambda)$ is a closed interval. Let

$$C := \left\{x \in L^2(T(\bar\lambda)) |\, x \geq 0\right\}.$$

Then we have $L^2[a,b] \subseteq L^2(T(\bar\lambda))$ since $(T(\bar\lambda)) \subseteq [a,b]$. Define

$$\theta(\lambda) := \int_{T(\bar\lambda)} \left(x^0 + \sum_{\ell=1}^{n} \lambda_\ell a_\ell\right)_+^2 dt = \int_{T(\bar\lambda)} \left(P_C(x^0 + \sum_{\ell=1}^{n} \lambda_\ell a_\ell)\right)^2 dt.$$

According to [38, Lemma 2.1], $\theta(\lambda)$ is continuously Gateaux differentiable and convex. Moreover,

$$\nabla\theta(\lambda) = Ay = F^-(\lambda).$$

Therefore, any matrix in the generalized Jacobian of the gradient mapping (which is required to be Lipschitz continuous) of a convex function must be positive semidefinite, see, for example, [31, Proposition 2.3]. Now we prove the second part. Suppose Assumption 4.3 holds for the sequence $\{\lambda^k\}$ which converges to *zero*. Then $F^-(\bar\lambda + \lambda^k)$ is differentiable because

$$\sum_{\ell=1}^{n} (\bar\lambda + \lambda^k)_\ell a_\ell < 0 \quad \text{for all } t \in T(\bar\lambda) \text{ and } \tau > 0.$$

Hence,

$$\lim_{k\to\infty} \nabla F^-(\bar\lambda + \lambda^k) = 0 \in \partial F^-(\bar\lambda).$$

□

We then have

Corollary 4.5 *For any $\lambda \in \mathbb{R}^n$, $V(\lambda) \in \partial F(\lambda)$.*

Proof. It follows from Lemma 4.4 that $0 \in \partial F^-(\bar\lambda)$ and from (35) that $V(\bar\lambda) = \nabla F^+(\bar\lambda)$. The relation (33) then implies $V(\bar\lambda) \in \partial F(\bar\lambda)$. Since $\bar\lambda$ is arbitrary we are done. □

We need another assumption for our regularity result.

Assumption 4.6 $b_\ell \neq 0$ *for all* $\ell = 1, \ldots, n$.

Lemma 4.7 *Suppose Assumptions (4.1), (4.3) and (4.6) hold and let λ^* be the solution of (27). then every element of $\partial F(\lambda^*)$ is positive definite.*

Proof. We have proved that

$$\partial F(\lambda^*) = \partial F^-(\lambda^*) + \nabla F^+(\lambda^*) = \partial F^-(\lambda^*) + V(\lambda^*)$$

and every element in $\partial F^-(\lambda^*)$ is positive semidefinite. It is enough to prove $\nabla F^+(\lambda^*)$ is positive definite. We recall that at the solution

$$b_i = F_i(\lambda^*) = \int_a^b \left(x^0 + \sum_{\ell=1}^n \lambda_\ell^* a_\ell \right)_+ a_i dt, \qquad \forall i = 1, \ldots, n.$$

The assumption (4.6) implies that $(x^0 + \sum_{\ell=1}^n \lambda_\ell^* a_\ell)_+$ does not vanish identically at the support of each a_i. Then Lemma 4.2 implies that $\nabla F^+(\lambda^*) = V(\lambda^*)$ is positive definite. □

Illustration to problem (2). An essential assumption for problem (2) is that the second divided difference is positive, i.e., $d_i > 0$ for all $i = 1, \ldots, n$. Hence, Assumption (4.6) is automatically valid. It is easy to see that Assumptions (4.1) and (4.3) are also satisfied for B-splines. It follows from the above argument that the Newton method (26) is well defined near the solution. However, to prove its convergence we need the semismoothness property of F, which is addressed below.

4.3 Semismoothness

As we see from Theorem 3.4 that the property of semismoothness plays an important role in convergence analysis of nonsmooth Newton's method (21). In our application it involves functions of following type:

$$\Phi(\lambda) := \int_0^1 \phi(\lambda, t) dt \tag{36}$$

where $\phi : \mathbb{R}^n \times [a, b] \mapsto \mathbb{R}$ is a locally Lipschitz mapping. The following development is due to D. Ralph [41] and relies on a characterization of semismoothness using the Clarke generalized directional derivative.

Definition 4.8 *[6] Suppose $\psi : \mathbb{R}^n \mapsto \mathbb{R}$ is locally Lipschitz. The generalized directional derivative of ψ which, when evaluated at λ in the direction h, is given by*

$$\psi^\circ(\lambda; h) := \limsup_{\substack{\beta \to \lambda \\ \delta \downarrow 0}} \frac{\psi(\beta + \delta h) - \psi(\beta)}{\delta}.$$

The different quotient when upper limit is being taken is bounded above in light of Lipschitz condition. So $\psi^\circ(\lambda; h)$ is well defined finite quantity. An important property of ψ° is that for any h,

$$\psi^\circ(\lambda; h) = \max\{\langle \xi, h \rangle \mid \xi \in \partial\psi(\lambda)\}. \tag{37}$$

We now have the following characterization of semismoothness.

Lemma 4.9 *[41] A locally Lipschitz function $\psi : \mathbb{R}^n \mapsto \mathbb{R}$ is semismooth at $\bar{\lambda}$ if and only if ψ is directionally differentiable and*

$$\begin{aligned} &\psi(\lambda) + \psi^\circ(\lambda; \bar{\lambda} - \lambda) - \psi(\bar{\lambda}) \le o(\|\lambda - \bar{\lambda}\|), \ \textit{and} \\ &\psi(\lambda) - \psi^\circ(\lambda; -\bar{\lambda} + \lambda) - \psi(\bar{\lambda}) \ge o(\|\lambda - \bar{\lambda}\|). \end{aligned} \tag{38}$$

The equivalence remains valid if the inequalities are replaced by equalities.

Proof. Noticing that (37) implies $-\psi^\circ(\lambda, -h) = \min_{\xi \in \partial\psi(\lambda)} h^T \xi$, the conditions in (38) are equivalent to

$$\psi(\lambda) + [-\psi^\circ(\lambda; -\bar{\lambda} + \lambda), \psi^\circ(\lambda; \bar{\lambda} - \lambda)] - \psi(\bar{\lambda}) = o(\|\lambda - \bar{\lambda}\|).$$

Combining with the directional differentiability of ψ, this set-valued equation clearly implies the semismoothness of ψ at $\bar{\lambda}$ because for any $\xi \in \partial\psi(\lambda)$, we have

$$\xi^T(\bar{\lambda} - \lambda) \in [-\psi^\circ(\lambda; -\bar{\lambda} + \lambda), \psi^\circ(\lambda; \bar{\lambda} - \lambda)].$$

Conversely, if ψ is semismooth at $\bar{\lambda}$ then for any λ we take an element $\xi \in \partial\psi(\lambda)$ (respectively) to obtain

$$\psi^\circ(\lambda, \bar{\lambda} - \lambda) = \xi^T(\bar{\lambda} - \lambda) \qquad \text{(respectively } -\psi^\circ(\lambda; -\bar{\lambda} + \lambda) = \xi^T(-\bar{\lambda} + \lambda)).$$

The existence of such ξ follows from compactness of $\partial\psi(\lambda)$. Then the required inequalities follows from the semismoothness of ψ at $\bar{\lambda}$. □

Now we have our major result concerning the function in (36).

Proposition 4.10 *[41] Let $\phi : \mathbb{R}^n \times [0,1] \mapsto \mathbb{R}$. Suppose for every $t \in [0,1]$ $\phi(\cdot, t)$ is semismooth at $\lambda \in \mathbb{R}^n$. Then Φ defined in (36) is also semismooth at λ.*

Proof. The directional differentiability of Φ follows from the first part of [18, Proposition 3.1]. Now we use Lemma 4.9 to prove the semismoothness of Φ. To this purpose it is enough to establish the following relation:

$$\int_0^1 \Big(\phi(\lambda, t) + \phi^\circ((\lambda, t); (\bar{\lambda} - \lambda, 0)) - \phi(\bar{\lambda}, t)\Big)\, dt = o(\|\lambda - \bar{\lambda}\|). \tag{39}$$

This implies

$$\Phi(\lambda) - \Phi^\circ(\lambda; \bar{\lambda} - \lambda) - \Phi(\lambda) \le o(\|\lambda - \bar{\lambda}\|)$$

because the first principles give

$$\Phi^\circ(\lambda; \bar{\lambda} - \lambda) \le \int_0^1 \phi^\circ((\lambda, t); (\bar{\lambda} - \lambda, 0)) dt.$$

If in (39) we replace $\phi^\circ((\lambda, t); (\bar{\lambda} - \lambda, 0))$ by $-\phi^\circ((\lambda, t); (-\bar{\lambda} + \lambda, 0))$ and follow an argument that is almost identical to the subsequent development, we obtain the counter condition

$$\Phi(\lambda) - \Phi^\circ(\lambda; -\bar{\lambda} + \lambda) - \Phi(\lambda) \ge o(\|\lambda - \bar{\lambda}\|)$$

and the proof is sealed in Lemma 4.9.

Now let U be the closed unit ball in $\mathbb{R}^n$ and

$$e(\cdot, y) = \phi(y) + \phi^\circ(y; \cdot - y) - \phi(\cdot), \qquad y \in \mathbb{R}^n \times [0,1].$$

Let $\epsilon > 0$ we will find $\delta > 0$ such that if $\lambda \in \bar{\lambda} + \delta U$ then

$$\int_0^1 e((\bar{\lambda}, t), (\lambda, t))dt \leq \epsilon \|\lambda - \bar{\lambda}\|.$$

Since ϵ can be made arbitrarily small, verifying existence of δ is equivalent to verifying (39).

For any $\delta > 0$ let

$$\Delta(\delta) := \left\{ t \in [0,1] |\ e((\bar{\lambda}, t), (\lambda, t)) \leq \frac{\epsilon}{2} \|\lambda - \bar{\lambda}\|,\ \forall\, \lambda \in \bar{\lambda} + \delta U \right\}.$$

For each $\lambda \in \mathbb{R}^n$ the mapping $t \mapsto e((\bar{\lambda}, t), (\lambda, t))$ is measurable, hence the set

$$\left\{ t |\ e((\bar{\lambda}, t), (\lambda, t)) \leq \frac{\epsilon}{2} \|\lambda - \bar{\lambda}\| \right\}$$

is also measurable. Thus, $\Delta(\delta)$, the interior of measurable sets, is itself measurable. Obviously, $\Delta(\delta) \subseteq \Delta(\delta')$ if $\delta \geq \delta'$. And for fixed $t \in [0,1]$, semismoothness gives, via Lemma 4.9, that

$$\frac{e((\bar{\lambda}, t), (\lambda, t))}{\|\lambda - \bar{\lambda}\|} \to 0 \text{ as } 0 \neq \lambda - \bar{\lambda} \to 0,$$

i.e., for all small enough $\delta > 0$, $t \in \Delta(\delta)$.

Let $\Omega(\delta) := [0,1] \setminus \Delta(\delta)$. The properties of $\Delta(\delta)$ yields (a) measurability of $\Omega(\delta)$, (b) $\Omega(\delta) \supseteq \Omega(\delta')$ if $\delta \geq \delta'$, and (c) for each t and all small enough $\delta > 0$, $t \notin \Omega(\delta)$. In particular, $\cap_{\delta > 0} \Omega(\delta) = \emptyset$ and it follows that the measure of $\Omega(\delta)$, $\text{meas}(\Omega(\delta))$, converges to 0 as $\delta \to 0_+$.

Let L be the Lipschitz constant of ϕ in a neighborhood of $(\bar{\lambda}, 0)$, so that for each λ near $\bar{\lambda}$,

$$\begin{aligned} e((\bar{\lambda}, t), (\lambda, t)) &\leq |\phi(\lambda, t) - \phi(\bar{\lambda}, t)| + |\phi^\circ(\lambda, t); (\bar{\lambda} - \lambda, 0))| \\ &\leq 2L\|(\lambda - \bar{\lambda}, 0)\| \;=\; 2L\|\lambda - \bar{\lambda}\| \end{aligned}$$

using the 2-norm. To sum up,

$$\begin{aligned} \int_0^1 e((\bar{\lambda}, t), (\lambda, t))dt &= \left(\int_{\Omega(\delta)} + \int_{\Delta(\delta)} \right) e((\bar{\lambda}, t), (\lambda, t))dt \\ &\leq (2L\|\lambda - \bar{\lambda}\|)\text{meas}(\Omega(\delta)) + (\|\lambda - \bar{\lambda}\|\epsilon/2)\text{meas}(\Delta(\delta)) \\ &\leq \|\lambda - \bar{\lambda}\|(2L\text{meas}(\Omega(\delta)) + \epsilon/2). \end{aligned}$$

Choose $\delta > 0$ small enough such that $\text{meas}(\Omega(\delta)) < \epsilon/(4L)$, and we are done. $\square$

Corollary 4.11 *Under Assumption 4.1, the functions F_j defined in (28) are each semismooth.*

Proof. For each $t \in [a, b]$, the mapping $\phi_j : \mathbb{R}^n \mapsto \mathbb{R}$ by

$$\phi_j(\lambda, t) = a_j(t)(x^0 + \sum_{\ell=1}^{n} \lambda_\ell a_\ell)_+$$

is piecewise linear with respect to λ, and hence is semismooth. Then Proposition 4.10 implies that each F_j defined in (28) is semismooth since $F_j(\lambda) = \int_a^b \phi_j(\lambda, t)dt$. $\square$

Now we are ready to use Theorem 3.4 of Qi and Sun [40] to establish the superlinear convergence of the Newton method (29) for the equation (27).

Theorem 4.12 *Suppose that Assumptions (4.1), (4.3) and (4.6) hold. Then Newton's method (29) for (27) is superlinearly convergent provided that the initial point λ^0 is close enough to the unique solution λ^*.*

Proof. Three major elements for the use of Theorem 3.4 have been established: (i) $V(\lambda) \in \partial F(\lambda)$ for any $\lambda \in \mathbb{R}^n$ (see, corollary 4.5), (ii) F is regular at λ^* (see, Lemma 4.7), and (iii) F is semismooth since each F_j is semismooth (see, Corollary 4.11). The result follows the direct application of Theorem 3.4 to the equation (27). $\square$

Illustration to (26). The superlinear convergence of the method (26) is a direct consequence of Theorem 4.12 because all the assumptions for Theorem 4.12 are satisfied for the convex best interpolation problem (1). This recovers the main result in [18]. Refinement of some results in [18] by taking into account of special structures of the B-splines leads to the quadratic convergence analysis conducted in [19].

4.4 Application to Inequality Constraints

Now we consider the approximation problem given by inequality constraints:

$$K = C \cap \{x | Ax \leq b\}.$$

Under the strong CHIP assumption, we have solution characterization (16) and (17), which we restate below for easy reference.

$$\lambda \geq 0, \ w := AP_C(x^0 + A^*\lambda) - b \geq 0, \ \lambda^T w = 0. \tag{40}$$

Again for computational consideration we assume that C is the cone of positive functions so that $P_C(x) = x_+$. Below we design Newton's method for (40) and study when it is superlinearly convergent. To do this, we use the well-known Fischer-Burmeister NCP function, widely studied in nonlinear complementarity problems [23, 42], to reformulate (40) as a system (semismooth) equations.

Recall the Fischer-Burmeister function is given by

$$\phi_{FB}(a,b) := a + b - \sqrt{a^2+b^2}.$$

Two important properties of ϕ_{FB} are

$$\phi_{FB}(a,b) = 0 \iff a \geq 0,\ b \geq 0,\ ab = 0$$

and the square ϕ_{FB}^2 is continuously differentiable, though ϕ_{FB} is not differentiable. Define

$$\Phi_{FB}(\lambda, w) := \begin{pmatrix} \phi_{FB}(\lambda_1, w_1) \\ \vdots \\ \phi_{FB}(\lambda_n, w_n) \end{pmatrix}$$

and

$$W(\lambda, w) := \begin{pmatrix} AP_C(x^0 + A^*\lambda) - w - b \\ \Phi_{FB}(\lambda, w) \end{pmatrix}.$$

Then it is easy to see that (40) is equivalent to the nonsmooth equation

$$W(\lambda, w) = 0.$$

Since W is locally Lipschitz, direct calculation gives

$$\partial W(\lambda, w) \subseteq \left\{ \begin{pmatrix} V(\lambda) & -I \\ D(\lambda, w) & E(\lambda, w) \end{pmatrix} \;\middle|\; \begin{array}{l} V(\lambda) \in \partial F(\lambda) \\ D(\lambda, w), E(\lambda, w) \text{ satisfy (42) and (43)} \end{array} \right\}. \tag{41}$$

$D(\lambda, w)$ and $E(\lambda, w)$ are diagonal matrices whose ℓth diagonal element is given by

$$D_\ell(\lambda, w) := 1 - \frac{\lambda_\ell}{\|(\lambda_\ell, w_\ell)\|}, \quad E_\ell(\lambda, w) := 1 - \frac{w_\ell}{\|(\lambda_\ell, w_\ell)\|} \tag{42}$$

if $(\lambda_\ell, w_\ell) \neq 0$ and by

$$D_\ell(\lambda, w) = 1 - \xi_\ell,\ E_\ell(\lambda, w) = 1 - \rho_\ell,\ \forall (\xi_\ell, \rho_\ell) \in \mathbb{R}^2 \text{ such that } \|(\xi_\ell, \rho_\ell)\| \leq 1 \tag{43}$$

if $(\lambda_\ell, w_\ell) = 0$.

Lemma 4.13 *Suppose every element $V(\lambda)$ in $\partial F(\lambda)$ is positive definite. Then every element of $\partial W(\lambda, w)$ is nonsingular.*

Proof. Let $M(\lambda, w)$ be an element of the right side set in (41) and let $(y, z) \in \mathbb{R}^{2n}$ be such that $M(y, z) = 0$. Then there exist $V(\lambda) \in \partial F(\lambda)$ and $D(\lambda, w)$ and $E(\lambda, w)$ satisfying (42) and (43) such that

$$V(\lambda)y - z = 0 \quad \text{and} \quad D(\lambda, w)y + E(\lambda, w)z = 0.$$

Since $V(\lambda)$ is nonsingular, it yields that

$$(DV^{-1} + E)z = 0.$$

It is well known from the NCP theory [10, Theorem 21] that the matrix $(DV^{-1}+E)$ is nonsingular because V^{-1} is positive definite according to the assumption. Hence, $z=0$, implying $y=0$. This establishes the nonsingularity of all elements in $\partial W(\lambda, w)$. $\square$

Newton's method for (40) can be developed as follows

$$(\lambda^+, w^+) - (\lambda, w) = -M^{-1}W(\lambda, w), \quad M \in \partial W(\lambda, w). \tag{44}$$

We have proved that each F_j is semismooth (Corollary 4.11). Using the fact that composite of semismooth functions is semismooth and the Fischer-Burmeister function is strongly semismooth, we know that W is semismooth function. Suppose (λ^*, w^*) is a solution of (40).

Assumption 4.14 *Each $b_\ell > 0$ for $\ell = 1, \ldots, n$.*

Lemma 4.15 *Suppose Assumption (4.1), (4.3) and (4.14) hold. Then every element in $\partial W(\lambda^*, w^*)$ is nonsingular.*

Proof. We note that at the solution it holds

$$AP_C(x^0 + A^*\lambda^*) = b + w^*.$$

Since $w^*_\ell \geq 0$, we see that $b_\ell + w^*_\ell > 0$. Following the proof of Lemma 4.7 we can prove that each element V in $\partial F(\lambda^*)$ is positive definite, and hence each element of $\partial W(\lambda^*, w^*)$ is nonsingular by Lemma 4.15. $\square$

All preparation is ready for the use of Theorem 3.4 to state the superlinear convergence of the method (44). The proof is similar to Theorem 4.12.

Theorem 4.16 *Suppose Assumptions (4.1), (4.3) and (4.14) hold. Then the Newton method (44) is superlinearly convergent provided that the initial point (λ^0, w^0) is sufficiently close to (λ^*, w^*).*

We remark that the quadratic convergence is also possible if we could establish the strong semismoothness of W at (λ^*, w^*). A sufficient condition for this property is that each F_j is strongly semismooth since the Fischer-Burmeister function is automatically strongly semismooth.

4.5 Globalization

In the previous subsections, Newton's method is developed for nonsmooth equations arising from constrained interpolation and approximation problems. It is locally superlinearly convergent under reasonable conditions. It is also worth of mentioning it globalization scheme that makes the Newton method globally convergent.

The first issue to be resolved is that we need an objective function for the respective problems. Natural choices for objective functions are briefly described below with outline of an algorithm scheme, but without global convergence analysis. It is easy to see (following discussion in [38, 18]) that the function f given by

$$f(\lambda) := \int_a^b \left(x^0 + \sum_{\ell=1}^n \lambda_\ell a_\ell \right)_+^2 dt - \sum_{\ell=1}^n \lambda_\ell b_\ell$$

severs this purpose because

$$\nabla f(\lambda) = F(\lambda) - b.$$

Since f is convex, $\|\nabla f(\lambda)\| = \|F(\lambda) - b\|$ can be used to monitor the convergence of global methods. We present below a global method, which globalizes the method (29) and has been shown extremely efficient for the convex best interpolation problem (1).

Algorithm 4.17 *(Damped Newton method)*

(S.0) (Initialization) Choose $\lambda^0 \in \mathbb{R}^n$, $\rho \in (0,1), \sigma \in (0, 1/2)$, *and tolerance* $\mathrm{tol} > 0$. $k := 0$.

(S.1) (Termination criterion) If $\epsilon_k = \|F(\lambda^k) - d\| \le \mathrm{tol}$ *then stop. Otherwise, go to (S.2).*

(S.2) (Direction generation) Let s^k *be a solution of the following linear system*

$$(V(\lambda^k) + \epsilon_k I)s = -\nabla f(\lambda^k). \tag{45}$$

(S.3) (Line search) Choose m_k *as the smallest nonnegative integer* m *satisfying*

$$f(\lambda^k + \rho^m s^k) - f(\lambda^k) \le \sigma \rho^m \nabla f(\lambda^k)^T s^k. \tag{46}$$

(S.4) (Update) Set $\lambda^{k+1} = \lambda^k + \rho^{m_k} s^k$, $k := k + 1$, *return to step (S.1).*

Since $V(\lambda)$ is positive semidefinite, the matrix $(V(\lambda) + \epsilon I)$ is positive definite for $\epsilon > 0$. Hence the linear equation (45) is well defined and the direction s^k is a descent direction for the objective function f. The global convergence analysis for Algorithm 4.17 is standard and can be found in [19].

Globalized version for the method (44) can be developed as well, but with some notable differences. To this case, the objective function $f(\lambda, w)$ is given by

$$f(\lambda, w) := \int_a^b \left(x^0 + \sum_{\ell=1}^n \lambda_\ell a_\ell \right)_+^2 dt - \sum_{\ell=1}^n \lambda_\ell (b + w) + \|\Phi_{FB}(\lambda, w)\|^2.$$

This function is also continuously differentiable, but not convex because $\|\Phi_{FB}(\lambda, w)\|^2$ is not convex although continuously differentiable. We also note that the gradient of $f(\lambda, w)$ is not $W(\lambda, w)$ any more. A global method based on f can be developed by following the scheme in [10].

5 Open Problems

It is obvious from Section 2 and Section 4 that there is a big gap between theoretical results and Newton-type algorithms for constrained interpolation problems. For example, the solution characterizations appeared in Theorems 2.2, 2.6, and 2.7 are for general convex sets (i.e., C is a closed convex set), however, the Newton method well-developed so far is only on the particular case yet the most important case that C is the cone of positive functions. This is due to the fact that the projection is an essential ingredient when solving the interpolation problem, and that the projection on the cone of positive functions is easy to calculate.

There are many problems that are associated to the projections onto other convex sets including cones. We only discuss two of them which we think are most interesting and likely to be (at least partly) solved by the techniques developed in this paper. The first one is the case that C is a closed polyhedral set in X, i.e.,

$$C := \{x \in X |\ \langle c_i, x\rangle \le r_i, \quad i = 1, \ldots, m\}$$

where $c_i \in X$ and $r_i \in \mathbb{R}$. We note that cones are not necessarily polyhedral. It follows from [11, Examples 10.7 and 10.9] that the sets $\{C, \cap H_j\}$ and $\{C, \cap \mathcal{H}_j\}$ both have strong CHIP. Hence the solution characterization theorems are applicable to the polyhedral case. Questions related to P_C include differentiability, directional differentiability, generalized Jacobian and semismoothness of the mapping AP_C, and most importantly how to design Newton's method for this case.

The second is the problem of interpolating a finite set of points with a curve constrained to lie between two piecewise linear splines (with knots at the abscissae of the given points). The objective is to minimize the 2-norm of the second derivative of the interpolant. Let (t_i, y_i) be given data points in $\mathbb{R}^2$ with

$$t_0 < t_1 < \ldots < t_n, \ \ \phi(t_i) < y_i < \psi(t_i) \text{ for } i = 1, \ldots, n.$$

Hence ϕ and ψ are given piecewise linear functions (or more generally lower and upper semicontinuous functions, respectively) such that

$$\inf_{t \in [t_0, t_n]} (\psi(t) - \phi(t)) > 0.$$

The constraint is

$$\hat{C} := \{x \in W^{2,2}[t_0, t_n] |\ \phi(t) \le x(t) \le \psi(t)\}$$

and

$$H := \{x \in W^{2,2}[t_0, t_n] |\ x(t_i) = y_i, \ \imath = 1, \ldots, n\}.$$

This problem can be reformulated as a constrained interpolation problem from a convex set in certain Hilbert space [15, 2]. Questions similar to that for the first problem remain unsolved for this interpolation problem from a strip.

Acknowledgment. The author would like to thank Danniel Ralph for his constructive comments on the topic and especially for his kind offer of his material [41] on semismoothness of integral functions being included in this survey (i.e., Sec. 4.3). It is also

interesting to see how his approach can be extended to cover the strongly semismooth case.

References

[1] L.-E. Andersson and T. Elfving, *An algorithm for constrained interpolation*, SIAM J. Sci. Statist. Comput. 8 (1987), pp. 1012–1025.

[2] L.-E. Andersson and T. Elfving, *Best constrained approximation in Hilbert space and interpolation by cubic splines subject to obstacles*, SIAM J. Sci. Comput. 16 (1995), pp. 1209–1232.

[3] H.H. Bauschke, J.M. Borwein, and P. Tseng, *Bounded linear regularity, strong CHIP, and CHIP are distinct properties*, J. Convex Anal. 7 (2000), pp. 395–412.

[4] H.H. Bauschke, J.M. Borwein, and W. Li, *Strong conical hull intersection property, bounded linear regularity, Jameson's property* (G)*, and error bounds in convex optimization*, Math. Program. 86 (1999), pp. 135–160.

[5] J. Borwein and A. S. Lewis, *Partially finite convex programming I: Quasi relative interiors and duality theory,* Math. Programming 57 (1992), pp. 15–48.

[6] F. H. Clarke, *Optimization and Nonsmooth Analysis,* John Wiley & Sons, New

[7] C.K. Chui, F. Deutsch, and J.D. Ward, *Constrained best approximation in Hilbert space*, Constr. Approx., 6 (1990), pp. 35–64.

[8] C.K. Chui, F. Deutsch, and J.D. Ward, *Constrained best approximation in Hilbert space II*, J. Approx. Theory, 71 (1992), pp. 213–238.

[9] C. de Boor, *A Practical Guide to Splines*, Springer-Verlag, New York, 1978.

[10] T. De Luca, F. Facchinei, and C. Kanzow, *A semismooth equation approach to the solution of nonlinear complementarity problems*, Math. Programming 75 (1996), pp. 407–439.

[11] F. Deutsch, *Best approximation in inner product spaces.* CMS Books in Mathematics 7. Springer-Verlag, New York, 2001.

[12] F. Deutsch, W. Li, and J.D. Ward, *A dual approach to constrained interpolation from a convex subset of Hilbert space*, J. Approx. Theory 90 (1997), pp. 385–414.

[13] F. Deutsch, W. Li, and J.D. Ward, *Best approximation from the intersection of a closed convex set and a polyhedron in Hilbert space, weak Slater conditions, and the strong conical hull intersection property*, SIAM J. Optim. 10 (1999), pp. 252–268

[14] F. Deutsch, V.A. Ubhaya, J.D. Ward, and Y. Xu, *Constrained best approximation in Hilbert space. III. Applications to n-convex functions*, Constr. Approx. 12 (1996), pp. 361–384.

[15] A.L. Dontchev, *Best interpolation in a strip*, J. Approx. Theory 73 (1993), pp. 334–342.

[16] A.L. Dontchev and B. D. Kalchev, *Duality and well-posedness in convex interpolation*, Numer. Funct. Anal. and Optim. 10 (1989), pp. 673-689.

[17] A.L. Dontchev and I. Kolmanovsky, *Best interpolation in a strip. II. Reduction to unconstrained convex optimization,* Comput. Optim. Appl. 5 (1996), pp. 233–251.

[18] A.L. Dontchev, H.-D. Qi, and L. Qi, *Convergence of Newton's method for convex best interpolation*, Numer. Math. 87 (2001), pp. 435–456.

[19] A.L. Dontchev, H.-D. Qi, and L. Qi, *Quadratic convergence of Newton's method for convex interpolation and smoothing*, Constr. Approx. 19 (2003), pp. 123–143.

[20] A.L. Dontchev, H.-D. Qi, L. Qi, and H. Yin, *A Newton method for shape-preserving spline interpolation*, SIAM J. Optim. 13 (2002), pp. 588–602.

[21] F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems*, Vol. I & II. Springer-Verlag, New York, 2003.

[22] J. Favard, *Sur l'interpolation*, J. Math. Pures Appl., 19 (1940), pp. 281–306.

[23] A. Fischer, *A special Newton-type optimization method*, Optimization, 24 (1992), pp. 269-284.

[24] M.S. Gowda and M. Teboulle, *A comparison of constraint qualifications in infinite-dimensional convex programming*, SIAM J. Control Optim. 28 (1990), pp. 925–935.

[25] U. Hornung, *Interpolation by smooth functions under restriction on the derivatives,* J. Approx. Theory 28 (1980), pp. 227–237.

[26] G. Iliev and W. Pollul, *Convex interpolation by functions with minimal L_p norm $(1 < p < \infty)$ of the kth derivative,* Mathematics and mathematical education (Sunny Beach, 1984), pp. 31–42, Bulg. Akad. Nauk, Sofia, 1984.

[27] L. D. Irvine, S. P. Marin and P. W. Smith, *Constrained interpolation and smoothing*, Constr. Approx. 2 (1986), pp. 129–151.

[28] V. Jeyakumar, *Infinite-dimensional convex programming with applications to constrained approximation*, J. Optim. Theory Appl. 75 (1992), pp. 569–586.

[29] V. Jeyakumar and D.T. Luc, *Approximate Jacobian matrices for nonsmooth continuous maps and C^1-optimization*, SIAM J. Control Optim. 36 (1998), pp. 1815–1832.

[30] V. Jeyakumar and H. Wolkowicz, *Generalizations of Slater's constraint qualification for infinite convex programs*, Math. Programming 57 (1992), pp. 85–101.

[31] H. Jiang and L. Qi, *Local uniqueness and Newton-type methods for nonsmooth variational inequalities*, J. Math. Analysis and Appl. 196 (1995), pp. 314–331.

[32] D. Klatte and B. Kummer, *Nonsmooth equations in optimization. Regularity, calculus, methods and applications.* Nonconvex Optimization and its Applications, 60. Kluwer Academic Publishers, Dordrecht, 2002.

[33] B. Kummer, *Newton's method for nondifferentiable functions.* Advances in mathematical optimization, 114–125, Math. Res., 45, Akademie-Verlag, Berlin, 1988.

[34] C. Li and X.Q. Jin, *Nonlinearly constrained best approximation in Hilbert spaces: the strong chip and the basic constraint qualification*, SIAM J. Optim. 13 (2002), pp. 228–239.

[35] C. Li and K.F. Ng, *On best approximation by nonconvex sets and perturbation of nonconvex inequality systems in Hilbert spaces*, SIAM J. Optim. 13 (2002), pp. 726–744.

[36] C. Li and K.F. Ng, *Constraint qualification, the strong chip, and best approximation with convex constraints in Banach spaces*, SIAM J. Optim. 14 (2003), pp. 584–607.

[37] C. A. Micchelli, P. W. Smith, J. Swetits and J. D. Ward, *Constrained L_p approximation*, Constr. Approx. 1 (1985), pp. 93–102.

[38] C. A. Micchelli and F. I. Utreras, *Smoothing and interpolation in a convex subset of a Hilbert space,* SIAM J. Sci. Statist. Comput. 9 (1988), pp. 728–747.

[39] R. Mifflin, *Semismoothness and semiconvex functions in constrained optimization*, SIAM J. Control Optim. 15 (1977), pp. 959–972.

[40] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Programming 58 (1993), pp. 353–367.

[41] D. Ralph, *Personal communication*, May, 2002.

[42] D. Sun and L. Qi, *On NCP-functions*, Comput. Optim. Appl. 13 (1999), pp. 201–220.

[43] H. Xu, *Set-valued approximations and Newton's methods*, Math. Program. 84 (1999), pp. 401–420.