ITERATIVE METHODS FOR k-HESSIAN EQUATIONS

GERARD AWANOU

ABSTRACT. On a domain of the n-dimensional Euclidean space, and for an integer $k = 1, \ldots, n$, the k-Hessian equations are fully nonlinear elliptic equations for $k > 1$ and consist of the Poisson equation for $k = 1$ and the Monge-Ampère equation for $k = n$. We analyze for smooth non degenerate solutions a 9-point finite difference scheme. We prove that the discrete scheme has a locally unique solution with a quadratic convergence rate. In addition we propose new iterative methods which are numerically shown to work for non smooth solutions. A connection of the latter with a popular Gauss-Seidel method for the Monge-Ampère equation is established and new Gauss-Seidel type iterative methods for 2-Hessian equations are introduced.

1. INTRODUCTION

Let Ω be a bounded, connected open subset of $\mathbb{R}^n, n \geq 2$ with boundary denoted $\partial\Omega$. Let $u \in C^2(\Omega)$ and for $x \in \Omega$, let $D^2u(x) = \left((\partial^2 u(x))(\partial x_i \partial x_j) \right)_{i,j=1,\ldots,n}$ denote its Hessian. We denote the eigenvalues of $D^2u(x)$ by $\lambda_i(x), i = 1, \ldots, n$. For $1 \leq k \leq n$, the k-Hessian operator is defined as

$$S_k(D^2u) = \sum_{i_1 < \cdots < i_k} \lambda_{i_1} \cdots \lambda_{i_k}.$$

We note that $S_1(D^2u) = \Delta u$ is the Laplacian operator and $S_n(D^2u) = \det D^2u$ is the Monge-Ampère operator. For $k \geq 2$, we are interested in the numerical approximation of solutions of the Dirichlet problem for the k-Hessian equation

$$(1.1) \qquad S_k(D^2u) = f \text{ in } \Omega, u = g \text{ on } \partial\Omega,$$

with f and g given and $f \geq 0$.

1.1. Local existence, uniqueness and quadratic convergence rate for a finite difference discretization.
Let u^0 be a sufficiently close initial guess to the smooth solution u of (1.1). Consider the iterative method

$$(1.2) \qquad \operatorname{div}\left(\{S_k^{ij}(D^2u^0)\} Du^{m+1} \right) = \operatorname{div}\left(\{S_k^{ij}(D^2u^0)\} Du^m \right) + f - S_k(D^2u^m) \text{ in } \Omega$$

$$u^{m+1} = g \text{ on } \partial\Omega,$$

where $\{S_k^{ij}(D^2u^0)\}$ is a matrix which generalizes the cofactor matrix of D^2u^0.

We prove the convergence of (1.2) at the continuous level in Hölder spaces. A discrete version of (1.2) is also shown to converge to a solution of a 9-point stencil discretization of (1.1). This establishes the local existence and uniqueness of a discrete solution. In addition the convergence rate of the discretization is shown to be quadratic.

It is reasonable to expect that the discrete version of the iterative method (1.2) will retrieve the correct solution when it is smooth and non degenerate. As with Newton's method it is not effective for non smooth and degenerate solutions. For these, we advocate iterative methods like the subharmonicity preserving iterations described below. The discrete version of (1.2) is used in this paper to prove the local solvability of the 9-point scheme when u is smooth and non degenerate. These results form a building block of a theory which explains why standard discretizations work for non smooth solutions [4]. In addition results for smooth solutions are also needed for the analysis of hybrid schemes where the 9 point scheme is used in part of the region where the solution is smooth and a monotone scheme elsewhere [2].

1.2. **Newton's method.** If one is only interested in smooth solutions, Newton's method is the most appropriate method. We analyze the convergence of Newton's method for solving (1.1) when it has a smooth solution.

1.3. **Numerical work for subharmonicity preserving iterations.** A smooth function u is said to be k-convex if $S_l(D^2u) \geq 0, 1 \leq l \leq k$. Convexity of a function can be shown to be equivalent to n-convexity, Lemma 2.5. It is of interest in some applications to be able to handle (1.1) when it has a non smooth k-convex solution. It has only been recently understood, c.f. [4] for the Monge-Ampère equation, that what is needed is a numerical method provably convergent for smooth solutions and numerically robust to handle non smooth solutions. The approach in [4] is to regularize the data and use approximation by smooth functions. The key to numerically handle non smooth solutions of (1.1) is to preserve k-convexity in the iterations. For discrete k-convexity we simply require discrete analogues of the condition $S_l(D^2u) \geq 0$ with a natural discretization of D^2u. We refer to [1] where this approach was first used for the discretization of n-convexity.

Consider the iterative method

$$(1.3) \qquad \Delta u^{m+1} = \left((\Delta u^m)^k + \frac{1}{c(k,n)}(f - S_k(D^2u^m)) \right)^{\frac{1}{k}} \text{ in } \Omega, u^{m+1} = g \text{ on } \partial\Omega,$$

with $c(k,n) = \binom{n}{k}/n^k$.

If D^2u has positive eigenvalues, we have the inequality

$$(1.4) \qquad\qquad\qquad S_k(D^2u) \leq c(k,n)(\Delta u)^k,$$

which follows from the Maclaurin inequalities, [13, Proposition 1.1 (v i)].

For $k = 2$, (1.4) also holds with *no convexity assumption* on u, [20, Lemma 15.11]. Explicitly $c(2,3) = 1/3$. Also, $c(n,n) = 1/n^n$ which gives

$$\det D^2u \leq \frac{1}{n^n}(\Delta u)^n,$$

a direct consequence of the arithmetic mean - geometric mean inequality.

If one starts with an initial guess u^0 such that $\Delta u^0 \geq 0$, (1.3) enforces $\Delta u^m \geq 0$ for all m. Indeed recall that $f \geq 0$ and assume that $\Delta u^m \geq 0$. Then by (1.4) $1/c(k,n)S_k(D^2u^m) \leq (\Delta u^m)^k$, and using (1.3) it follows that $(\Delta u^{m+1})^k \geq 0$. In other words, starting with an initial guess u^0 with $\Delta u^0 \geq 0$, (1.3) enforces subharmonicity

in arbitrary dimension for smooth convex solutions and subharmonicity for 2-Hessian equations with no convexity assumption on u. In addition for 2-Hessian equations, the limit solution solves $S_2(D^2u) = f \geq 0$. That is, the sequence u^{m+1} defined by (1.3) has a formal limit which solves $\Delta u \geq 0$ and $S_2(D^2u) \geq 0$. Thus (1.3) enforces 2-convexity in arbitrary dimension for 2-Hessian equations.

Another class of iterative methods we introduce in this paper are Gauss-Seidel type iterative methods. The Gauss-Seidel methods are more efficient than (1.3) for large scale problems.

The simplicity of the methods discussed in this paper and the facility with which they can be implemented, make them attractive to researchers interested in Monge-Ampère equations. The other major motivation to study the subharmonicity preserving iterations is that they can be adapted to the finite element context and have been numerically shown in that context to be robust for non smooth solutions.

In two dimension, (1.3) appears to perform well in the degenerate case $f \geq 0$ as discrete k-convexity is enforced in the iterations. The situation is different in three dimension with $k = 2$. We were not able to reproduce the solution $u(x, y, z) = |x-1/2|$ by solving (1.1) with $k = 2$ and using (1.3). Here, since u does not depend on z, we have $f(x, y, z) = 0$ as in the two dimensional case. However, for $n = 3$ and $k = 3$, we can preserve convexity in the degenerate case by using the sequence of nonlinear 2-Hessian equations

$$
(1.5) \qquad S_2(D^2u^{m+1}) = 3\left(\left(\frac{1}{3}S_2(D^2u^m) \right)^{\frac{3}{2}} + f - \det D^2u^m \right)^{\frac{2}{3}},
$$

with $u^{m+1} = g$ on $\partial\Omega$. Each of these equations is solved iteratively by (1.3) with $k = 2, n = 3$. We note that $\left(\frac{1}{3}S_2(D^2u^m) \right)^{\frac{3}{2}} - \det D^2u^m \geq 0$ when $S_2(D^2u^m) > 0$, [20, Lemma 15.12]. Starting with an initial guess which satisfies $S_2(D^2u^0) > 0$ and setting $\det D^2u^m = 0$ in (1.5) whenever $S_2(D^2u^m) = 0$, we obtain a double sequence iterative method which at the limit enforce $\Delta u \geq 0, S_2(D^2u) \geq 0$, and $\det D^2u = f \geq 0$.

The reason for setting $\det D^2u^m = 0$ in (1.5) whenever $S_2(D^2u^m) = 0$ is motivated by the observation that in the case $f = 0$, if $S_2(D^2u^m) = 0$, $S_2(D^2u^{m+1})$ is ill-defined or complex valued if $\det D^2u^m > 0$. While (1.3) may be inexact for degenerate 2-Hessian equations, its use inside a double iterative method appears effective. This is reminiscent of inexact Uzawa algorithms.

1.4. **Relation with other work.** The k-Hessian equations have mainly applications in conformal geometry and physics. The Monge-Ampère operator has received recently a lot of interest from numerical analysts. For $n = 3$ and $k = 2$, the numerical resolution of (1.1) has been considered in [22], where it was referred to as the σ_2 problem. The iterative method (1.3) generalizes an iterative method introduced in [5] for the two dimensional Monge-Ampère equation. The latter corresponds to the choice $k = n = 2$ and the constant $c(2, 2) = 1/4$ replaced by $1/2$. The 2-Hessian equation has also been considered recently in [12] from the point of view of monotone schemes.

We will see that if the central finite difference discretization of (1.3) is solved by a Gauss-Seidel iterative method, one recovers a Gauss-Seidel iterative method which has been used by many authors to solve the two dimensional Monge-Ampère equation. We will refer to the latter method as the 2D Gauss-Seidel method for Monge-Ampère equation. It has been used in the numerical simulation of Ricci flow [15], as a smoother in multigrid methods for the balance vortex model in meteorology, [8, 7] and has been recently shown numerically to capture the viscosity solution of the 2D Monge-Ampère equation [5]. The connection between (1.3) and the 2D Gauss-Seidel method for the Monge-Ampère equation is what enables us to introduce new Gauss-Seidel type iterative methods for k-Hessian equations.

The ingredients of our proof of the convergence rate for the finite difference discretization are discrete Schauder estimates and a suitable generalization of the combined fixed point iterative method used in [10]. Schauder estimates were also used in the proof of convergence of Newton's method at the continuous level [21].

1.5. **Organization of the paper.** The paper is organized as follows: In the next section, we give some notations, recall the Schauder estimates and their discrete analogues. In section 3 we prove our main results on the quadratic convergence rate of a finite difference discretization of (1.1) and in section 4 we prove the convergence of Newton's method. In section 5 we introduce new Gauss-Seidel type iterative methods and their connections with the subharmonicity preserving iterations (1.3). Section 6 is devoted to numerical results. We conclude with some remarks. The reader interested only in the Monge-Ampère equation, or for a first reading, may assume that $k = n$.

2. NOTATION AND PRELIMINARIES

2.1. **Hölder spaces and Schauder estimates.** For a nonnegative integer r or for $r = \infty$, we denote by $C^r(\Omega)$ the set of all functions having all derivatives of order $\leq r$ continuous on Ω and by $C^r(\overline{\Omega})$, the set of all functions in $C^r(\Omega)$ whose derivatives of order $\leq r$ have continuous extensions to $\overline{\Omega}$. For a multi-index $\beta = (\beta_1, \ldots, \beta_n) \in \mathbb{N}^n$, put $|\beta| = \beta_1 + \ldots + \beta_n$. We use the notation $D^\beta u(x)$ for the partial derivative $(\partial/\partial x_1)^{\beta_1} \ldots (\partial/\partial x_n)^{\beta_n} u(x)$.

The norm in $C^r(\Omega)$ is given by

$$||u||_{r;\Omega} = \sum_{j=0}^{r} |u|_{j;\Omega}, \quad |u|_{j;\Omega} = \sup_{|\beta|=j} \sup_{\Omega} |D^\beta u(x)|.$$

We denote by $|x|$ the Euclidean norm of $x \in \mathbb{R}^n$. A function u is said to be uniformly Hölder continuous with exponent $\alpha, 0 < \alpha \leq 1$ in Ω if the quantity

$$\sup_{x \neq y} \frac{|u(x) - u(y)|}{|x - y|^\alpha},$$

is finite. The space $C^{r,\alpha}(\overline{\Omega})$ consists of functions whose r-th order derivatives are uniformly Hölder continuous with exponent α in Ω. It is a Banach space with norm

$$||u||_{r,\alpha;\Omega} = ||u||_{r;\Omega} + [u]_{r,\alpha;\Omega},$$

where
$$[u]_{r,\alpha;\Omega} = \sup_{|\beta|=r}\sup_{x\neq y}\frac{|D^{\beta}u(x) - D^{\beta}u(y)|}{|x - y|^{\alpha}}.$$

The norms $||\,||_{r;\Omega}$ and $||\,||_{r,\alpha;\Omega}$ are naturally extended to vector fields and matrix fields by taking the supremum over all components. We make the standard convention of using C for a generic constant. For $A = (a_{ij})_{i,j=1,\dots,n}$ and $B = (b_{ij})_{i,j=1,\dots,n}$ we recall that $A : B = \sum_{i,j=1}^{n} a_{ij}b_{ij}$. We will often use the following property

(2.1) $$||fg||_{0,\alpha;\Omega} \leq C||f||_{0,\alpha;\Omega}||g||_{0,\alpha;\Omega}, \text{ for } f, g \in C^{0,\alpha}(\overline{\Omega}),$$

from which it follows that if A, B are matrix fields

(2.2) $$||A : B||_{0,\alpha;\Omega} \leq C\sum_{i,j=1}^{n} ||a_{ij}||_{0,\alpha;\Omega}||b_{ij}||_{0,\alpha;\Omega}.$$

We first state a global regularity result for the solution of strictly elliptic equations, which follows from [14, Theorems 6.14, 6.6 and Corollary 3.8].

Theorem 2.1. *Assume $0 < \alpha < 1$. Let Ω be a $C^{2,\alpha}$ domain in $\mathbb{R}^n$ and $f, a^{ij} \in C^{\alpha}(\overline{\Omega})$, $\phi \in C^{2,\alpha}(\overline{\Omega})$. We consider the strictly elliptic operator*

(2.3) $$Lu = \sum_{i,j=1}^{n} a^{ij}(x)\frac{\partial^2}{\partial x_i \partial x_j}u(x),$$

with coefficients satisfying for positive constants λ, Λ,

$$\sum_{i,j=1}^{n} a^{ij}(x)\zeta_i\zeta_j \geq \lambda \sum_{l=1}^{n} \zeta_l^2, \zeta_l \in \mathbb{R}, \text{ } and |a^{i,j}|_{0,\alpha;\Omega} \leq \Lambda.$$

Then the solution u of the equation

$$Lu = f \text{ } in \,\Omega, u = \phi \text{ } on \,\partial\Omega,$$

satisfies

$$||u||_{2,\alpha;\Omega} \leq C(||\phi||_{2,\alpha;\Omega} + ||f||_{0,\alpha;\Omega}),$$

where C depends on $n, \alpha, \lambda, \Lambda, \Omega, \sup_{\partial\Omega}|\phi|,$ and $\sup_{\Omega}|f|/\lambda$.

We will make the slight abuse of language of also denoting by $S_k(x), x = (x_1, \dots, x_n)$ the kth elementary symmetric polynomial of the variable x, i.e.

$$S_k(\lambda) = \sum_{i_1 < \cdots < i_k} \lambda_{i_1} \cdots \lambda_{i_k}.$$

A function $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$ with Hessian D^2u having eigenvalues $\lambda_i, i = 1, \dots, n$ is said to be k-admissible if $S_j(\lambda) > 0, j = 1, \dots, k$. Solutions of the k-Hessian equation will be required to be k-admissible, thus requiring $f > 0$.

Moreover, let $\kappa = (\kappa_1, \dots, \kappa_{n-1})$ denote the principal curvatures of $\partial\Omega$.

Definition 2.2. *The domain Ω is said to be $(k-1)$-convex if there exists $c_0 > 0$ such that*

$$S_{k-1}(\kappa) \geq c_0 > 0 \text{ } on \,\partial\Omega.$$

We then have, ([24, Theorems 3.3 and 3.4])

Theorem 2.3. *Assume that Ω is $(k-1)$-convex, $\partial\Omega \in C^{3,1}$, $f \in C^{1,1}(\overline{\Omega})$, inf $f > 0$, $g \in C^{3,1}(\overline{\Omega})$. Then there is a unique k-admissible solution $u \in C^{3,1}(\overline{\Omega})$ to the Dirichlet problem* (1.1).

We will need some identities for the k-Hessian operator $S_k(D^2u)$ which are derived explicitly for example in [13, p. 5–6]. See also [24]. For a symmetric matrix $A = (a_{ij})_{i,j} = 1, \ldots, n$ with eigenvalues $\lambda_i, i = 1, \ldots, n$, let us also denote by $S_k(A)$ the k-th elementary symmetric polynomial of λ. This is equivalent to say that $S_k(A)$ is the sum of all $k \times k$ principal minors of A. Using the permutation definition of the determinant, we have

$$(2.4) \qquad S_k(A) = \frac{1}{k!} \sum_{1 \le i_1, \cdots, i_k \le n} \delta^{j_1, \cdots, j_k}_{i_1, \cdots, i_k} a_{i_1 j_1} \cdots a_{i_k j_k},$$

where $\delta^{j_1, \cdots, j_k}_{i_1, \cdots, i_k}$ is the generalized Kronecker delta which takes the value $+1$ if $i_1, \cdots, i_k$ differs from $j_1, \cdots, j_k$ by an even permutation and the value -1 in the case of an odd permutation. In other words, for a choice of $i_1, \ldots, i_k$, $\delta^{j_1, \cdots, j_k}_{i_1, \cdots, i_k}$ is the signature of the permutation σ defined by $\sigma(i_l) = j_l, l = 1, \ldots, k$. This implies that we only consider the case where the sets $\{i_1, \ldots, i_k\}$ and $\{j_1, \ldots, j_k\}$ are identical. Moreover we define $\delta^{j_1, \cdots, j_k}_{i_1, \cdots, i_k}$ to be 0 if $\{i_1, \ldots, i_k\} \ne \{j_1, \ldots, j_k\}$. Note also that $\{i_1, \ldots, i_k\}$ is a subset of k elements of $\{1, \ldots, n\}$.

We have

$$S_k^{ij}(A) := \frac{\partial}{\partial a_{ij}} S_k(A) = \frac{1}{(k-1)!} \sum_{1 \le i, i_1, \cdots, i_{k-1} \le n} \delta^{j, j_1, \cdots, j_{k-1}}_{i, i_1, \cdots, i_{k-1}} a_{i_1 j_1} \cdots a_{i_{k-1} j_{k-1}},$$

and so $S_k(A) = \frac{1}{k} \sum_{i,j=1}^n S_k^{ij}(A) a_{ij}$ by the k-homogeneity of S_k and Euler's theorem for homogeneous functions. Here $\{j_1, \ldots, j_{k-1}\}$ is the image of the set of $k-1$ elements $\{i_1, \ldots, i_{k-1}\}$ not containing i by a permutation.

Let us denote by $\{S_k^{ij}(A)\}$ the symmetric matrix with entries $S_k^{ij}(A)$. We can write $S_k(A) = 1/k \, \{S_k^{ij}(A)\} : A$, that is $S_k(D^2v) = \frac{1}{k}\{S_k^{ij}(D^2v)\} : D^2v$. Using (2.4) and observing that the expression of $S_k(A)$ can be written in terms of a multilinear map, we obtain

$$(2.5) \qquad S_k'(D^2v)D^2w = \{S_k^{ij}(D^2v)\} : D^2w.$$

Let us denote by $\{S_k^{ij}(A)\}'$ the Fréchet derivative of the mapping $A \to \{S_k^{ij}(A)\}$. Since $\{S_k^{ij}(A)\}'(B)$ is a sum of terms each of which is a product of $k-2$ terms from A and is linear in B, we have

$$(2.6) \qquad \|\{S_k^{ij}(D^2v)\}'D^2w\|_{0;\Omega} \le C|v|_{2;\Omega}^{k-2}|w|_{2;\Omega}.$$

Using (2.2) and (2.6) we also have

$$(2.7) \qquad \|\{S_k^{ij}(D^2v)\}'D^2w\|_{0,\alpha;\Omega} \le C|v|_{2,\alpha;\Omega}^{k-2}|w|_{2,\alpha;\Omega}.$$

Finally we note that

Lemma 2.4. *Let v be a C^2 strictly convex function with Hessian having smallest eigenvalue uniformly bounded below by a constant $a > 0$. Then for $\eta = a/(2n)$, we have w strictly convex, whenever $\|w - v\|_{C^2(\Omega)} < \eta$.*

Proof. It follows from [16, Theorem 1 and Remark 2 p. 39] that for two symmetric $n \times n$ matrices A and B,

(2.8) $$|\lambda_l(A) - \lambda_l(B)| \leq n \max_{i,j} |A_{ij} - B_{ij}|, l = 1, \ldots, n.$$

It follows that for $u, v \in C^2(\Omega)$,

(2.9) $$|\lambda_1(D^2 u(x)) - \lambda_1(D^2 v(x))| \leq n\|w - v\|_{C^2(\Omega)}.$$

The result then follows. $\qquad\square$

We conclude this section with the equivalence of n-convexity and convexity in the usual sense.

Lemma 2.5. *A C^2 function u is convex if and only if it is n-convex.*

Proof. If u is C^2, $\lambda_i \geq 0$ on Ω for all i and thus $S_l(D^2 u) \geq 0, l = 1, \ldots, n$.

Conversely let us assume that A is a symmetric matrix with $S_l(A) \geq 0, l = 1, \ldots, n$. We show that its eigenvalues λ_i are all positive. Let

$$p(\lambda) = \lambda^n + c_1 \lambda^{n-1} + \ldots + c_n,$$

denote the characteristic polynomial of A. It can be shown [17, Theorem 1.2.12] that

$$c_l = (-1)^l S_l(A), l = 1, \ldots, n.$$

We show that if $\lambda_i < 0$ then $p(\lambda_i) \neq 0$. We have

$$
\begin{aligned}
p(\lambda_i) &= \lambda_i^n + c_1 \lambda_i^{n-1} + \ldots + c_n \\
&= \lambda_i^n + \sum_{l=1}^{n} (-1)^l S_l(A) \lambda_i^{n-l} \\
&= (-1)^n \left((-\lambda_i)^n + \sum_{l=1}^{n} (-1)^{l-n} S_l(A) \lambda_i^{n-l} \right) \\
&= (-1)^n \left((-\lambda_i)^n + \sum_{l=1}^{n} S_l(A)(-\lambda_i)^{n-l} \right).
\end{aligned}
$$

Since $-\lambda_i > 0$ and $S_l(A) \geq 0$ for all l, we have $(-1)^n p(\lambda_i) \geq 0$. Moreover since $\sum_{l=1}^{n} S_l(A)(-\lambda_i)^{n-l} \geq 0$ and $-\lambda_i > 0$ we have $(-1)^n p(\lambda_i) \neq 0$. We conclude that $\lambda_i \geq 0$ for all i. This completes the proof. $\qquad\square$

2.2. **Discrete Schauder estimates and related tools.** We will study the numerical approximation of (1.1)–(1.3) by standard finite difference discretizations. For simplicity, we consider a cuboidal domain $\Omega = (0,1)^n \subset \mathbb{R}^n$. Let $0 < h < 1$ with $1/h \in \mathbb{Z}$. Put

$$
\begin{aligned}
\mathbb{Z}_h &= \{x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n : x_i/h \in \mathbb{Z}\} \\
\Omega_0^h &= \Omega \cap \mathbb{Z}_h, \Omega^h = \overline{\Omega} \cap \mathbb{Z}_h, \partial\Omega^h = \partial\Omega \cap \mathbb{Z}_h = \Omega^h \setminus \Omega_0^h.
\end{aligned}
$$

Let $e^i, i = 1, \ldots, n$ denote the i-th unit vector of $\mathbb{R}^n$. We define the following first order difference operators on the space $\mathcal{M}(\Omega^h)$ of grid functions $v^h(x), x \in \mathbb{Z}_h$,

$$\partial_+^i v^h(x) := \frac{v^h(x + he^i) - v^h(x)}{h},$$

$$\partial_-^i v^h(x) := \frac{v^h(x) - v^h(x - he^i)}{h},$$

$$\partial_h^i v^h(x) := \frac{v^h(x + he^i) - v^h(x - he^i)}{2h}.$$

Higher order difference operators are obtained by combining the above difference operators. For a multi-index $\beta = (\beta_1, \ldots, \beta_n) \in \mathbb{N}^n$, we define

$$\partial_+^\beta v^h := \partial_+^{\beta_1} \cdots \partial_+^{\beta_n} v^h.$$

The operators ∂_-^β and ∂_h^β are defined similarly. Note that

(2.10) $$\partial_+^i \partial_-^i v^h(x) = \frac{v^h(x + he^i) - 2v^h(x) + v^h(x - he^i)}{h^2},$$

(2.11)
$$\partial_h^i \partial_h^j v^h(x) = \frac{1}{4h^2} \left\{ v^h(x + he^i + he^j) + v^h(x - he^i - he^j) \right.$$
$$\left. - v^h(x + he^i - he^j) - v^h(x - he^i + he^j) \right\}, i \neq j.$$

The second order derivatives $\partial^2 v / \partial x_i \partial x_j$ are discretized using (2.10) and (2.11) for $i \neq j$. This gives a discretization of the Hessian $D^2 u$ which we denote by $\mathcal{H}_d(u^h)$.

Thus the discrete version of (1.1) takes the form

(2.12) $$S_k(\mathcal{H}_d u^h(x)) = f(x), x \in \Omega_0^h, u^h(x) = g(x) \text{ on } \partial\Omega^h.$$

The discrete Laplacian takes the form

(2.13) $$\Delta_d(u^h) = \sum_{i=1}^n \partial_+^i \partial_-^i u^h.$$

We consider a discrete uniformly elliptic linear operator with low order terms

$$L_d v^h(x) = \sum_{i,j=1}^n a^{ij}(x) \partial_-^i \partial_+^j v^h(x) + \sum_{i=1}^n b^i(x) \partial_+^i v^h(x), x \in \Omega_0^h,$$

i.e. the matrix $(a^{ij}(x))_{i,j=1,\ldots,n}$ is uniformly positive definite. We now define discrete analogues of the Hölder norms and semi-norms following [18]. Let $[\xi, \eta]$ denote the set of points $\zeta \in \Omega^h$ such that $\xi_j \leq \zeta_j \leq \eta_j, j = 1, \ldots, n$. Then for $v^h \in \mathcal{M}(\Omega^h), 0 < \alpha < 1$, we define

$$|v^h|_{j;\Omega_0^h} = \max \left\{ |\partial_+^\beta v^h(\xi)|, |\beta| = j, [\xi, \xi + \beta] \subset \Omega^h \right\},$$

$$[v^h]_{j,\alpha;\Omega_0^h} = \max \left\{ \frac{|\partial_+^\beta v^h(\xi) - \partial_+^\beta v^h(\eta)|}{(|\xi - \eta|)^\alpha}, |\beta| = j, \xi \neq \eta, [\xi, \xi + \beta] \cup [\eta, \eta + \beta] \subset \Omega^h \right\},$$

$$||v^h||_{p;\Omega_0^h} = \max_{j \leq p} |v^h|_{j;\Omega_0^h},$$

$$||v^h||_{p,\alpha;\Omega_0^h} = ||v^h||_{p;\Omega_0^h} + [v^h]_{p,\alpha;\Omega_0^h}.$$

The above norms are extended canonically to vector fields and matrix fields by taking the maximum over all components. For $j = 0$, we have discrete analogues of the maximum and $C^{0,\alpha}$ norms.

For a domain $O \subset \mathbb{R}^n$, we denote by $\mathcal{D}_h(O)$ the set of mesh functions on $\mathbb{R}^n$ which vanish outside O. If $v^h = 0$ on $\partial\Omega^h$, extending v^h by 0 to $\mathbb{Z}_h$, we obtain $v^h \in \mathcal{D}_h(\Omega)$. The following theorem then follows from [23, Lemma 3.4].

Theorem 2.6. *Assume $0 < \alpha < 1$ and $v^h = 0$ on $\partial\Omega^h$. Then there are constants C and h_0 such that for $v^h \in \mathcal{M}(\Omega^h), h \leq h_0$*

(2.14)
$$||v^h||_{2,\alpha;\Omega_0^h} \leq C||L_d\, v^h||_{0,\alpha;\Omega_0^h},$$

with the constant C independent of h.

Since

$$\partial_+^i \partial_-^i v^h(x) = \partial_+^i \partial_+^i v^h(x - he^i) \,\text{and}$$

$$\partial_h^j \partial_h^i v^h(x) = \frac{1}{4}\Big(\partial_+^j \partial_+^i v^h(x) + \partial_+^j \partial_+^i v^h(x - he^i) + \partial_+^j \partial_+^i v^h(x - he^j)$$

$$+ \partial_+^j \partial_+^i v^h(x - he^i - he^j)\Big),$$

we have $\max\{||\partial_+^i \partial_-^i v^h||_{0,\alpha;\Omega_0^h}, ||\partial_h^j \partial_h^i v^h||_{0,\alpha;\Omega_0^h}, i,j = 1, \ldots, n\} \leq ||v^h||_{2,\alpha;\Omega_0^h}$ and hence the above theorem also applies when the second order derivatives (2.10) and (2.11) are used in the definition of $||.||_{2,\alpha;\Omega_0^h}$.

By Taylor series expansions, it is not difficult to verify that for $v \in C^2(\Omega)$

$$|v|_{j;\Omega_0^h} \leq |v|_{2;\Omega}, j \leq 2.$$

Moreover, for $v \in C^{4,\alpha}(\Omega)$,

(2.15)
$$||D^2 v - \mathcal{H}_d(v)||_{0;\Omega_0^h} \leq Ch^2 |v|_{4;\Omega},$$

and

$$[D^2 v - \mathcal{H}_d(v)]_{0,\alpha;\Omega_0^h} \leq Ch^2 [v]_{4,\alpha;\Omega}.$$

To see that the last inequality holds, it is enough to consider a function of one variable $v \in C^{4,\alpha}(-1, 1)$ and estimate $[v''(x) - (v(x + h) - 2v(x) + v(x - h))/h^2]_{0,\alpha}$. Now,

$$v''(x) - \frac{v(x+h) - 2v(x) + v(x-h)}{h^2} = -\frac{h^2}{24}(v^{(4)}(x+t_1 h) + v^{(4)}(x-t_2 h)), t_1, t_2 \in [0, 1].$$

Next we note that, using the definition, the $C^{0,\alpha}$ norm of $v^{(4)}(x + t_1 h)$ is bounded above by the $C^{0,\alpha}$ norm of $v^{(4)}$. The result then follows.

We have for $v \in C^{4,\alpha}(\Omega)$,

(2.16)
$$||D^2 v - \mathcal{H}_d(v)||_{0,\alpha;\Omega_0^h} \leq Ch^2 ||v||_{4,\alpha;\Omega}.$$

Lemma 2.7. *We have for $u \in C^{4,\alpha}(\Omega)$*

$$||S_k(D^2 u) - S_k(\mathcal{H}_d(u))||_{0,\alpha;\Omega_0^h} \leq Ch^2 |u|_{2;\Omega}^{k-1} ||u||_{4,\alpha;\Omega}.$$

Proof. By the mean value theorem, using (2.5), we have for some t in $[0,1]$, and $x \in \Omega_0^h$,

$$S_k(D^2u)(x) - S_k(\mathcal{H}_d(u))(x) = S_k'(tD^2(u)(x) + (1-t)\mathcal{H}_d(u)(x)) : (D^2u(x)$$
$$- \mathcal{H}_d(u)(x))$$

$$= \sum_{i,j=1}^{n} S_k^{ij}(tD^2(u)(x) + (1-t)\mathcal{H}_d(u)(x))(D^2u(x)$$
$$- \mathcal{H}_d(u)(x))_{ij}.$$

Using (2.2), it follows that

$$||S_k(D^2u) - S_k(\mathcal{H}_d(u))||_{0,\alpha;\Omega_0^h} \leq C(|u|_{2;\Omega} + |u|_{2;\Omega_0^h})^{k-1}||D^2u - \mathcal{H}_d(u)||_{0,\alpha;\Omega_0^h}$$
$$\leq Ch^2|u|_{2;\Omega}^{k-1}||u||_{4,\alpha;\Omega}.$$

$\square$

3. APPROXIMATIONS BY LINEAR ELLIPTIC PROBLEMS

In this section, we prove the convergence of the iterative method (1.2) and its discrete version. As indicated in the introduction, we also obtain the existence and uniqueness of the solution of the discrete version of (1.1), i.e. (2.12), as well as error estimates.

3.1. **Convergence at the operator level.** We assume that there is a unique k-admissible solution $u \in C^{2,\alpha}(\overline{\Omega})$ of (1.1) for $0 < \alpha < 1$. Let $u^0 \in C^{2,\alpha}(\overline{\Omega})$ such that

$$(3.1) \qquad\qquad ||u - u^0||_{2,\alpha;\Omega} < \delta.$$

For $k = n$, using an eigenvalue argument, it is not difficult to prove that the cofactor matrix is uniformly positive definite under the assumption $f \geq f_0 > 0$ for a constant f_0. We assume that the matrix $\{S_k^{ij}(D^2u)\}$ is uniformly positive definite. We claim that this holds if $u \in C^2(\overline{\Omega})$ and there is $c_3 > 0$ such that

$$c_3 \leq S_l(D^2u), 1 < l \leq k.$$

We then have

$$(3.2) \qquad\qquad c_3 \leq S_l(D^2u) \leq c_4, 1 < l \leq k,$$

for a constant c_4. The proof is essentially given as [13, Theorem 1.3]. We define

$$S_k^i(\lambda) := \frac{\partial}{\partial \lambda_i} S_k(\lambda).$$

First we note from the proof of [13, Theorem 1.3] that the eigenvalues of $\{S_k^{ij}(D^2u)\}$ are given by $S_k^i(\lambda(D^2u)), 1 \leq i \leq n$. On the other hand, since $S_l(D^2u) \geq c_3 > 0, 1 < l \leq k$, we have by [6, Proposition 1.1]

$$\frac{\partial}{\partial \lambda_i} S_k(\lambda)^{\frac{1}{k}} > 0 \text{ for } \lambda = \lambda(D^2u).$$

Finally, as $S_k(D^2u) \leq c_4$ and $u \in C^2(\overline{\Omega})$, the result follows.

By the continuity of the smallest eigenvalue of a matrix as a function of its entries, $\{S_k^{ij}(D^2u^0)\}$ is also uniformly positive definite for $|u - u^0|_{2;\Omega}$ sufficiently small.

Next, $\{S_k^{ij}(D^2u^0)\}$ is a symmetric matrix and divergence free by [13, Formula 1.10]. Thus we obtain

$$(3.3) \qquad \operatorname{div}\left(\{S_k^{ij}(D^2u^0)\}Dv)\right) = \{S_k^{ij}(D^2u^0)\} : D^2v.$$

We have

Theorem 3.1. *Under the assumption that there is a unique k-admissible solution $u \in C^{2,\alpha}(\overline{\Omega})$ of (1.1) for $0 < \alpha < 1$, the sequence defined by (1.2) converges to u for u^0 sufficiently close to u.*

Proof. We define the operator $R : C^{2,\alpha}(\overline{\Omega}) \to C^{2,\alpha}(\overline{\Omega})$ by

$$-\operatorname{div}\left(\{S_k^{ij}(D^2u^0)\}D(v - Rv)\right) = -S_k(D^2v) + f \text{ in } \Omega$$

$$R(v) = g \text{ on } \partial\Omega.$$

By Theorem 2.1, the operator R is well defined. We show that for $\rho > 0$ sufficiently small, R is a strict contraction in the ball $B_\rho(u) = \{v \in C^{2,\alpha}(\overline{\Omega}), ||u - v||_{2,\alpha;\Omega} < \rho\}$. For $v, w \in B_\rho(u)$ we have using (3.3)

$$\operatorname{div}\left(\{S_k^{ij}(D^2u^0)\}D(Rv - Rw)\right) = \operatorname{div}\left(\{S_k^{ij}(D^2u^0)\}D(v - w)\right) + S_k(D^2w) - S_k(D^2v)$$

$$= -\{S_k^{ij}(D^2u^0)\} : (D^2w - D^2v) + S_k(D^2w) - S_k(D^2v).$$

Next, by the mean value theorem and using (2.5), we have for some t in $[0, 1]$,

$$S_k(D^2w) - S_k(D^2v) = \{S_k^{ij}(tD^2w + (1-t)D^2v)\} : D^2(w - v)$$

$$= \{S_k^{ij}(t(D^2w - D^2u^0) + (1-t)(D^2v - D^2u^0) + D^2u^0)\} : D^2(w - v).$$

We use (2.7) to estimate the $C^{0,\alpha}$ norm of

$$A = \{S_k^{ij}(t(D^2w - D^2u^0) + (1-t)(D^2v - D^2u^0) + D^2u^0)\} - \{S_k^{ij}(D^2u^0)\}.$$

For $0 \leq s \leq 1$ to be specified below, put

$$\alpha_{st} = st(D^2w - D^2u^0) + s(1-t)(D^2v - D^2u^0) + D^2u^0.$$

We have

$$(3.4) \qquad |\alpha_{st}|_{0,\alpha;\Omega} \leq ||u^0 - v||_{2,\alpha;\Omega} + ||u^0 - w||_{2,\alpha;\Omega} + ||u^0||_{2,\alpha;\Omega}.$$

By the mean value theorem, for some $s \in [0, 1]$ we have

$$A = \{S_k^{ij}(\alpha_{st})\}'(t(D^2w - D^2u^0) + (1-t)(D^2v - D^2u^0)),$$

and thus by (2.7)

$$(3.5) \qquad ||A||_{0,\alpha;\Omega} \leq C|\alpha_{st}|_{0,\alpha;\Omega}^{k-2}(||u^0 - v||_{2,\alpha;\Omega} + ||u^0 - w||_{2,\alpha;\Omega}).$$

By Schauder estimates (Theorem 2.1), (2.2), (3.4) and (3.5) we obtain

$$\|R(v) - R(w)\|_{2,\alpha;\Omega} \leq C\|A\|_{0,\alpha;\Omega}\|D^2(v-w)\|_{0,\alpha;\Omega}$$

$$\leq C(\|u^0 - v\|_{2,\alpha;\Omega} + \|u^0 - w\|_{2,\alpha;\Omega} + \|u^0\|_{2,\alpha;\Omega})^{k-2}$$

(3.6)
$$(\|u^0 - v\|_{2,\alpha;\Omega} + \|u^0 - w\|_{2,\alpha;\Omega})\|v - w\|_{2,\alpha;\Omega}$$

$$\leq C(\rho + \delta + \|u^0\|_{2,\alpha;\Omega})^{k-2}(\rho + \delta)\|v - w\|_{2,\alpha;\Omega},$$

where δ measures how close u^0 is to u (3.1). Thus, for ρ and δ sufficiently small, R is a strict contraction in $B_\rho(u)$.

It remains to show that R maps $B_\rho(u)$ into itself. We note by the definition of R and unicity of the solution of (1.1), a fixed point of R solves (1.1). Let $v \in B_\rho(u)$,

$$\|u - Rv\|_{2,\alpha;\Omega} = \|Ru - Rv\|_{2,\alpha;\Omega} \leq \|u - v\|_{2,\alpha;\Omega} \leq \rho,$$

which shows that R maps $B_\rho(u)$ into itself. The existence of a fixed point follows from the Banach fixed point theorem. Moreover, the sequence defined by $u^{m+1} = R(u^m)$, i.e. the sequence defined by (1.2), converges for ρ and δ sufficiently small to u.   □

3.2. **Finite difference discretization.** Next we consider the following discrete version of (1.2)

$$\{S_k^{ij}(\mathcal{H}_d\, u^{0,h})\} : \mathcal{H}_d u^{m+1,h} = \{S_k^{ij}(\mathcal{H}_d\, u^{0,h})\} : \mathcal{H}_d u^{m,h}$$

(3.7)
$$+ f - S_k(\mathcal{H}_d\, u^{m,h}) \text{ in } \Omega_0^h$$

$$u^{m+1,h} = g \text{ on } \partial\Omega^h.$$

Under the assumptions of Theorem 3.5 below, we show that (2.12) has a unique solution to which the above sequence converges. Moreover, the convergence rate is $O(h^2)$. Define

(3.8) $$B_\rho(u) = \{v^h \in \mathcal{M}(\Omega^h), \|v^h - u\|_{2,\alpha;\Omega_0^h} \leq \rho\}.$$

Lemma 3.2. *Let $S^h : \mathcal{M}(\Omega^h) \to \mathcal{M}(\Omega^h)$ be a strict contraction with contraction factor less than 1/2, i.e. for $v^h, w^h \in \mathcal{M}(\Omega^h)$*

$$\|S^h(v^h) - S^h(w^h)\|_{2,\alpha;\Omega_0^h} \leq \frac{1}{2}\|v^h - w^h\|_{2,\alpha;\Omega_0^h}.$$

Let us also assume that S^h does not move the center u of the ball $B_\rho(u)$ too far, i.e.

$$\|S^h(u) - u\|_{2,\alpha;\Omega_0^h} \leq C_0 h^2.$$

Then S^h maps $B_\rho(u)$ into itself for $\rho = 2C_0 h^2$. Moreover S^h has a unique fixed point u^h in $B_\rho(u)$ with the error estimate

$$\|u - u^h\|_{2,\alpha;\Omega_0^h} \leq 2C_0 h^2.$$

Proof. For $v^h \in B_\rho(u)$,

$$\|S^h(v^h) - u\|_{2,\alpha;\Omega_0^h} \leq \|S^h(v^h) - S^h(u)\|_{2,\alpha;\Omega_0^h} + \|S^h(u) - u\|_{2,\alpha;\Omega_0^h}$$

$$\leq \frac{1}{2}\|v^h - u\|_{2,\alpha;\Omega_0^h} + C_0 h^2$$

$$\leq \frac{\rho}{2} + C_0 h^2 \leq \frac{\rho}{2} + \frac{\rho}{2} = \rho.$$

This proves that S^h maps $B_\rho(u)$ into itself. The existence of a fixed point follows from the Banach fixed point theorem. The convergence rate follows from the observation that

$$||u - u^h||_{2,\alpha;\Omega_0^h} \leq ||u - S^h(u)||_{2,\alpha;\Omega_0^h} + ||S^h(u) - S^h(u^h)||_{2,\alpha;\Omega_0^h}$$

$$\leq C_0 h^2 + \frac{1}{2}||u^h - u||_{2,\alpha;\Omega_0^h}.$$

$\square$

Remark 3.3. *For h sufficiently small, $\mathcal{H}_d(u)$ is sufficiently close to D^2u and hence $\{S_k^{ij}(\mathcal{H}_d u)\}$ is positive definite, a property which also holds for $\{S_k^{ij}(\mathcal{H}_d u^{0,h})\}$ for $u^{0,h}$ sufficiently close to u. The arguments are similar to the ones of Lemma 2.4. See also Lemma 3.4 below.*

Lemma 3.4. *Let u be a k-admissible solution of (1.1). Assume that $\inf f > 0$ and $u \in C^4(\Omega)$. Then for h sufficiently small, $\Delta_d(u) \geq c_0 > 0$ where $c_0 = 1/2((\inf f)/c(k,n))^{1/k}$. Moreover, if u is a strictly convex function, then for $\rho = O(h^2)$, $\mathcal{H}_d(u)$ is a positive matrix and v^h is a discrete convex function, when $v^h \in B_\rho(u)$.*

Proof. Since the eigenvalues of a matrix are continuous functions of its entries (as roots of the characteristic polynomial), for a matrix $A = (a_{ij})$ with $S_k(A) > 0$, we have for $\epsilon > 0$, the existence of $\gamma > 0$ depending only on the space dimension n such that $|S_k(B) - S_k(A)| < \epsilon$ when $\sup_{ij}|b_{ij} - a_{ij}| < \gamma$. This implies $S_k(B) > S_k(A) - \epsilon$. Thus with $\epsilon = S_k(A)/2$, we have $S_k(B) > S_k(A)/2$.

For h sufficiently small we have $Ch^2|u|_{4;\Omega} < \gamma$ and thus since $S_k(D^2u) = f > \inf f > 0$, by (2.15) $S_k(\mathcal{H}_d(u)) \geq 1/2\inf f$. By (1.4)

$$\Delta_d(u) \geq \frac{1}{2}((\inf f)/c(k,n))^{1/k}.$$

Let $v^h \in B_\rho(u)$. Then by definition of $B_\rho(u)$ and (2.15)

$$||\mathcal{H}_d(v^h) - \mathcal{H}_d(u)||_{0,\alpha;\Omega_0^h} \leq ||\mathcal{H}_d(v^h) - D^2u||_{0,\alpha;\Omega_0^h} + ||D^2u - \mathcal{H}_d(u)||_{0,\alpha;\Omega_0^h}$$

$$\leq \rho + Ch^2|u|_{4;\Omega},$$

which can be made smaller than γ for h and ρ sufficiently small. Thus given that $\mathcal{H}_d(u)$ is positive definite, the same holds for $\mathcal{H}_d(v^h)$. $\square$

Theorem 3.5. *Assume that $u \in C^{4,\alpha}(\overline{\Omega})$ is k-admissible. Choose $u^{0,h}$ such that $||u^{0,h} - u||_{2,\alpha;\Omega_0^h} = O(h^2)$. For h sufficiently small, (2.12) has a locally unique solution u^h which satisfies $\Delta_d(u^h) \geq 0$ and u^h converges to the unique solution u of (1.1) as $h \to 0$ at the rate $O(h^2)$.*

Proof. It follows from the assumptions that $\inf f > 0$. We define the operator $R^h : \mathcal{M}(\Omega^h) \to \mathcal{M}(\Omega^h)$ by

$$-\{S_k^{ij}(\mathcal{H}_d u^{0,h})\} : \mathcal{H}_d(v^h - R^h v^h) = -S_k(\mathcal{H}_d v^h) + f \text{ in } \Omega_0^h$$

$$R^h(v^h) = g \text{ on } \partial\Omega^h,$$

and show that R^h has a unique fixed point in $B_\rho(u)$ for $\rho = O(h^2)$. By Remark 3.3 the above problem is then well defined. It follows from (3.3) that the operator R^h is a discrete version of the operator R used in the proof of Theorem 3.1. Thus, as in the proof of Theorem 3.1 we obtain

$$\{S_k^{ij}(\mathcal{H}_d\, u^{0,h})\} : \mathcal{H}_d(R^h v^h - R^h w^h) = S_k(\mathcal{H}_d\, w^h) - S_k(\mathcal{H}_d\, v^h)$$
$$+ \{S_k^{ij}(\mathcal{H}_d\, u^{0,h})\} : \mathcal{H}_d\,(v^h - w^h).$$

And thus by the mean value theorem and discrete Schauder estimates, as in the proof of Theorem 3.1

(3.9) $\|R^h(v^h) - R^h(w^h)\|_{2,\alpha;\Omega_0^h} \le$
$$C(\rho + \delta_h + \|u^{0,h}\|_{2,\alpha;\Omega_0^h})^{k-2}(\rho + \delta_h)\|v^h - w^h\|_{2,\alpha;\Omega_0^h}.$$

Next, note that with (2.16) applied to u one has $|u|_{2,\alpha;\Omega_0^h} \le C\|u\|_{4,\alpha;\Omega}$. It follows that $\|u^{0,h}\|_{2,\alpha;\Omega_0^h} \le \|u\|_{2,\alpha;\Omega_0^h} + \delta_h \le C\|u\|_{4,\alpha;\Omega} + \delta_h$. We recall that by assumption $\|u^{0,h} - u\|_{2,\alpha;\Omega_0^h} = O(h^2)$. Thus R^h is a strict contraction in $B_\rho(u)$ for $\rho = O(h^2)$ and h sufficiently small. Moreover, the contraction factor can be made smaller than $1/2$ by choosing h sufficiently small.

Since $f = S_k(D^2 u)$, by the discrete Schauder estimates Theorem 2.6 and Lemma 2.7

$$\|R^h(u) - u\|_{2,\alpha;\Omega_0^h} \le C\|S_k(D^2 u) - S_k(\mathcal{H}_d(u))\|_{0,\alpha;\Omega_0^h} \le Ch^2.$$

By Lemma 3.2 we conclude that R^h has a fixed point u^h in $B_\rho(u)$ with the claimed convergence rate.

The claimed property of u^h follows from the fact that $u^h \in B_\rho(u)$ and Lemma 3.4.

$\square$

4. Newton's method

As in the previous section, we assume that $\{S_k^{ij}(D^2 u)\}$ is uniformly positive definite. By Remark 3.3, for h sufficiently small, there exists $m' > 0$ such that for $v^h \in B_\rho(u)$, $\{S_k^{ij}(\mathcal{H}_d v^h)\}$ has smallest eigenvalue greater than m'. We consider for $u^{0,h} \in B_\rho(u)$ the sequence of iterates

(4.1)
$$\{S_k^{ij}(\mathcal{H}_d u^{m,h})\} : (\mathcal{H}_d u^{m+1,h} - \mathcal{H}_d u^{m,h}) = f - S_k(\mathcal{H}_d u^{m,h}) \text{ in } \Omega_0^h$$
$$u^{m+1,h} = g \text{ in } \partial\Omega^h.$$

We note that (4.1) defines $u^{m+1,h}$ as the solution of a discrete second order elliptic equation in non divergence form, which is uniformly elliptic for $u^{m,h} \in B_\rho(u)$ for h sufficiently small.

Theorem 4.1. *The sequence defined by* (4.1) *satisfies*

(4.2)
$$\|u^{m+1,h} - u^h\|_{2,\alpha;\Omega_0^h} \le C\|u^{m,h} - u^h\|^2_{2,\alpha;\Omega_0^h},$$

for ρ and h sufficiently small and where u^h denotes the solution of (2.12) *in $B_\rho(u), \rho = O(h^2)$.*

Proof. Put

(4.3) $$B = \{S_k^{ij}(\mathcal{H}_d u^{m,h})\} : (\mathcal{H}_d u^{m+1,h} - \mathcal{H}_d u^h).$$

We have by (2.12)

$$B = \{S_k^{ij}(\mathcal{H}_d u^{m,h})\} : (\mathcal{H}_d u^{m,h} - \mathcal{H}_d u^h) + S_k(\mathcal{H}_d u^h) - S_k(\mathcal{H}_d u^{m,h})$$

(4.4) $$= \left(\{S_k^{ij}(\mathcal{H}_d u^{m,h})\} - \{S_k^{ij}(\mathcal{H}_d u^h)\} \right) : (\mathcal{H}_d u^{m,h} - \mathcal{H}_d u^h)$$

$$+ \{S_k^{ij}(\mathcal{H}_d u^h)\} : (\mathcal{H}_d u^{m,h} - \mathcal{H}_d u^h) + S_k(\mathcal{H}_d u^h) - S_k(\mathcal{H}_d u^{m,h}).$$

Put

(4.5) $$B_1 = \left(\{S_k^{ij}(\mathcal{H}_d u^{m,h})\} - \{S_k^{ij}(\mathcal{H}_d u^h)\} \right) : (\mathcal{H}_d u^{m,h} - \mathcal{H}_d u^h),$$

and

(4.6) $$B_2 = \{S_k^{ij}(\mathcal{H}_d u^h)\} : (\mathcal{H}_d u^{m,h} - \mathcal{H}_d u^h) + S_k(\mathcal{H}_d u^h) - S_k(\mathcal{H}_d u^{m,h}).$$

By the mean value theorem, (2.5) and (2.7), we have

$$B_1 = \left(\{S_k^{ij}(t\mathcal{H}_d u^{m,h} + (1-t)\mathcal{H}_d u^h)\}'(\mathcal{H}_d u^{m,h} - \mathcal{H}_d u^h) \right) : (\mathcal{H}_d u^{m,h} - \mathcal{H}_d u^h),$$

for $t \in [0,1]$ and thus

(4.7) $$\begin{aligned} ||B_1||_{0,\alpha;\Omega_0^h} &\leq C(||u^h||_{2,\alpha;\Omega_0^h} + ||u^{m,h}||_{2,\alpha;\Omega_0^h})^{k-2}||u^{m,h} - u^h||_{2,\alpha;\Omega_0^h}^2 \\ &\leq C(||u||_{2,\alpha;\Omega_0^h} + \rho)^{k-2}||u^{m,h} - u^h||_{2,\alpha;\Omega_0^h}^2 \\ &\leq C(||u||_{2,\alpha;\Omega_0^h} + \rho)^{k-2}||u^{m,h} - u^h||_{2,\alpha;\Omega_0^h}^2. \end{aligned}$$

We also have by the mean value theorem

$$\begin{aligned} B_2 &= \{S_k^{ij}(\mathcal{H}_d u^h)\} : (\mathcal{H}_d u^{m,h} - \mathcal{H}_d u^h) \\ &\quad + \{S_k^{ij}(t\mathcal{H}_d u^h + (1-t)\mathcal{H}_d u^{m,h})\} : (\mathcal{H}_d u^h - \mathcal{H}_d u^{m,h}) \\ &= \left(\{S_k^{ij}(\mathcal{H}_d u^h)\} - \{S_k^{ij}(t\mathcal{H}_d u^h + (1-t)\mathcal{H}_d u^{m,h})\} \right) : (\mathcal{H}_d u^{m,h} - \mathcal{H}_d u^h) \\ &= \left(\{S_k^{ij}((1-s)\mathcal{H}_d u^h + st\mathcal{H}_d u^h + s(1-t)\mathcal{H}_d u^{m,h})\}' \right. \end{aligned}$$

(4.8)

$$\left. ((1-t)(\mathcal{H}_d u^h - \mathcal{H}_d u^{m,h})) \right) : (\mathcal{H}_d u^{m,h} - \mathcal{H}_d u^h),$$

for $s,t \in [0,1]$. As for B_1 we obtain

(4.9) $$||B_2||_{0,\alpha;\Omega_0^h} \leq C(||u||_{2,\alpha;\Omega_0^h} + \rho)^{k-2}||u^{m,h} - u^h||_{2,\alpha;\Omega_0^h}^2.$$

Combining (4.3)–(4.8) and using Schauder estimates, we obtain (4.2). □

Choosing $\rho = O(h^2)$ we have $C\rho < 1$ for h sufficiently small. We conclude that $u^{m+1,h} \in B_\rho(u)$ when $u^{m,h} \in B_\rho(u)$ and the quadratic convergence rate of Newton's method.

Remark 4.2. *Having established that the discrete problem has a locally unique solution and that v^h is a discrete convex function for v^h sufficiently close to u, the convergence of Newton's method also follows from the verification of standard assumptions given in* [19, p. 68]. *See* [11] *for an example of verification of the standard assumptions for a wide stencil discretization.*

5. GAUSS-SEIDEL ITERATIVE METHODS

It is a natural idea to solve (2.12) by a nonlinear Gauss-Seidel method, that is solve (2.12) for $u^h(x)$ and solve the resulting nonlinear equations by a Gauss-Seidel method. Although this seems a daunting task for arbitrary k, we show that for $k = 2$, this takes a very elegant form. We then establish a connection between the resulting nonlinear Gauss-Seidel iterative method for 2-Hessian equations and the discrete version of (1.3), i.e.

$$(5.1) \qquad \Delta_d u^{m+1,h} = \left((\Delta_d u^{m,h})^k + \frac{1}{c(k,n)} (f - S_k(\mathcal{H}_d u^{m,h})) \right)^{\frac{1}{k}} \text{ in } \Omega_0^h$$
$$u^{m+1,h} = g \text{ on } \partial\Omega^h,$$

when the Gauss-Seidel method is used to solve the Poisson equations.

5.1. Nonlinear Gauss-Seidel method for 2-Hessian equations. We start with the identity

$$(5.2) \qquad \Delta_d u^h = \left((\Delta_d u^h)^2 + \frac{1}{c(2,n)} (f - S_2(\mathcal{H}_d u^h)) \right)^{\frac{1}{2}},$$

and show that the right hand side is independent of $u^h(x)$. Note that by (2.11), $\partial_h^i \partial_h^j u^h(x), i \neq j$ is independent of $u^h(x)$ and by (2.13),

$$\frac{\partial(\Delta_d u^h(x))}{\partial(u^h(x))} = \sum_{i=1}^{n} -\frac{2}{h^2} = -\frac{2n}{h^2}.$$

Since $\partial S_k(A)/\partial z = \sum_{i,j=1}^{n} (\partial S_k(A)/\partial a_{ij})(\partial a_{ij}/\partial z)$, we conclude that

$$\frac{\partial}{\partial(u^h(x))} S_2(\mathcal{H}_d u^h(x)) = \sum_{\substack{i,j=1 \\ i \neq j}}^{n} S_2^{ij}(\mathcal{H}_d u^h(x)) \frac{\partial}{\partial(u^h(x))} \partial_h^i \partial_h^j u^h(x)$$

$$+ \sum_{i=1}^{n} S_2^{ii}(\mathcal{H}_d u^h(x)) \frac{\partial}{\partial(u^h(x))} \partial_+^i \partial_-^i u^h(x)$$

$$= -\frac{2}{h^2} \sum_{i=1}^{n} S_2^{ii}(\mathcal{H}_d u^h(x)) = -\frac{2}{h^2} \sum_{i=1}^{n} \sum_{\substack{1 \leq p \leq n \\ p \neq i}} \delta_{ip}^{ip} \partial_+^p \partial_-^p u^h(x)$$

$$= -\frac{2}{h^2} \sum_{i=1}^{n} \sum_{p \neq i} \partial_+^p \partial_-^p u^h(x) = -\frac{2}{h^2}(n-1)\Delta_d u^h(x)$$

$$= -\frac{2}{h^2}(2n) c(2,n)\Delta_d u^h(x) = -\frac{4n}{h^2} c(2,n)\Delta_d u^h(x),$$

and we recall that the definition of δ_{ip}^{ip} was given in section 2.1. This gives

$$\frac{\partial}{\partial(u^h(x))}\left((\Delta_d u^h(x))^2 + \frac{1}{c(2,n)}(f - S_2(\mathcal{H}_d u^h(x)))\right) = 0.$$

We can therefore rewrite (5.2) as

$$
\begin{aligned}
(5.3) \qquad u^h(x) = \frac{h^2}{2n}\Bigg[&\sum_{i=1}^{n} \frac{u^h(x + he^i) + u^h(x - he^i)}{h^2} \\
&- \left((\Delta_d u^h(x))^2 + \frac{1}{c(2,n)}(f - S_2(\mathcal{H}_d u^h(x)))\right)^{\frac{1}{2}}\Bigg],
\end{aligned}
$$

where the solution with $\Delta_d u^h \geq 0$ has been selected. For $n = 2$, this is the identity which was solved in [15, 8, 7, 5] by a Gauss-Seidel iterative method, as indicated in the introduction. For $n \geq 3$, this provides new iterative methods for the 2-Hessian equations.

Henceforth, we shall assume that a row ordering of the elements of Ω^h is chosen. Note that if we apply the Gauss-Seidel method to the problem (5.1), we obtain a double sequence $u^{m,p,h}$ defined by

$$
\begin{aligned}
u^{m+1,p+1,h}(x) = \frac{h^2}{2n}\Bigg[&\sum_{i=1}^{n} \frac{u^{m+1,p,h}(x + he^i) + u^{m+1,p+1,h}(x - he^i)}{h^2} \\
&- \left((\Delta_d u^{m,h}(x))^2 + \frac{1}{c(2,n)}(f - S_2(\mathcal{H}_d u^{m,h}(x)))\right)^{\frac{1}{2}}\Bigg],
\end{aligned}
$$

This leads us to consider the double sequence $u_p^{m,h}$ defined by

$$
\begin{aligned}
u_{p+1}^{m+1,h}(x) = \frac{h^2}{2n}\Bigg[&\sum_{i=1}^{n} \frac{u_p^{m+1,h}(x + he^i) + u_{p+1}^{m+1,h}(x - he^i)}{h^2} \\
&- \left((\Delta_d u_{p*}^{m,h}(x))^2 + \frac{1}{c(2,n)}(f - S_2(\mathcal{H}_d u_{p*}^{m,h}(x)))\right)^{\frac{1}{2}}\Bigg],
\end{aligned}
$$

where $\Delta_d u_{p*}^{m,h}(x)$ and $S_2(\mathcal{H}_d u_{p*}^{m,h}(x))$ are the actions of the discrete Laplace and 2-Hessian operators on $u_p^{m,h}$ updated with the most recently computed values.

Formally, as $m \to \infty$, this gives the nonlinear Gauss-Seidel method

$$
\begin{aligned}
(5.4) \qquad u_{p+1}^h(x) = \frac{h^2}{2n}\Bigg[&\sum_{i=1}^{n} \frac{u_p^h(x + he^i) + u_{p+1}^h(x - he^i)}{h^2} \\
&- \left((\Delta_d u_{p*}^h(x))^2 + \frac{1}{c(2,n)}(f - S_2(\mathcal{H}_d u_{p*}^h(x)))\right)^{\frac{1}{2}}\Bigg],
\end{aligned}
$$

where as above $\Delta_d u_{p*}^h(x)$ and $S_2(\mathcal{H}_d u_{p*}^h(x))$ are the actions of the discrete Laplace and 2-Hessian operators on u_p^h updated with the most recently computed values of u_{p+1}^h. In particular, the right hand side of (5.4) does not depend on u_{p+1}^h since as shown above, the right hand side of (5.3) does not depend on $u^h(x)$.

6. Numerical results

We give numerical results for the σ_2 problem, i.e. for $k = 2, n = 3$ using the subharmonicity preserving iterations. Although our theoretical results only cover smooth solutions, as indicated in the abstract and in the introduction, the subharmonicity preserving iterations appear able to handle non smooth solutions. The initial guess in all of our numerical experiments is taken as the finite difference approximation of the solution of the Poisson equation $\Delta u = 2\sqrt{f}$ in Ω with $u = g$ on $\partial\Omega$.

We use the following test functions on the unit cube $[0, 1]^3$:

Test 1: A smooth solution which is strictly convex, $u(x, y, z) = e^{x^2+y^2+z^2}$ so that $f(x, y, z) = 4(3 + x^2 + y^2 + z^2)e^{2(x^2+y^2+z^2)}$ and $g(x, y, z) = e^{x^2+y^2+z^2}$ on $\partial\Omega$.

Test 2: A smooth solution which is 2-convex but not convex. It is known that for a radial function $u(x) = \phi(r), r = |x|, x \in \mathbb{R}^n$ the eigenvalues of D^2u are given by $\lambda_1 = \phi''(r)$ with multiplicity 1 and $\lambda_2 = \phi'(r)/r$ with multiplicity $n - 1$. See for example [9, Lemma 2.1]. It follows that with $u(x, y, z) = \ln(a + x^2 + y^2 + z^2)$, we have $\phi(r) = \ln(a + r^2)$ and we get $\Delta u = \frac{6a+2r^2}{(a+r^2)^2} \geq 0$, $S_2(D^2u) = 4\frac{3a-r^2}{(a+r^2)^3} \geq 0$, $\det D^2u = 2\frac{a-r^2}{(a+r^2)^2}$, in $[0, 1]^3$. With $a = 2$, $\det D^2u$ takes negative values in $[0, 1]^3$.

Test 3: A solution not in $H^2(\Omega)$, $u(x, y, z) = -\sqrt{3 - x^2 - y^2 - z^2}$ so that $f(x, y, z) = -(x^2 + y^2 + z^2 - 9)/(-3 + x^2 + y^2 + z^2)^2$ and $g(x, y, z) = -\sqrt{3 - x^2 - y^2 - z^2}$ on $\partial\Omega$.

Test 4: No exact solution is known. Here $f(x, y, z) = 1$ and $g(x, y, z) = 0$.

Test 5: A degenerate three dimensional Monge-Ampère equation. We take $f(x, y, z) = 0$ and $g(x, y, z) = |x - 1/2|$. We use the double iterative method based on (1.5).

Numerically, the solution computed may not satisfy $S^2D^2u^m \geq 0$. At those points we set both $S_2(D^2u^m)$ and $\det D^2u^m$ to 0 in (1.5). If the numerical value of $S_2(D^2u^m)$ is negative, then 0 is a better approximate value. Since $S_2(D^2u^m)$ is computed from u^m, the numerical value of $\det D^2u^m$ would also be inaccurate. Since u^m is expected to be an approximate solution of u for which $\det D^2u \geq 0$, a better approximation of $\det D^2u^m$ at any stage where the latter is negative is also 0. It would be interesting to analyze the effect of these rounding off errors on the overall numerical convergence of the method. For example, one may analyze the convergence of the inexact double iteration. Similar situations appear with inexact Newton's methods and inexact Uzawa algorithms.

The right hand side $f(x, y, z)$ can be computed from the exact solution $u(x, y, z)$ using the definition of $S_2(D^2u)$ as the sum of the 2×2 principal minors.

For all tests but Test 3, we used the direct solver (5.1). For $h = 2^6$, we run out of memory with (5.1). For Test 3, the Gauss-Seidel method was used since there is no memory issue for the latter with $h = 2^6$. As expected, we have quadratic convergence (as $h \to 0$) for the smooth solutions of Tests 1 and 2 while enough data is not available to give the convergence rate for the singular solution of Test 3.

	h				
	$1/2^1$	$1/2^2$	$1/2^3$	$1/2^4$	$1/2^5$
Error	$6.2328\ 10^{-2}$	$2.6556\ 10^{-2}$	$7.7836\ 10^{-3}$	$2.0616\ 10^{-3}$	$5.2449\ 10^{-4}$
Rate		1.23	1.77	1.92	1.97

TABLE 1. Maximum error with Test 1.

	h				
	$1/2^1$	$1/2^2$	$1/2^3$	$1/2^4$	$1/2^5$
Error	$6.5241\ 10^{-4}$	$5.0653\ 10^{-4}$	$1.3850\ 10^{-4}$	$3.5587\ 10^{-5}$	$9.1276\ 10^{-6}$
Rate		0.36	1.87	1.96	1.96

TABLE 2. Maximum error with Test 2.

	h		
	$1/2^4$	$1/2^5$	$1/2^6$
Error	$1.1084\ 10^{-3}$	$9.7971\ 10^{-4}$	$7.6618\ 10^{-4}$
Rate		0.18	0.35

TABLE 3. Maximum error with Test 3.
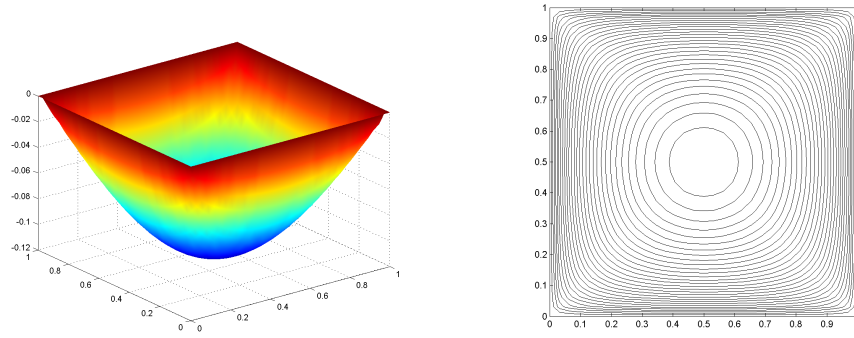
FIGURE 1. Test 4, $h = 1/2^5$. Graph and contour in plane $z = 1/2$.
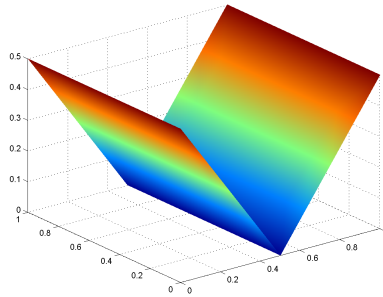
FIGURE 2. Test 5, $h = 1/2^4$. Graph in the plane $z = 1/2$.

In [5], it was argued based on numerical evidence that the Gauss-Seidel method (5.4) is faster than a certain variant of the direct solver (5.1) for singular solutions. In our implementation we saw evidence of the contrary, that is, the Gauss-Seidel method is less efficient. We note that the Gauss-Seidel method requires much more loops which are not efficient in MATLAB.

7. Concluding Remarks

Remark 7.1. *Although the pseudo-transient and time marching methods introduced in [3] work as well for k-Hessian equations, and apply to more general fully nonlinear equations, the subharmonicity preserving iterative methods introduced in this paper are parameter free. All these type of methods can be accelerated with fast Poisson solvers and multigrid methods.*

Remark 7.2. *When it comes to numerical methods for fully nonlinear equations, there are two types of convergence to study. Since the equations are nonlinear, they must be solved iteratively. One must then address the convergence to the discrete solution of the iterative methods used. The second type of convergence is the convergence of the numerical solution to the exact solution as the discretization parameter converges to 0. We have addressed both types of convergence in this paper.*

Remark 7.3. *Existence of a discrete solution and convergence (as the mesh size $h \to 0$), for finite difference discretization of smooth solutions of fully nonlinear equations, are not often discussed. It is clear that convergence does not simply follow from the consistency of standard finite difference discretization of the second order derivatives. For viscosity solutions, convergence of monotone, stable and consistent schemes follows immediately from the theory of Barles and Souganidis.*

Remark 7.4. *The iterative method (1.3) can be viewed as a linearization of the fully nonlinear equation (1.1). It is possible to linearize (1.1) in ways different from (1.2) and (1.3). See for example the methods described in [3]. The iterative method (1.3) has been shown numerically to select discrete solutions which converge to non smooth solutions. Since (1.3) consists of a sequence of Poisson equations, the numerical solution of (1.1) can now be tackled with any good numerical method.*

References

[1] N. E. Aguilera and P. Morin. Approximating optimization problems over convex functions. *Numer. Math.*, 111(1):1–34, 2008.
[2] G. Awanou. Convergence of a hybrid scheme for the elliptic Monge-Ampère equation. http://homepages.math.uic.edu/ awanou/up.html.
[3] G. Awanou. Pseudo transient continuation and time marching methods for Monge-Ampère type equations. *Adv. Comput. Math.*, 41(4):907–935, 2015.
[4] G. Awanou. On standard finite difference discretizations of the elliptic Monge-Ampère equation. *J. Sci. Comput.*, 69(2):892–904, 2016.

[5] J.-D. Benamou, B. D. Froese, and A. M. Oberman. Two numerical methods for the elliptic Monge-Ampère equation. *M2AN Math. Model. Numer. Anal.*, 44(4):737–758, 2010.

[6] L. Caffarelli, L. Nirenberg, and J. Spruck. The Dirichlet problem for nonlinear second-order elliptic equations. III. Functions of the eigenvalues of the Hessian. *Acta Math.*, 155(3-4):261–301, 1985.

[7] Y. Chen. *Efficient and robust solvers for Monge-Ampère equations*. PhD thesis, Clarkson University, 2010.

[8] Y. Chen and S. R. Fulton. An adaptive continuation-multigrid method for the balanced vortex model. *J. Comput. Phys.*, 229(6):2236–2248, 2010.

[9] P. L. Felmer and A. Quaas. On critical exponents for the Pucci's extremal operators. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 20(5):843–865, 2003.

[10] X. Feng and M. Neilan. Analysis of Galerkin methods for the fully nonlinear Monge-Ampère equation. *J. Sci. Comput.*, 47(3):303–327, 2011.

[11] B. Froese and A. Oberman. Convergent finite difference solvers for viscosity solutions of the elliptic Monge-Ampère equation in dimensions two and higher. *SIAM J. Numer. Anal.*, 49(4):1692–1714, 2011.

[12] B. D. Froese, A. M. Oberman, and T. Salvador. Numerical methods for the 2-Hessian elliptic partial differential equation. *IMA Journal of Numerical Analysis*, 2016.

[13] N. Gavitone. *Hessian equations, quermassintegrals and symmetrization*. PhD thesis, Universita di Napoli Federico II, 2009.

[14] D. Gilbarg and N. S. Trudinger. *Elliptic partial differential equations of second order*. Classics in Mathematics. Springer-Verlag, Berlin, 2001. Reprint of the 1998 edition.

[15] M. Headrick and T. Wiseman. Numerical Ricci-flat metrics on K3. *Classical and Quantum Gravity*, 22(23):4931–4960, DEC 7 2005.

[16] A. J. Hoffman and H. W. Wielandt. The variation of the spectrum of a normal matrix. *Duke Math. J.*, 20:37–39, 1953.

[17] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, 1990. Corrected reprint of the 1985 original.

[18] C. G. L. Johnson. Estimates near plane portions of the boundary for discrete elliptic boundary problems. *Math. Comp.*, 28:909–935, 1974.

[19] C. T. Kelley. *Iterative methods for linear and nonlinear equations*, volume 16 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1995. With separately available software.

[20] G. M. Lieberman. *Second order parabolic differential equations*. World Scientific Publishing Co. Inc., River Edge, NJ, 1996.

[21] G. Loeper and F. Rapetti. Numerical solution of the Monge-Ampère equation by a Newton's algorithm. *C. R. Math. Acad. Sci. Paris*, 340(4):319–324, 2005.

[22] D. C. Sorensen and R. Glowinski. A quadratically constrained minimization problem arising from PDE of Monge-Ampère type. *Numer. Algorithms*, 53(1):53–66, 2010.

[23] V. Thomée. On the convergence of difference quotients in elliptic problems. In *Numerical Solution of Field Problems in Continuum Physics (Proc. Sympos. Appl. Math., Durham, N.C., 1968), SIAM-AMS Proc., Vol. II*, pages 186–200. Amer. Math. Soc., Providence, R. I., 1970.

[24] X.-J. Wang. The k-Hessian equation. In *Geometric analysis and PDEs*, volume 1977 of *Lecture Notes in Math.*, pages 177–252. Springer, Dordrecht, 2009.

DEPARTMENT OF MATHEMATICS, STATISTICS, AND COMPUTER SCIENCE, M/C 249. UNIVERSITY OF ILLINOIS AT CHICAGO, CHICAGO, IL 60607-7045, USA