

PSEUDO TIME CONTINUATION AND TIME MARCHING METHODS FOR MONGE-AMPÈRE TYPE EQUATIONS

GERARD AWANOU

ABSTRACT. We discuss the performance of three numerical methods for the fully nonlinear Monge-Ampère equation. The first two are pseudo time continuation methods while the third is a pure pseudo time marching algorithm. The pseudo time continuation methods are shown to converge for smooth data on a uniformly convex domain. We give numerical evidence that they perform well for the non-degenerate Monge-Ampère equation. The pseudo time marching method applies in principle to any nonlinear equation. Numerical results with this approach for the degenerate Monge-Ampère equation are given as well as for the Pucci and Gauss-curvature equations.

1. INTRODUCTION

We are interested in numerical solutions of equations of type

$$(1.1) \quad F(x, u(x), Du(x), D^2u(x)) = 0,$$

on a convex bounded domain Ω of \mathbb{R}^n with boundary $\partial\Omega$ and Dirichlet boundary conditions $u = g$ and F real valued. Here u is a real valued function and $Du(x), D^2u(x)$ denote its gradient vector and Hessian matrix respectively.

The fully nonlinear Monge-Ampère equation is given by

$$(1.2) \quad \det D^2u = f \text{ in } \Omega, \quad u = g \text{ on } \partial\Omega,$$

where f, g are given functions with $f > 0$ in the non degenerate case, otherwise we assume that $f \geq 0$. We will also assume that $f \in C(\Omega)$ and $g \in C(\partial\Omega)$.

For a symmetric matrix M with eigenvalues e_1, \dots, e_n and $0 < \lambda \leq \Lambda$, we recall the Pucci extremal operators [12]

$$\mathcal{M}^+[M] = \Lambda \sum_{e_i > 0} e_i + \lambda \sum_{e_i < 0} e_i, \quad \mathcal{M}^-[M] = \lambda \sum_{e_i > 0} e_i + \Lambda \sum_{e_i < 0} e_i.$$

One can then consider the Pucci equations

$$\mathcal{M}^+[D^2u(x)] = f(x), \quad \mathcal{M}^-[D^2u(x)] = f(x).$$

The Pucci equations appear in stochastic control where the control variable is the diffusion coefficient. They also play an important role in the theory of fully nonlinear equations. If u is a solution of $F(x, D^2u(x)) = 0$ with F uniformly elliptic [12], then u is a subsolution and a supersolution of equations which do not depend on F but on λ, Λ and f . Any result valid for these classes of equations is also valid for any fully nonlinear uniformly elliptic equation, [12] p. 16. Let us consider the Pucci equation

$$\mathcal{M}^+[D^2u(x)] = f(x),$$

and put $\alpha = \Lambda/\lambda \geq 1$ and denote by λ^+, λ^- the maximum and minimum eigenvalues of D^2u respectively. In two dimensions, following [15, 14, 11, 27, 21], we will consider the Pucci equation,

$$(1.3) \quad \alpha\lambda^+ + \lambda^- = f(x), \quad \alpha \geq 1.$$

We note that in principle the vanishing moment methodology [18] is applicable to the Pucci equation. However, the operator F is not differentiable in this case as opposed to the Monge-Ampère equation for example, hence Newton's method cannot be used. One can use that methodology following [15, 14, 11], where a Pucci equation is written in terms of the Monge-Ampère operator.

Another equation of interest is the prescribed Gauss curvature equation,

$$(1.4) \quad \det D^2u - K(x)(1 + |Du|^2)^{\frac{n+2}{2}} = 0.$$

Given Ω, K and g , one seeks a convex function solution of the equation.

Starting with [9, 14], interest has grown for finite element methods which are able to capture viscosity solutions of second order fully nonlinear equations. In the context of non-smooth solutions, for proven convergence results, wide stencils finite difference have been used for the Monge-Ampère and the Pucci equations [27]. See also [28]. This paper is the sequel to [5] where we did a comparative study of three methods suitable for finite dimensional computations of solutions of the Monge-Ampère equation. We refer to the above paper for recent references and an introduction to the spline element method which is also used here to illustrate the numerical results.

In this paper, we consider three numerical methods for the Monge-Ampère equation. The first two are pseudo-time continuation methods. Given a fully nonlinear elliptic equation, $F(u) = 0$ with F differentiable, we consider the sequence of problems

$$(\nu L + F'(u_k))(u_{k+1} - u_k) = -F(u_k),$$

where L is a linear operator which we take in this paper as the identity or the Laplace operator. In the latter case, L may be viewed as a preconditioner and in this case, the computations are speeded up compared to the case where L is the identity operator. In the case of the Monge-Ampère equation, $F(u) = \det D^2u - f$, and the solution of the problem is reduced to solving a sequence of elliptic equations whose solutions are sought here in the spline space of degree d and smoothness $r \geq 1$, Section 2. Note that for the Monge-Ampère equation [5]

$$F'(u_k)(u_{k+1} - u_k) = \operatorname{div}((\operatorname{cof} D^2u_k)(Du_{k+1} - Du_k)) = (\operatorname{cof} D^2u_k) : (D^2u_{k+1} - D^2u_k).$$

Given an initial guess u_0 , we are led to the sequence of approximate solutions

$$(1.5) \quad \nu L\theta_k + (\operatorname{cof} D^2u_k) : D^2\theta_k = (f - f_k), \quad f_k = \det D^2u_k, \quad \theta_k = u_{k+1} - u_k.$$

As initial guess one may take the solution of the Poisson equation $\Delta u = 2\sqrt{f}$ in $\Omega, u = g$ on $\partial\Omega$.

Method 1 corresponds to the case where L is the identity operator and Method 2 to the case where L is the Laplace operator. The third method is a pseudo time marching method. Given $\nu > 0$, we consider the sequence of iterates

$$(1.6) \quad -\nu\Delta u_{k+1} = -\nu\Delta u_k + F(x, u_k(x), Du_k(x), D^2u_k), \quad u_{k+1} = g \text{ on } \partial\Omega.$$

It can be interpreted as an Euler discretization of the pseudo-time dependent equation $\frac{\partial \Delta u}{\partial t} + F(x, u(x), Du(x), D^2u) = 0$, or as a Laplacian preconditioner of a simple pseudo time marching algorithm, [21] $u_{k+1} = u_k + \Delta^{-1}(F(x, u_k(x), Du_k(x), D^2u_k(x)))$. See also a remark in [26]. The simple pseudo-time marching algorithm also performs well for numerical solutions in some cases for ν sufficiently large. However the use of the Laplacian preconditioner besides the speed of computation, also helps select a convex solution for the two dimensional Monge-Ampère equation. Indeed, we will see that Method 3 and to some extent Method 2 enforces $\Delta u > 0$, which when combined with $\det D^2u = f \geq 0$ and C^1 continuity implies numerical convexity [24], Lemma 1. The use of the spline element method is also motivated by its higher order of accuracy and its robustness in some limiting situations [4, 5]. We give a convergence result for the Monge-Ampère equation which says that if νI , with I the identity matrix is a good approximation of the Hessian D^2u , the discrete versions of (1.6) will converge for the mesh size h sufficiently small.

For two dimensional problem, with Method 3, convergence deteriorates with the spline element method, if accuracy of more than 10^{-9} is sought. However this is not the case with the standard finite difference discretization of (1.6). The latter performs well in three dimensions for non smooth solutions. We give these numerical results as well. They indicate that compatible discretizations are needed for (1.6), for example the standard finite difference method satisfies a discrete maximum principle.

It is believed that some methods are able to capture the viscosity solutions and other can't. This paper gives evidence that at least in two dimensions, by relaxation, a method which works only for solutions in $H^2(\Omega)$ can be expected to perform in the non smooth case for a class of problems which include the Monge-Ampère, Pucci, Gauss curvature and a class of Isaacs-Bellman equation.

We use the standard notation for the Sobolev spaces $W^{m,p}(\Omega)$ with norms and semi-norms respectively $||\cdot||_{m,p}$ and $|\cdot|_{m,p}$. In the case $p = 2$, we use the notation $H^m(\Omega)$ and the norms and semi-norms are denoted respectively $||\cdot||_m$ and $|\cdot|_m$. We will use C for a generic constant but will index specific constants.

The paper is organized as follows: we first give a brief description of the spline element method, then we give the convergence results for the iterative methods. Next, we give numerical results for the Monge-Ampère, Pucci and Gauss curvature equation and conclude with some remarks.

2. SPLINE ELEMENT DISCRETIZATION

We refer to [2, 6, 7, 8, 22, 4] for a description of the spline element method. In the simplest case, it can be described as follows. Let $u \in V = H_0^m(\Omega)$, $m \geq 1$ solve a variational problem $a(u, v) = f(v)$ with the conditions of the Lax-Milgram lemma satisfied. Take

$$V_h := S_d^r(\mathcal{T}) = \{s \in C^r(\Omega), s|_t \in \mathcal{P}_d, \forall t \in \mathcal{T}\},$$

for a triangulation \mathcal{T} of the domain and \mathcal{P}_d the space of polynomials of degree less than or equal to d . The space $S_d^r(\mathcal{T})$ is the spline space of smoothness r and degree d . For $r = 0$ and $d = 1$ we have the space of piecewise linear continuous functions. Next choose a representation of polynomials such that $V_h = \{c \in \mathbb{R}^N, Rc = G\}$ for some

integer N , matrix R and vector G . The discrete problem is find $c \in V_h, c^T K d = F^T d$ for all $d \in V_h$ for a suitable stiffness matrix K and a load vector F . Introducing a Lagrange multiplier, we are lead to saddle point problems

$$\begin{pmatrix} K & R^T \\ R & 0 \end{pmatrix} \begin{pmatrix} \mathbf{c} \\ \lambda \end{pmatrix} = \begin{bmatrix} F \\ G \end{bmatrix},$$

which are solved by a version of the augmented Lagrangian algorithm

$$(2.1) \quad \left(K + \frac{1}{\mu} R^T R\right) c^{(l+1)} = K^T c^{(l)} + \frac{1}{\mu} R^T G, \quad l = 1, 2, \dots$$

The convergence properties of the iterative method were given in [3].

3. CONVERGENCE OF THE ITERATIVE METHODS

We first describe the convergence properties of the pseudo-time continuation methods. We consider a damped version of (1.5), namely

$$(3.1) \quad \begin{aligned} (\nu L + F'(u_k))(u_{k+1} - u_k) &= -\frac{1}{\tau} F(u_k), \\ \nu L \theta_k + (\text{cof } D^2 u_k) : D^2 \theta_k &= \frac{1}{\tau} (f - f_k), \quad f_k = \det D^2 u_k, \quad \theta_k = u_{k+1} - u_k, \end{aligned}$$

where $\tau > 0$ is a damping parameter. In the numerical experiments, we used $\tau = 1$. We will need the following global regularity result, [29].

Theorem 3.1. *Let Ω be a uniformly convex domain in R^n , with boundary in C^3 . Suppose $g \in C^3(\bar{\Omega})$, $\inf f > 0$, and $f \in C^\alpha$ for some $\alpha \in (0, 1)$. Then (1.2) has a convex solution u which satisfies the a priori estimate*

$$\|u\|_{C^{2,\alpha}(\bar{\Omega})} \leq C,$$

where C depends only on $n, \alpha, \inf f, \Omega, \|f\|_{C^\alpha(\bar{\Omega})}$ and $\|g\|_{C^3}$.

According to [29], all assumptions in the above theorem are sharp. We have the following analogue of Theorem 2.1 in [26].

Theorem 3.2. *Let Ω be a uniformly convex domain in R^n , with boundary in C^3 . Let $0 < m \leq f \leq M, f \in C^\alpha$ for some $m, M > 0$ and $\alpha \in (0, 1)$. Assume also that $g \in C^3(\bar{\Omega})$. Then, there exists $\tau \geq 1$ depending on m, f , such that if u_k is the sequence defined by (1.5), it converges in $C^{2,\beta}$ to a solution u of (1.2) for u_0 sufficiently close to u and for every $\beta < \alpha$.*

Proof. We outlined the proof in [5]. See also [26]. An apparently different proof was given in [19] for L the identity operator. One shows by induction that there are constants $C_1, C_2 > 0$ such that

$$(3.2) \quad \frac{1}{C_1} \det D^2 u \leq \det D^2 u_k \leq C_1 \det D^2 u \quad \text{and} \quad \|\det D^2 u - \det D^2 u_k\|_{C^\alpha} \leq C_2,$$

for τ sufficiently large and ν in an appropriate range. This implies that the sequence $f_k = \det D^2 u_k$ is bounded in C^α and by Theorem (3.1), the sequence u_k is bounded in $C^{2,\alpha}$. Arzela-Ascoli's theorem is then used to prove that the sequence is precompact in $C^{2,\beta}, \beta < \alpha$. Since (1.2) has at most two solutions, [13] p. 324, by requiring u_0

sufficiently close to u , we assure that the solution is locally unique. The remaining part of the proof consists in verifying the induction hypothesis (3.2).

First choose f_0 (or equivalently u_0) and C_1, C_2 such that (3.2) is satisfied. And assume that it holds for f_k . We show that it holds for f_{k+1} . We have

$$f_{k+1} = \det D^2 u_{k+1} = \det(D^2 u_k + D^2 \theta_k).$$

Since $F'[u]$ is linear in u , $F''[u]$ does not depend on u and we have

$$f_{k+1} = f_k + (\text{Cof } D^2 u_k) : D^2 \theta_k + \frac{1}{2} (\text{Cof } D^2 \theta_k) : D^2 \theta_k = f_k + \frac{1}{\tau} (f - f_k) - \nu L \theta_k + r_k.$$

Put $r_k = \frac{1}{2} (\text{Cof } D^2 \theta_k) : D^2 \theta_k$. By Theorem 3.1, the equation

$$\det D^2 u_k = \rho_k, \text{ in } \Omega, \quad u_k = g \text{ on } \partial\Omega,$$

has a strictly convex solution $u_k \in C^{2,\alpha}(\Omega)$ (since $\rho_k > 0$). Recall that L is either the identity operator or the Laplace operator. This implies that $D^2 u_k$ is strictly positive definite and hence $\nu L \theta_k + (\text{cof } D^2 u_k)$ is a uniformly elliptic matrix in C^α . Moreover since θ_k solves (3.1), by Schauder theory

$$(3.3) \quad \|\theta_k\|_{C^{2,\alpha}} \leq C \frac{\|f - f_k\|_{C^\alpha}}{\tau},$$

which implies $\|D^2 \theta_k\|_{C^\alpha} \leq C \frac{\|f - f_k\|_{C^\alpha}}{\tau}$. Now, each entry in $(\text{Cof } D^2 \theta_k)$ involves the product of $n - 1$ second derivatives of θ_k . We therefore have $\|r_k\|_{C^\alpha} \leq C \frac{\|f - f_k\|_{C^\alpha}^n}{\tau^n}$. Since

$$(3.4) \quad f - f_{k+1} = (1 - 1/\tau)(f - f_k) + \nu L \theta_k - r_k,$$

by assumption (3.2) and (3.3),

$$\|f - f_{k+1}\|_{C^\alpha} \leq \left(1 - \frac{1}{\tau} + \frac{\nu C_3}{\tau} + \frac{C_4 C_2^n}{\tau^n}\right) \|f - f_k\|_{C^\alpha}.$$

For $\nu < 1/C_3$, $1 - C_3 \nu > 0$ and with τ sufficiently large, $C_4 C_2^n / \tau^{n-1} \leq 1 - C_3 \nu$, and we have $\|f - f_{k+1}\|_{C^\alpha} \leq C_2$. It remains to verify that $\frac{1}{C_1} f \leq f_{k+1} \leq C_1 f$. The induction assumption for f_k implies that $C_1 \geq 1$. If $C_1 = 1$, $f_k = f$ and from (3.1) $f_{k+1} = f$. We can therefore assume that $C_1 > 1$. Using (3.4), (3.3) and (3.2), we obtain $\|\theta_k\|_{0,\infty} \leq C_6/\tau$, $\|r_k\|_{0,\infty} \leq C_5/\tau^n$ and we have

$$f - f_{k+1} \leq \left(1 - \frac{1}{\tau}\right) \left(1 - \frac{1}{C_1}\right) f + \frac{\nu C_6}{\tau} + \frac{C_5}{\tau^n}.$$

Note that $\nu C_6 + C_5/\tau^{n-1} \leq (1 - 1/C_1)f$ implies $f - f_{k+1} \leq (1 - 1/C_1)f$. Since $f > m > 0$ by assumption, for $\nu < (1 - 1/C_1)m/C_6$ and τ sufficiently large, the result holds. We obtain similarly $f_{k+1} - f \leq (C_1 - 1)f$ for τ sufficiently large and ν in the appropriate range. \square

Next, we give convergence results for Method 3.

Theorem 3.3. *Let $u_0 \in C^0(\bar{\Omega}) \cap C^{2,\alpha}(\Omega)$ such that $\Delta u_0 > 0$ and assume that*

$$F(x, u(x), Du(x), D^2 u(x)) - f(x) \in C^\alpha, \text{ for } u \in C^{2,\alpha}.$$

Assume moreover that $F(x, u_0(x), Du_0(x), D^2u_0(x)) \leq f(x)$. Then, there is an increasing sequence ν_{k+1} such that the sequence defined by

$$(3.5) \quad -\nu_{k+1}\Delta u_{k+1} = -\nu_{k+1}\Delta u_k + F(x, u_k(x), Du_k(x), D^2u_k) - f(x), \quad u_{k+1} = g \text{ on } \partial\Omega.$$

is monotonically decreasing.

Proof. We have $-\nu_1\Delta u_1 = -\nu_1\Delta u_0 + F(x, u_0(x), Du_0(x), D^2u_0(x)) - f$, so $-\nu_1\Delta(u_1 - u_0) \leq 0$. Since the domain Ω is convex, it satisfies the exterior sphere condition [23]. Moreover, $g \in C(\partial\Omega)$ by assumption. By Theorem 6.13 in [20], $u_1 \in C^0(\bar{\Omega}) \cap C^2(\Omega)$. By the maximum principle, since $u_1 - u_0 \in C^0(\bar{\Omega}) \cap C^{2,\alpha}(\Omega)$,

$$u_1 \leq u_0.$$

We next show that for sufficiently large ν_{k+1} , $u_{k+1} \leq u_k$. As $\Delta u_0 > 0$ and $-\nu_1(\Delta u_1 - \Delta u_0) \leq 0$, we have $\Delta u_1 \geq \Delta u_0 > 0$. Assume by induction that ν_k has been chosen such that $\Delta u_k > 0$. Note that (3.12) can be written

$$(3.6) \quad \begin{aligned} \Delta u_{k+1} &= \Delta u_k - \frac{1}{\nu_{k+1}} (F(x, u_k(x), Du_k(x), D^2u_k) - f(x)), \text{ in } \Omega, \\ u_{k+1} &= g \text{ on } \partial\Omega. \end{aligned}$$

If $F(x, u_k(x), Du_k(x), D^2u_k) - f(x) \leq 0$ for all $x \in \Omega$, then $\Delta u_{k+1} \geq \Delta u_k$ and $u_{k+1} \leq u_k$. In general, let $\Gamma = \{x \in \Omega, F(x, u_k(x), Du_k(x), D^2u_k) - f(x) > 0\}$. Note that the right hand side of the first equation in (3.6) is of the type $A - 1/\nu B$ where A and B are both positive continuous functions hence bounded on $\bar{\Gamma}$, that is there are numbers m_A, M_A, m_B and M_B such that $m_A \leq A(x) \leq M_A$ and $m_B \leq B(x) \leq M_B$ for all x in Ω . Since $\Delta u_k \geq \Delta u_{k-1} \geq \dots \geq \Delta u_0 > f$, we have $m_A > 0$.

Now, for $\nu \geq \bar{\nu}_{k+1} = M_B/m_A$, $A - 1/\nu B \geq 0$ and we conclude that for $\nu_{k+1} = \max(\nu_k, \bar{\nu}_{k+1})$, $\Delta u_{k+1} \geq \Delta u_k$. This concludes the proof. \square

It is not clear whether the sequences ν_k and u_k are bounded from above and below respectively. The assumption $F(x, u_0(x), Du_0(x), D^2u_0(x)) \leq f(x)$ can be relaxed by choosing ν sufficiently large so that $\Delta u_1 \geq \Delta u_0$.

We can give a more precise result in the discrete case.

Recall that [5],

$$(3.7) \quad \det D^2u = \frac{1}{n} (\text{cof } D^2u) : D^2u = \frac{1}{n} \text{div} ((\text{cof } D^2u) Du).$$

This gives the following weak formulation, [5], of the Dirichlet problem for the Monge-Ampère equation: find $u \in H^n(\Omega)$, $u = g$ on $\partial\Omega$ such that

$$(3.8) \quad \int_{\Omega} (\text{cof } D^2u) Du \cdot Dw \, dx = -n \int_{\Omega} fw \, dx, \quad \forall w \in H^n(\Omega) \cap H_0^1(\Omega).$$

Or equivalently $\int_{\Omega} \det D^2u w \, dx = \int_{\Omega} fw \, dx$, $\forall w \in H^n(\Omega) \cap H_0^1(\Omega)$. We have

Lemma 3.4. *Assume $n = 2$ and Ω bounded. Let $v, w \in H^2(\Omega)$ and $\psi \in H_0^1(\Omega) \cap H^2(\Omega)$, then*

$$\int_{\Omega} (\det D^2v - \det D^2w)\psi \, dx = \int_0^1 \left\{ \int_{\Omega} ((\operatorname{cof}(1-t)D^2w + tD^2v)(Dv - Dw))D\psi \, dx \right\} dt,$$

and

$$\left| \int_{\Omega} (\det D^2v - \det D^2w)\psi \, dx \right| \leq C_0 \|v + w\|_2 \|v - w\|_2 \|\psi\|_2.$$

Moreover, if $v, w \in H^2(\Omega) \cap W^{2,\infty}(\Omega)$

$$\left| \int_{\Omega} (\det D^2v - \det D^2w)\psi \, dx \right| \leq C_0 \|v + w\|_{2,\infty} \|v - w\|_1 \|\psi\|_1.$$

Proof. We first recall the Mean Value Theorem. Let E and F be Banach spaces and X an open subset of E and let $F : X \rightarrow F$ be a differentiable map. If $F' : X \rightarrow L(E, F)$ is continuous, F is said to be of class C^1 and for all $a, x \in X$, we have

$$F(x) = F(a) + \int_0^1 DF[(1-t)a + tx](x-a) \, dt.$$

Next, let $F : C^\infty(\Omega) \rightarrow C^\infty(\Omega)$ denote the mapping $v \mapsto \det D^2v$. Then F is differentiable with

$$F'[u](v) = (\operatorname{cof} D^2u) : D^2v = \operatorname{div}((\operatorname{cof} D^2u)Dv).$$

Since $v \mapsto F'[v]$ is linear, F is of class C^1 and by the Mean Value Theorem

$$F(v) - F(w) = \int_0^1 \operatorname{div}((\operatorname{cof}(1-t)D^2w + tD^2v)(Dv - Dw)) \, dt.$$

Next, let $\psi \in \mathcal{D}(\Omega)$. By Fubini's theorem,

$$\begin{aligned} \int_{\Omega} (\det D^2v - \det D^2w)\psi \, dx &= \int_0^1 \left\{ \int_{\Omega} \operatorname{div}((\operatorname{cof}(1-t)D^2w + tD^2v)(Dv - Dw))\psi \, dx \right\} dt \\ &= \int_0^1 \left\{ \int_{\Omega} ((\operatorname{cof}(1-t)D^2w + tD^2v)(Dv - Dw))D\psi \, dx \right\} dt. \end{aligned}$$

Since for u, ψ smooth, $\psi = 0$ on $\partial\Omega$,

$$\begin{aligned} \left| \int_{\Omega} \det D^2u\psi \right| &= \left| \int_{\Omega} (\operatorname{cof} D^2u)Du \cdot D\psi \right| \leq C \|D^2u\|_0 \|Du\|_{0,4} \|D\psi\|_{0,4} \\ &\leq C \|u\|_2^2 \|\psi\|_2, \end{aligned}$$

by Hölder inequality and Sobolev embedding, we conclude by the density of $\mathcal{D}(\Omega)$ in $H_0^1(\Omega)$ that the above equality also holds for $v, w \in C^\infty(\Omega)$ and $\psi \in H_0^1(\Omega) \cap H^2(\Omega)$. Then by the density of $\mathcal{D}(\Omega)$ in the Sobolev spaces, it is also valid for $v, w \in H^2(\Omega)$. The result then follows. \square

Let V_h be the spline space of degree d and smoothness $r \geq 1$

$$S_d^r(\mathcal{T}) = \{p \in C^r(\Omega), p|_t \in P_d, \forall t \in \mathcal{T}\},$$

where P_d denotes the space of polynomials of degree d in two variables and \mathcal{T} denotes the triangulation of the domain Ω . We make the assumption that the triangulation is quasi-uniform in the sense that there is a constant $C > 0$ such that any triangle K , $h_K/\rho_K \leq C$, where h_K denotes the diameter of K and ρ_K the radius of the largest ball contained in K .

For $r = 0, 1$ and $d \geq 5$ since the Argyris finite element space is contained in $S_d^r(\mathcal{T})$, we will use error estimates of the Argyris interpolation operator Q . For $f \in W^{m+1,q}(\Omega)$,

$$(3.9) \quad \|f - Qf\|_{k,q} \leq Ch^{m+1-k}|f|_{m+1,q},$$

for $0 \leq k \leq \min\{m+1, 2\}$, $1 \leq q \leq \infty$. Put $V_0^h = V^h \cap H_0^1(\Omega)$ and note from (3.9),

$$(3.10) \quad \|Qu\|_{2,q} \leq C\|u\|_{2,q}, \quad u \in W^{2,q}(\Omega),$$

assuming $h \leq 1$. We also recall the following inverse inequality which may be viewed as a consequence of the assumption of uniform triangulation and of Markov inequality, [25] p. 2,

$$(3.11) \quad \|p\|_2 \leq \frac{C}{h}\|p\|_1, \forall p \in \mathcal{P}_d, d \geq 1.$$

Using a standard fixed point argument, we have

Theorem 3.5. *Assume $u \in H^{m+1}(\Omega)$, $m > 2$, and let A be a symmetric uniformly positive definite matrix which approximates D^2u in the sense that $\|\text{cof}(D^2Qu) - A\|_{0,\infty} \leq Ch^p$, $p > 1$. Then the spline element approximation in $S_d^r(\mathcal{T})$ of the element of the sequence*

$$(3.12) \quad -\text{div } A\nabla u_{k+1} = -\text{div } A\nabla u_k + \det D^2u_k - f, \quad u_{k+1} = g \text{ on } \partial\Omega.$$

converges for h sufficiently small to the solution of the discrete version of (3.8), namely: Find $u_h \in S_d^r(\mathcal{T})$, $d \geq 5$, $u_h = g$ on $\partial\Omega$ such that

$$(3.13) \quad -\frac{1}{2} \int_{\Omega} (\text{cof } D^2u_h) Du_h \cdot Dw_h \, dx = \int_{\Omega} fw_h \, dx, \quad \forall w_h \in V_0^h,$$

or equivalently $\int_{\Omega} \det D^2u_h w_h \, dx = \int_{\Omega} fw_h \, dx$, $\forall w_h \in V_0^h$. Moreover, we have the error estimate

$$(3.14) \quad \|u - u_h\|_1 \leq Ch^m|u|_{m+1}, \quad \|u - u_h\|_2 \leq Ch^{m-1}|u|_{m+1}.$$

Proof. Let $B[v, w]$ define a bilinear form on $H_0^1(\Omega) \times H_0^1(\Omega)$ by

$$(3.15) \quad B[v, w] = \int_{\Omega} ADv \cdot Dw,$$

and for a given $v_h \in V^h$, $v_h = g$ on $\partial\Omega$, define $T(v_h)$ as the unique solution of

$$(3.16) \quad B[v_h - T(v_h), \psi_h] = - \int_{\Omega} \det D^2v_h \psi_h \, dx + \int_{\Omega} f\psi_h \, dx, \quad \forall \psi_h \in V_0^h.$$

Since $v_h - T(v_h) \in V_0^h$, $T(v_h) \in V^h$, $T(v_h) = g$ on $\partial\Omega$. A fixed point of the nonlinear operator T corresponds to a solution of (3.13) and conversely if v_h is a solution of

(3.13), then v_h is a fixed point of T . We will show that T has a unique fixed point in a neighborhood of $Q(u)$. Put

$$B_h(\rho) = \{v_h \in V_h, v_h = g \text{ on } \partial\Omega, \|v_h - Qu\|_1 \leq \rho\}.$$

We first show that

$$(3.17) \quad \|Qu - T(Qu)\|_1 \leq C_1 h^m \|u\|_{2,\infty} |u|_{m+1},$$

then we show there exists $0 < \rho_0$ depending on h such that T is a contraction mapping in the ball $B_h(\rho_0)$ with a contraction factor $1/2$. We conclude by applying the Brouwer fixed point theorem in a suitable ball.

Put $w_h = Qu - T(Qu)$. Then $w_h \in H_0^1(\Omega)$ and using $\det D^2 u = f$,

$$B[w_h, w_h] = \int_{\Omega} (\det D^2 u - \det D^2 Qu) w_h dx.$$

By Sobolev embedding, since $n = 2$, for $m > 2$, elements of $H^{m+1}(\Omega)$ can be considered as elements in $W^{2,\infty}(\Omega)$. Then by Lemma 3.4, the coercivity of B on $H_0^1(\Omega)$, (3.10) and (3.9),

$$\|w_h\|_1^2 \leq C \|u\|_{2,\infty} \|u - Qu\|_1 \|w_h\|_1 \leq C_1 h^m \|u\|_{2,\infty} |u|_{m+1} \|w_h\|_1,$$

from which the claim follows.

For $v_h, w_h \in B_h(\rho_0)$, with ρ_0 yet to be determined, and $\psi_h \in V_0^h$,

$$\begin{aligned} B[T(v_h) - T(w_h), \psi_h] &= B[T(v_h) - v_h, \psi_h] + B[v_h - w_h, \psi_h] + B[w_h - T(w_h), \psi_h] \\ &= \int_{\Omega} (\det D^2 v_h - \det D^2 w_h) \psi_h dx + \int_{\Omega} A(Dv_h - Dw_h) D\psi_h dx. \end{aligned}$$

Using

$$\int_{\Omega} A(Dv_h - Dw_h) D\psi_h dx = \int_0^1 \int_{\Omega} A(Dv_h - Dw_h) D\psi_h dx dt,$$

Lemma 3.4, and with the observation that by the inverse inequality, $v_h, w_h \in H^2(\Omega)$, we have

$$\begin{aligned} B[T(v_h) - T(w_h), \psi_h] &= \int_0^1 \left\{ \int_{\Omega} (A + \text{cof}((1-t)D^2 w_h + tD^2 v_h))(Dv_h - Dw_h) D\psi_h dx \right\} dt \\ &= \int_0^1 \left\{ \int_{\Omega} (A + \text{cof} D^2 Qu)(Dv_h - Dw_h) D\psi_h dx \right\} dt \\ &\quad + \int_0^1 \left\{ \int_{\Omega} (\text{cof}((1-t)(D^2 w_h - D^2 Qu) + t(D^2 v_h - D^2 Qu)) \right. \\ &\quad \left. (Dv_h - Dw_h) D\psi_h dx \right\} dt. \end{aligned}$$

We conclude using again the coercivity of B on $H_0^1(\Omega)$, Lemma 3.4, and $\psi_h = T(v_h) - T(w_h)$ that

$$\begin{aligned} \|T(v_h) - T(w_h)\|_1^2 &\leq C \|A + \text{cof} D^2 Qu\|_{0,\infty} \|v_h - w_h\|_1 \|\psi_h\|_1 \\ &\quad + C (\|D^2 w_h - D^2 Qu\|_0 + \|D^2 v_h - D^2 Qu\|_2) \|v_h - w_h\|_2 \|\psi_h\|_2. \end{aligned}$$

By the inverse inequality (3.11), and the assumption in the theorem

$$\|T(v_h) - T(w_h)\|_1 \leq (C_2 h^p + C_3 \frac{\rho_0}{h^3}) \|v_h - w_h\|_1.$$

We require $C_2 h^p \leq 1/4$ and we choose ρ_0 such that for, $C_3 \rho_0 / h^3 \leq 1/4$ for example $\rho_0 = h^3 / (4C_3)$. It follows that T is a contraction mapping in the ball $B_h(\rho_0)$ with a contraction factor $1/2$.

Finally, note that with $\rho_0 = h^3 / (4C_3)$, in $B_h(\rho_0)$,

$$\|Tv_h - Qu\|_1 \leq \|Qu - T(Qu)\|_1 + \|TQu - Tv_h\|_1 \leq C_1 h^m \|u\|_{2,\infty} |u|_{m+1} + \frac{\|v_h - Qu\|_1}{2}.$$

Put $\rho_1 = 2C_1 h^m \|u\|_{2,\infty} |u|_{m+1}$. For $m > 1$ and h sufficiently small, $\rho_1 \leq \rho_0$, and T maps $B_h(\rho_1)$ into itself.

We conclude by the Brouwer fixed point theorem that T has a unique fixed point u_h in $B_h(\rho_1)$ to which the iterates $u_h^{k+1} = T(u_h^k)$ converge. Moreover

$$\begin{aligned} \|u - u_h\|_1 &\leq \|u - Qu\|_1 + \|Qu - u_h\|_1 = \|u - Qu\|_1 + \|Qu - T(u_h)\|_1 \\ &\leq Ch^m |u|_{m+1} + \rho_1 \leq (C + 2C_1 \|u\|_{2,\infty}) h^m |u|_{m+1}. \end{aligned}$$

and

$$\begin{aligned} \|u - u_h\|_2 &\leq \|u - Qu\|_2 + \|Qu - u_h\|_2 = \|u - Qu\|_2 + \|Qu - Tu_h\|_2 \\ &\leq Ch^{m-1} |u|_{m+1} + \frac{\rho_1}{h} \leq Ch^{m-1} |u|_{m+1}, \end{aligned}$$

using again an inverse estimate. \square

Using a duality argument, we have the following L^2 error estimate, for $m \geq 2$, and assuming $\|A - \text{cof}D^2u\|_{0,\infty} \leq Ch^p, p \geq 1$

$$\|u - u_h\|_0 \leq Ch^{m+1} |u|_{m+1}.$$

Assume that the domain is convex and let $w \in H^2(\Omega)$ be the solution of the problem

$$\text{div}(ADw) = u - u_h, \text{ in } \Omega, w = 0 \text{ on } \Omega.$$

By elliptic regularity, $\|w\|_2 \leq \|u - u_h\|_0$. We have

$$\begin{aligned} (3.18) \quad \|u - u_h\|_0^2 &= \int_{\Omega} (u - u_h) \text{div}(ADw) dx = \int_{\Omega} (ADw) \cdot D(u - u_h) dx \\ &= \int_{\Omega} A(Dw - DQw) \cdot D(u - u_h) dx + \int_{\Omega} (ADQw) \cdot D(u - u_h) dx. \end{aligned}$$

Recall that for $v_h \in V_0^h$

$$\begin{aligned} \int_{\Omega} \det D^2 u v_h dx &= \int_{\Omega} (\text{cof}D^2 u) Du \cdot Dv_h dx = \int_{\Omega} f v_h dx, \\ \int_{\Omega} \det D^2 u_h v_h dx &= \int_{\Omega} (\text{cof}D^2 u_h) Du_h \cdot Dv_h dx = \int_{\Omega} f v_h dx. \end{aligned}$$

And so by Lemma 3.4

$$\begin{aligned} \int_{\Omega} \det D^2 u v_h dx - \int_{\Omega} \det D^2 u_h v_h dx &= \int_0^1 \left\{ \int_{\Omega} ((\operatorname{cof}(1-t)D^2 u_h + tD^2 u) \right. \\ &\quad \left. (Du - Du_h)) Dv_h dx \right\} dt, \\ &= \int_0^1 \left\{ \int_{\Omega} ((\operatorname{cof}(1-t)(D^2 u_h - D^2 u) + D^2 u) \right. \\ &\quad \left. (Du - Du_h)) Dv_h dx \right\} dt. \end{aligned}$$

Subtracting from (3.18), using $v_h = Qw$, we obtain

$$\begin{aligned} \|u - u_h\|_0^2 &\leq C \|w - Qw\|_1 \|u - u_h\|_1 + C \|A - \operatorname{cof} D^2 u\|_{0,\infty} \|w\|_1 \|u - u_h\|_1 \\ &\quad + C \|u - u_h\|_2^2 \|Qw\|_2. \end{aligned}$$

Since $\|w - Qw\|_1 \leq Ch \|w\|_2$, $\|w\|_1 \leq \|w\|_2$, $\|Qw\|_2 \leq \|w\|_2$, by elliptic regularity, and assuming $\|A - \operatorname{cof} D^2 u\|_{0,\infty} \leq Ch^p$,

$$\|u - u_h\|_0 \leq Ch^{m+1} |u|_{m+1} + Ch^{m+p} |u|_{m+1} + Ch^{2m-2} |u|_{m+1},$$

from which the result follows.

4. NUMERICAL RESULTS

Unless otherwise indicated, all numerical simulations below are for $r = 1$ and the domain is $[0, 1]^n$, $n = 2, 3$. For $n = 2$, the computational domain is the unit square $[0, 1]^2$ which is first divided into squares of side length h . Then each square is divided into two triangles by the diagonal with negative slope. For $n = 3$, the initial tetrahedral partition \mathcal{T}_1 consists in six tetrahedra. Each tetrahedron is then uniformly refined into 8 subtetrahedra forming \mathcal{T}_k , $k = 2, 3$. In the tables, n_{int} denote the number of iterations.

4.1. Monge-Ampère. We used the following test functions suggested in [14, 10, 18].

Test 1: $u(x, y) = e^{(x^2+y^2)/2}$ so that $f(x, y) = (1 + x^2 + y^2)e^{(x^2+y^2)}$ and $g(x, y) = e^{(x^2+y^2)/2}$ on $\partial\Omega$.

Test 2: $u(x, y) = -\sqrt{2 - x^2 - y^2}$ so that $f(x, y) = 2/(2 - x^2 - y^2)^2$ and $g(x, y) = -\sqrt{2 - x^2 - y^2}$ on $\partial\Omega$.

Test 3: $f(x, y) = 1$ and $g(x, y) = 0$. No exact solution is known.

Test 4: $u(x, y, z) = e^{(x^2+y^2+z^2)/3}$ so that $f(x, y, z) = 8/81(3 + 2(x^2 + y^2 + z^2))e^{(x^2+y^2+z^2)}$ and $g(x, y, z) = e^{(x^2+y^2+z^2)/3}$ on $\partial\Omega$.

Test 5: $u(x, y, z) = -\sqrt{3 - x^2 - y^2 - z^2}$ so that $f(x, y, z) = 3(3 - x^2 - y^2 - z^2)^{-5/2}$ and $g(x, y, z) = -\sqrt{3 - x^2 - y^2 - z^2}$ on $\partial\Omega$.

Test 6: $f(x, y, z) = 1$ and $g(x, y, z) = 0$. No exact solution is known.

Test 7: $u(x, y) = |x - 1/2|$ with $g(x, y) = |x - 1/2|$ and $f(x, y) = 0$.

d	n_{it}	L^2 norm	H^1 norm	H^2 norm
$d = 3$	8	$1.0610 \cdot 10^{-3}$	$1.1101 \cdot 10^{-2}$	$1.6383 \cdot 10^{-1}$
$d = 4$	10	$3.5127 \cdot 10^{-5}$	$4.8553 \cdot 10^{-4}$	$9.0596 \cdot 10^{-3}$
$d = 5$	13	$4.1572 \cdot 10^{-6}$	$6.5142 \cdot 10^{-5}$	$1.9364 \cdot 10^{-3}$
$d = 6$	25	$1.9684 \cdot 10^{-7}$	$3.6401 \cdot 10^{-6}$	$1.4774 \cdot 10^{-4}$
$d = 7$	40	$2.2712 \cdot 10^{-8}$	$4.1495 \cdot 10^{-7}$	$2.2424 \cdot 10^{-5}$
$d = 8$	157	$1.3867 \cdot 10^{-9}$	$3.1763 \cdot 10^{-8}$	$2.2274 \cdot 10^{-6}$
$d = 9$	303	$3.7149 \cdot 10^{-9}$	$1.5998 \cdot 10^{-7}$	$1.1272 \cdot 10^{-5}$

TABLE 1. Test 1, Method 1, $h = 1/2, \nu = 0.015$

d	n_{it}	L^2 norm	H^1 norm	H^2 norm
$d = 3$	10	$1.2809 \cdot 10^{-4}$	$2.6554 \cdot 10^{-3}$	$8.9587 \cdot 10^{-2}$
$d = 4$	12	$1.6278 \cdot 10^{-6}$	$4.5619 \cdot 10^{-5}$	$1.7395 \cdot 10^{-3}$
$d = 5$	16	$1.1531 \cdot 10^{-7}$	$2.3916 \cdot 10^{-6}$	$1.3444 \cdot 10^{-4}$
$d = 6$	24	$1.7609 \cdot 10^{-9}$	$6.8523 \cdot 10^{-8}$	$5.5427 \cdot 10^{-6}$
$d = 7$	39	$1.7113 \cdot 10^{-10}$	$4.9437 \cdot 10^{-9}$	$5.3920 \cdot 10^{-7}$
$d = 8$	111	$3.9851 \cdot 10^{-10}$	$3.1860 \cdot 10^{-8}$	$4.6123 \cdot 10^{-6}$

TABLE 2. Test 1, Method 1, $h = 1/4, \nu = 0.015$

h	n_{it}	L^2 norm	H^1 norm
$1/2^0$	435	2.195410^{-2}	1.640910^{-1}
$1/2^1$	352	3.609710^{-3}	6.140510^{-2}
$1/2^2$	345	1.068510^{-3}	4.097810^{-2}
$1/2^3$	319	3.766610^{-4}	2.947810^{-2}

TABLE 3. Test 2: Method 1, $d = 3, \nu = 15, u(x, y) = -\sqrt{2 - x^2 - y^2} \notin H^2(\Omega)$

d	n_{it}	L^2 norm	H^1 norm	H^2 norm
$d = 3$	1	$1.2338 \cdot 10^{-2}$	$7.6984 \cdot 10^{-2}$	$4.4411 \cdot 10^{-1}$
$d = 4$	7	$1.6289 \cdot 10^{-3}$	$1.4719 \cdot 10^{-2}$	$1.3983 \cdot 10^{-1}$
$d = 5$	10	$1.5333 \cdot 10^{-3}$	$8.7312 \cdot 10^{-3}$	$6.0412 \cdot 10^{-2}$
$d = 6$	18	$1.2324 \cdot 10^{-4}$	$9.7171 \cdot 10^{-4}$	$1.0584 \cdot 10^{-2}$

TABLE 4. Test 4, Method 1, $\mathcal{I}_1, \nu = 0.015$

d	n_{it}	L^2 norm	H^1 norm	H^2 norm
$d = 3$	1	$3.1739 \cdot 10^{-3}$	$2.3005 \cdot 10^{-2}$	$2.4496 \cdot 10^{-1}$
$d = 4$	14	$3.2786 \cdot 10^{-4}$	$3.5626 \cdot 10^{-3}$	$5.2079 \cdot 10^{-2}$
$d = 5$	39	$2.4027 \cdot 10^{-5}$	$3.9210 \cdot 10^{-4}$	$8.8868 \cdot 10^{-3}$

TABLE 5. Test 4, Method 1, $\mathcal{I}_2, \nu = 0.015$

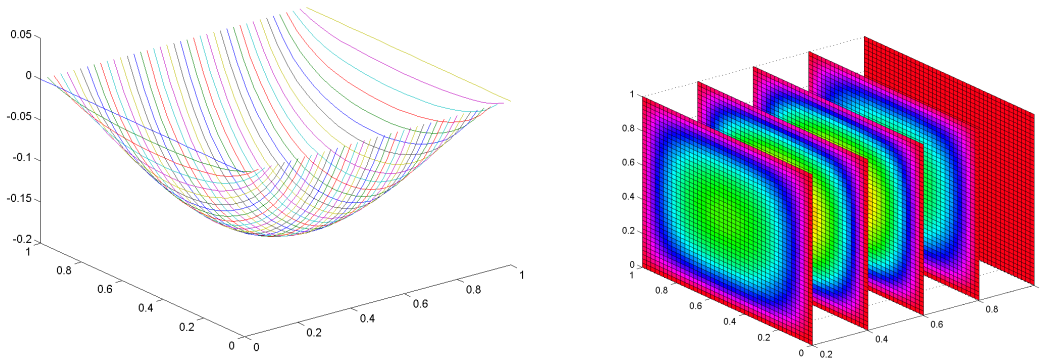


FIGURE 1. Test 6, Method 1, on \mathcal{I}_3 , $d = 5$, $\nu = 15$ graph $x = 1/2$, slices x -direction

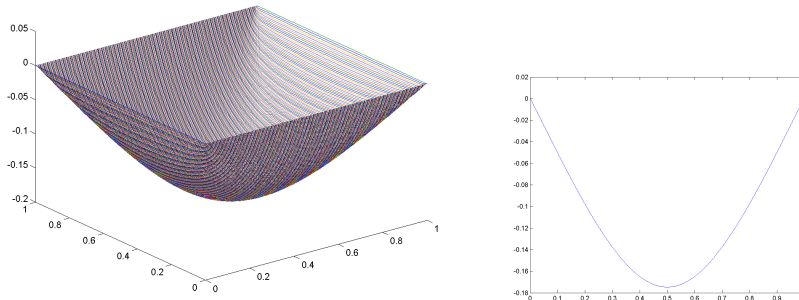
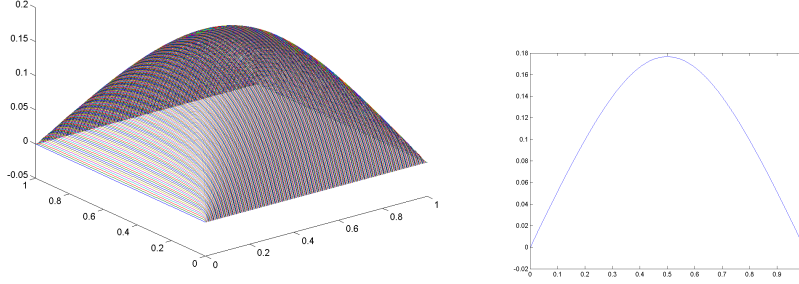


FIGURE 2. Test 3, $L = \Delta$, $h = 1/2^4$, $d = 5$, $\nu = 15$.

For non-smooth solutions, better results are obtained with L as the Laplace operator, Method 2. Even for smooth solutions, the computations are much faster. For example, for Test 6, one observes a negative curvature of the numerical solution near the corners. This is a problem with many numerical simulations of the Monge-Ampère equation, [14, 10, 18]. It is evidenced only when one plots the graph of the solution along the line $y = x$ in \mathbb{R}^2 . This problem disappears when we use $L = \Delta$ in two dimensions. This can be explained as follows: for plane problems, the scheme enforces element by element $\Delta u > 0$ and for a non degenerate problem $\det D^2 u = f > 0$. This implies that the numerical solution has Hessian positive definite element by element, which when combined with C^1 continuity gives numerical convexity [24], Lemma 1.

There are three instances where Method 3 outperforms the pseudo-time continuation methods. The concavity property of the concave solution in the case $f(x, y) = 1$, $g(x, y) = 0$ are better than the one obtained by the vanishing moment methodology, [5, 18]. The second example is its performance for Test 2. And finally, we give results for Test 7, a degenerate Monge-Ampère equation. We obtained similar results for the three dimensional analogue of Test 7.

FIGURE 3. Test 3, Method 3, $h = 1/2^4$, $d = 5$, $\nu = 50$.

h	n_{it}	L^2 norm	H^1 norm
$1/2^1$	826	$2.0721 \cdot 10^{-2}$	$1.5963 \cdot 10^{-1}$
$1/2^2$	652	$1.8579 \cdot 10^{-3}$	$5.4300 \cdot 10^{-2}$
$1/2^3$	644	$5.0438 \cdot 10^{-4}$	$3.7948 \cdot 10^{-2}$
$1/2^4$	588	$2.1132 \cdot 10^{-4}$	$2.8015 \cdot 10^{-2}$
$1/2^5$	547	$8.4871 \cdot 10^{-5}$	$2.0803 \cdot 10^{-2}$
$1/2^6$	2	$4.4597 \cdot 10^{-3}$	$9.9622 \cdot 10^{-2}$
$1/2^7$	1	NaN	NaN

TABLE 6. Test 2, Method 3, $d = 3$, $\nu = 50$

We note that for smooth solutions, if ν is not high enough, numerical results with Method 3 may be inaccurate. This may explain the loss of accuracy for $h = 1/2^6$. However, this is best evidenced by the performance of the method on a smooth solution, Tables 7, 8 and 9. In these tables, one observes a loss of accuracy for high value of d . As the number of degrees of freedom increases, it becomes expensive to find a suitable value of ν . For Test 2, with $h = 1/2^6$ and $\nu = 500$, we obtained errors in the L^2 and H^1 norms respectively $4.6611 \cdot 10^{-4}$ and $1.5931 \cdot 10^{-2}$. With $\nu = 700$, the errors were $7.9749 \cdot 10^{-4}$ and $1.6327 \cdot 10^{-2}$. We then used the following variant of Method 3: Put $G(x, D^2u(x)) = -\nu\Delta u + F(x, D^2u(x))$. For $m = 1, 2, \dots$, we consider truncating functions $\chi_m(r)$ defined by $\chi_m(r) = -m$ for $r < -m$, $\chi_m(r) = r$ for $-m \leq r \leq m$ and $\chi_m(r) = m$ for $r > m$ and the sequence of problems

$$(4.1) \quad -\nu\Delta u + \chi_m(G(x, D^2u(x))) = 0 \text{ in } \Omega, u = g \text{ on } \partial\Omega.$$

With $\nu = 150$ and $m = 25$, the results were identical to the ones previously obtained and for $h = 1/2^6$ the errors in the L^2 and H^1 norms were $3.5825 \cdot 10^{-5}$ and $1.5733 \cdot 10^{-2}$. Unfortunately, we were not able to make this trick work for high values of d or very small values for h even in the case of smooth solutions.

Note carefully that the monotone scheme in [27] has a directional error when a fixed stencil is used and has not been tested in the regime where the accuracy is of order 10^{-5} for the singular example of Test 2. Also, the condition number in (2.1) can be very large and we were not able to compute a numerical solution for $h = 1/2^7$. A

h	n_{it}	L^2 norm	H^1 norm	H^2 norm
$d = 3$	24	1.061010^{-3}	1.110110^{-2}	1.638310^{-1}
$d = 4$	24	3.512710^{-5}	4.855310^{-4}	9.059610^{-3}
$d = 5$	40	4.157210^{-6}	6.514210^{-5}	1.936410^{-3}
$d = 6$	2	6.205810^{-4}	7.523610^{-3}	2.128110^{-1}
$d = 7$	2	6.340810^{-4}	8.329910^{-3}	2.669910^{-1}
$d = 8$	2	6.398810^{-4}	8.763610^{-3}	3.060410^{-1}

TABLE 7. Test 1, Method 3, $h = 1/2, \nu = 5$

h	n_{it}	L^2 norm	H^1 norm	H^2 norm
$d = 3$	260	$1.0610 \ 10^{-3}$	$1.1101 \ 10^{-2}$	$1.6383 \ 10^{-1}$
$d = 4$	234	$3.5127 \ 10^{-5}$	$4.8553 \ 10^{-4}$	$9.0596 \ 10^{-3}$
$d = 5$	236	$4.1569 \ 10^{-6}$	$6.5142 \ 10^{-5}$	$1.9364 \ 10^{-3}$
$d = 6$	217	$1.9780 \ 10^{-7}$	$3.6411 \ 10^{-6}$	$1.4775 \ 10^{-4}$
$d = 7$	213	$2.1441 \ 10^{-8}$	$4.1381 \ 10^{-7}$	$2.2415 \ 10^{-5}$
$d = 8$	186	$1.0893 \ 10^{-8}$	$6.3270 \ 10^{-8}$	$1.6264 \ 10^{-6}$

TABLE 8. Test 1, Method 3, $h = 1/2, \nu = 50$

h	n_{it}	L^2 norm	H^1 norm	H^2 norm
$d = 3$	239	$1.2809 \ 10^{-4}$	$2.6554 \ 10^{-3}$	$8.9587 \ 10^{-2}$
$d = 4$	233	$1.6279 \ 10^{-6}$	$4.5620 \ 10^{-5}$	$1.7395 \ 10^{-3}$
$d = 5$	233	$1.1504 \ 10^{-7}$	$2.3915 \ 10^{-6}$	$1.3444 \ 10^{-4}$
$d = 6$	230	$2.1643 \ 10^{-9}$	$6.8741 \ 10^{-8}$	$5.5412 \ 10^{-6}$
$d = 7$	205	$2.9193 \ 10^{-9}$	$1.8321 \ 10^{-8}$	$4.1951 \ 10^{-7}$
$d = 8$	179	$1.5311 \ 10^{-8}$	$8.4643 \ 10^{-8}$	$6.7135 \ 10^{-7}$

TABLE 9. Test 1, Method 3, $h = 1/4, \nu = 50$

factor to take into account is that f is singular and for polynomial interpolation of f , values of f near the singularity point $(1, 1)$ have to be used which inevitably causes overflow. We set $f(1, 1) = 500$ for this test.

For Test 7, the scheme is able again to capture the singularity.

Method 3 did not converge for Test 5. Nor did Method 1 and Method 2. Although convergence at the continuous level (on smooth domains) is guaranteed. This can be explained as follows: in three dimensions, for a matrix A , $\text{tr}(A) \geq 0$ and $\det A \geq 0$ is not enough to characterize a semi-positive definite Hessian. Using the vanishing moment methodology in the framework of the spline element method, we also observe divergence for Test 5. The paper [18] did not report on that case. There could be an intricate relationship between the discretization parameter h and other parameters in the problem, including the parameter μ in (2.1). As discussed in the remarks excellent results are obtained with finite difference methods.

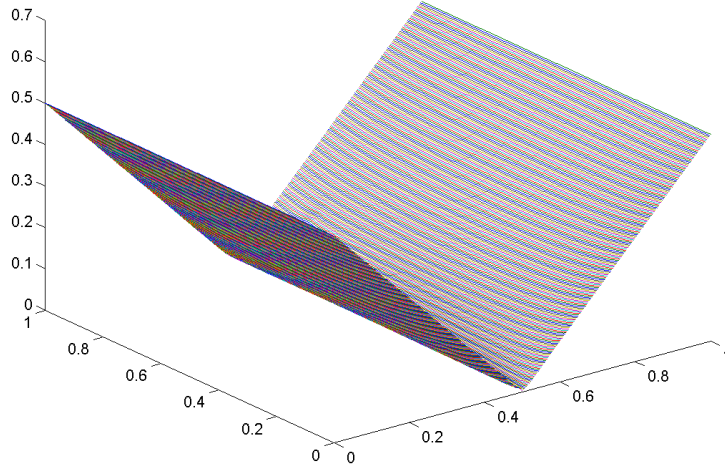


FIGURE 4. Test 7, Method 3, $h = 1/2^4$, $d = 5$, $\nu = 50$.

Remark 4.1. *With $A = \nu I$, with I the 2×2 identity matrix, Theorem 3.5 gives conditions under which the discrete version of (1.6) will converge. These conditions seem difficult to be met in practice. This may explain the deterioration of the convergence in Tables 9 for large scale problems and probably the reason the method does not perform well in three dimensions for non-smooth problems. We note that for $n = 3$, a different variational formulation was given in [17]. Convergence results similar to Theorem 3.5 can be analyzed in that case as well.*

4.2. Pucci equation. We use Method 3. Since there is no convexity requirement we could have simply used C^0 splines or Lagrange elements. The zero function was taken as initial guess. Recall, (1.3) that $\alpha \geq 1$. We consider two types of test functions, a smooth one and a problem for which no exact solution is known but cannot be in $H^2(\Omega)$.

Test 8: An exact radial solution, $-((x + 1)^2 + (y + 1)^2)^{1/2 - \alpha/2}$

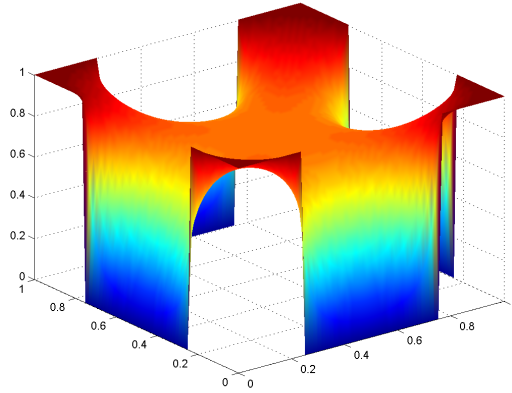
Test 9: The solution is not in $H^2(\Omega)$ and no exact formula is known. The function $g(x_1, x_2)$ is 1 except on the set $\{(x_1, x_2) \in [0, 1]^2, x_1 = 0, 1, 1/4 < x_2 < 3/4\} \cup \{(x_1, x_2) \in [0, 1]^2, x_2 = 0, 1, 1/4 < x_1 < 3/4\}$ where it is 0.

There does not seem to be a straightforward connection between the number of iterations and the value of ν .

However with finite differences, the higher ν , the higher is the number of iterations. Also, as noted in [11], the higher α , the more hyperbolic the problem becomes and more difficult to solve with a Laplacian based solver.

In figure 4.2, we plot the section of the solution on the line $x = 1/2$ with increasing values of $\alpha = 2, 2.5, 3, 3.5$, verifying numerically the discrete comparison principle.

h	n_{it}	L^2 norm	H^1 norm	H^2 norm
$1/2^0$	54	$1.7400 \cdot 10^{-3}$	$8.8611 \cdot 10^{-3}$	$6.5398 \cdot 10^{-2}$
$1/2^1$	55	$2.1264 \cdot 10^{-4}$	$2.9641 \cdot 10^{-3}$	$5.4038 \cdot 10^{-2}$
$1/2^2$	58	$2.5409 \cdot 10^{-5}$	$7.3783 \cdot 10^{-4}$	$2.8321 \cdot 10^{-2}$
$1/2^3$	56	$3.0881 \cdot 10^{-6}$	$1.7844 \cdot 10^{-4}$	$1.3870 \cdot 10^{-2}$
$1/2^4$	56	$3.9522 \cdot 10^{-7}$	$4.3379 \cdot 10^{-5}$	$6.7526 \cdot 10^{-3}$
$1/2^5$	56	$5.7712 \cdot 10^{-8}$	$1.0643 \cdot 10^{-5}$	$3.3088 \cdot 10^{-3}$
$1/2^6$	58	$1.0558 \cdot 10^{-8}$	$2.6307 \cdot 10^{-6}$	$1.6330 \cdot 10^{-3}$
$1/2^7$	58	$2.3611 \cdot 10^{-9}$	$6.5122 \cdot 10^{-7}$	$8.0750 \cdot 10^{-4}$

TABLE 10. Test 8: Method 3, $d = 3, r = 1, \nu = 5, \alpha = 2.5$ FIGURE 5. Test 9, $d = 3, \nu = 25$ graph of the solution with $\alpha = 2.5$

We also found that with random perturbation on the right hand side f , the scheme still reproduces a smooth quadratic solution.

4.3. Gauss curvature equation. Following [18], we considered the Gauss curvature equation (1.4) on the domain $[-0.57, 0.57]^2$ with boundary conditions $g(x, y) = x^2 + y^2 - 1$ and ask what is the maximum value of K for which there exists a convex solution. The vanishing moment method breaks down for $K = 2.2$. With Method 3, we are able to capture a convex solution for K as large as 11.2. The initial guess was taken as the solution of the Monge-Ampère equation $\det D^2u = K$.

Remark 4.2. *The iterations were stopped when the L^∞ norm of the difference between two iterates is less than 10^{-10} or when that value is bigger than the previously computed.*

Remark 4.3. *We could have performed the simulations with finite differences even for the two dimensional Monge-Ampère equation due to a remarkable property of the*

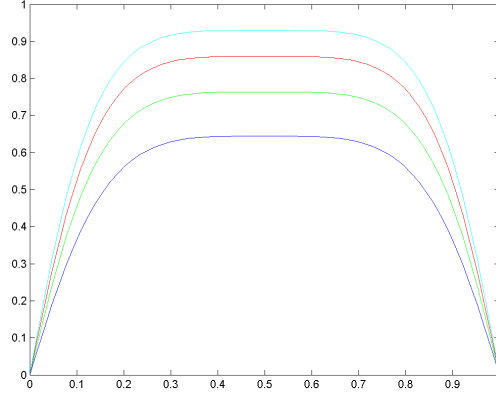


FIGURE 6. Test 9, $d = 3, \nu = 25$ sections of the solutions with increasing values of $\alpha = 2, 2.5, 3, 5$

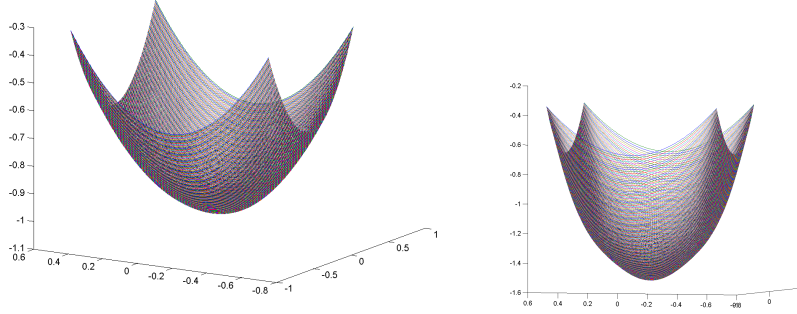


FIGURE 7. Gauss curvature equation by Method 3, $\nu = 50, h = 1/2, d = 3$ and with $K = 2.2$ and $K = 11.2$ respectively

discrete Hessian, that is the symmetric matrix with entries $\mathcal{D}_{k,l}u, k, l = 1, \dots, 2$, where

$$\begin{aligned} \mathcal{D}_{1,1}u_{ij} &= \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h^2}, \\ \mathcal{D}_{2,2}u_{ij} &= \frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{h^2} \text{ and} \\ \mathcal{D}_{1,2}u_{ij} &= \frac{u_{i+1,j+1} + u_{i-1,j-1} - u_{i-1,j+1} - u_{i+1,j-1}}{4h^2}. \end{aligned}$$

In [1], it is shown that the limit of a sequence of grid functions, with semi-positive definite Hessian, which converges in a suitable norm is a convex function. The obvious criticism of finite differences is the difficulty to deal with non rectangular domains. However, for Test 8 with the Monge-Ampère equation, we obtained better results with finite differences. We also give results for Test 8 with finite difference methods. The structure of Table 4.3 is similar to Table 4, p. 14 in [27]. The scheme used here is more accurate by several order of magnitude and is conceptually simpler. As pointed out in the introduction and the numerical results section, the variant of Method 3,

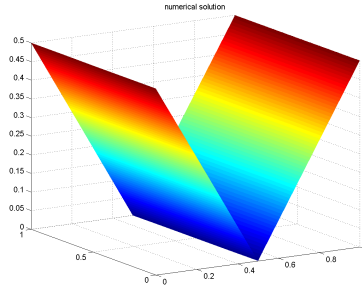


FIGURE 8. By finite differences, Test 8 Monge-Ampère, Method 3, $h = 1/2^2, \nu = 5$

α	h				
	$1/2^4$	$1/2^5$	$1/2^6$	$1/2^7$	$1/2^8$
2	$1.8717 \cdot 10^{-5}$	$4.6818 \cdot 10^{-6}$	$1.1715 \cdot 10^{-6}$	$2.9291 \cdot 10^{-7}$	$7.3233 \cdot 10^{-8}$
2.5	$3.2042 \cdot 10^{-5}$	$8.0433 \cdot 10^{-6}$	$2.0108 \cdot 10^{-6}$	$5.0288 \cdot 10^{-7}$	$1.2572 \cdot 10^{-7}$
3	$4.7162 \cdot 10^{-5}$	$1.1836 \cdot 10^{-5}$	$2.9584 \cdot 10^{-6}$	$7.3989 \cdot 10^{-7}$	$1.8497 \cdot 10^{-7}$
3.5	$6.2880 \cdot 10^{-5}$	$1.5700 \cdot 10^{-5}$	$3.9248 \cdot 10^{-6}$	$9.8150 \cdot 10^{-7}$	$2.4537 \cdot 10^{-7}$

TABLE 11. Test 8, $d = 3, \nu = 5$

namely (4.1) performs well for Test 5 with finite difference methods. Surprisingly, with $m = 25, \nu = 50$, we obtained the same results with the three dimensional iterative method introduced in [5],

$$\Delta u_{k+1} = ((\Delta u_k)^3 + 9(f - \det D^2 u_k))^{\frac{1}{3}}.$$

The errors in the maximum norm were given by $3.0976 \cdot 10^{-3}, 1.0432 \cdot 10^{-3}, 1.4169 \cdot 10^{-3}, 1.3766 \cdot 10^{-3}, 1.1017 \cdot 10^{-3}, 8.3671 \cdot 10^{-4}, 3.6635 \cdot 10^{-5}$ for $h = 1/2^k, k = 2, \dots, 8$ respectively.

Remark 4.4. We have also experimented numerically with Method 3 for the Bellman-Isaacs equation using finite differences. The Bellman-Isaacs equations are given by $F(x, D^2 u) := \sup_{\beta \in \mathcal{B}} \inf_{\alpha \in \mathcal{A}} (L_{\alpha, \beta} u(x) - f_{\alpha, \beta}(x)) = 0$ for given sets \mathcal{A} and \mathcal{B} , given functions $f_{\alpha, \beta}$ and $L_{\alpha, \beta}$ a family of second order operators. Take $L_{\alpha, \beta} u(x) = \text{tr}(\sigma \sigma' D^2 u(x))$ and $f_{\alpha, \beta}(x)$ chosen as $L_{\alpha, \beta} v(x), v(x)$ exact solution. The scheme performs well with a smooth solution $u_1(x, y) = \sin(x) \sin(y)$ even with the symmetric matrix σ given by $\sigma_{11} = x^2, \sigma_{12} = \frac{1}{2}xy, \sigma_{22} = y^2$, for which no fixed narrow stencil can work in general [16]. However, with a non smooth solution such as $u_2(x, y) = \sin(y/2) \sin(x/2)$ if $-\pi \leq x < 0$ and $u_2(x, y) = \sin(y/2) \sin(x/4)$ if $0 \leq x \leq \pi$ and the matrix σ given by $\sigma_{11} = \sin(x + y), \sigma_{12} = \sigma_{23} = \beta, \sigma_{13} = \sigma_{22} = 0$, we could obtain convergent results by choosing the domain as $[-1/2^m, 1/2^m]$ for fixed m and ν high. This suggests the need for a domain decomposition approach.

Remark 4.5. The operator L in (1.5) may be taken as the biharmonic operator. In which case, we add the boundary condition $\Delta u_{k+1} = 1/\nu^2$. The resulting algorithm may not be analyzed with the techniques discussed in this paper.

Remark 4.6. *Have we approximated the viscosity solutions in the non-smooth case? Method 3 approximates the solution of a fully nonlinear equation by C^2 functions. This by itself defines a notion of weak solution. It would be interesting to find out if the limit is always unique and the connections with other definitions of weak solutions.*

ACKNOWLEDGMENTS

The author acknowledges discussions with F. Celiker, B. Cockburn, W. Gangbo, R. Glowinski, M.J. Lai, R. Nochetto, A. Oberman and A. Regev. The author was supported in part by NSF grant DMS-0811052 and the Sloan Foundation. This research was supported in part by the Institute for Mathematics and its Applications with funds provided by the National Science Foundation.

REFERENCES

1. Néstor E. Aguilera and Pedro Morin, *Approximating optimization problems over convex functions*, Numer. Math. **111** (2008), no. 1, 1–34.
2. G. Awanou, *Energy methods in 3D spline approximations of the Navier-Stokes equations*, Ph.D. Dissertation, University of Georgia, Athens, Ga, 2003.
3. G. M. Awanou and M. J. Lai, *On convergence rate of the augmented Lagrangian algorithm for nonsymmetric saddle point problems*, Appl. Numer. Math. **54** (2005), no. 2, 122–134.
4. Gerard Awanou, *Robustness of a spline element method with constraints*, J. Sci. Comput. **36** (2008), no. 3, 421–432.
5. ———, *Spline element method for the Monge-Ampère equation*, Submitted, 2010.
6. Gerard Awanou and Ming-Jun Lai, *Trivariate spline approximations of 3D Navier-Stokes equations*, Math. Comp. **74** (2005), no. 250, 585–601 (electronic).
7. Gerard Awanou, Ming-Jun Lai, and Paul Weston, *The multivariate spline method for scattered data fitting and numerical solution of partial differential equations*, Wavelets and splines: Athens 2005, Mod. Methods Math., Nashboro Press, Brentwood, TN, 2006, pp. 24–74.
8. Victoria Baramidze and Ming-Jun Lai, *Spherical spline solution to a PDE on the sphere*, Wavelets and splines: Athens 2005, Mod. Methods Math., Nashboro Press, Brentwood, TN, 2006, pp. 75–92.
9. Jean-David Benamou and Yann Brenier, *A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem*, Numer. Math. **84** (2000), no. 3, 375–393.
10. Jean-David Benamou, Brittany Froese, and Adam Oberman, *Two numerical methods for the elliptic Monge-Ampère equation*, 2010.
11. L. A. Caffarelli and R. Glowinski, *Numerical solution of the Dirichlet problem for a Pucci equation in dimension two. Application to homogenization*, J. Numer. Math. **16** (2008), no. 3, 185–216.
12. Luis A. Caffarelli and Xavier Cabré, *Fully nonlinear elliptic equations*, American Mathematical Society Colloquium Publications, vol. 43, American Mathematical Society, Providence, RI, 1995.
13. R. Courant and D. Hilbert, *Methods of mathematical physics. Vol. II*, Wiley Classics Library, John Wiley & Sons Inc., New York, 1989, Partial differential equations, Reprint of the 1962 original, A Wiley-Interscience Publication.
14. E. J. Dean and R. Glowinski, *Numerical methods for fully nonlinear elliptic equations of the Monge-Ampère type*, Comput. Methods Appl. Mech. Engrg. **195** (2006), no. 13-16, 1344–1386.
15. Edward J. Dean and Roland Glowinski, *On the numerical solution of a two-dimensional Pucci's equation with Dirichlet boundary conditions: a least-squares approach*, C. R. Math. Acad. Sci. Paris **341** (2005), no. 6, 375–380.
16. K. Debrabant and E.R. Jakobsen, *Semi-lagrangian schemes for linear and fully nonlinear diffusion equations*, Submitted, 2010.

17. X. Feng and M. Neilan, *Analysis of Galerkin methods for the fully nonlinear Monge-Ampère equation*, To appear, J. Scientific Computing 2010, 2009.
18. ———, *Vanishing moment method and moment solutions for second order fully nonlinear partial differential equations*, J. Sci. Comput. **38** (2009), no. 1, 74–98.
19. B.D. Froese and A.M. Oberman, *Convergent finite difference solvers for viscosity solutions of the elliptic Monge-Ampere equation in dimensions two and higher*, Submitted, 2010.
20. David Gilbarg and Neil S. Trudinger, *Elliptic partial differential equations of second order*, Classics in Mathematics, Springer-Verlag, Berlin, 2001, Reprint of the 1998 edition.
21. Roland Glowinski, *Numerical methods for fully nonlinear elliptic equations*, ICIAM 07—6th International Congress on Industrial and Applied Mathematics, Eur. Math. Soc., Zürich, 2009, pp. 155–192.
22. Xian-Liang Hu, Dan-Fu Han, and Ming-Jun Lai, *Bivariate splines of various degrees for numerical solution of partial differential equations*, SIAM J. Sci. Comput. **29** (2007), no. 3, 1338–1354 (electronic).
23. Jürgen Jost, *Partial differential equations*, second ed., Graduate Texts in Mathematics, vol. 214, Springer, New York, 2007.
24. Ming-Jun Lai, *Convex preserving scattered data interpolation using bivariate C^1 cubic splines*, J. Comput. Appl. Math. **119** (2000), no. 1-2, 249–258, Dedicated to Professor Larry L. Schumaker on the occasion of his 60th birthday.
25. Ming-Jun Lai and Larry L. Schumaker, *Spline functions on triangulations*, Encyclopedia of Mathematics and its Applications, vol. 110, Cambridge University Press, Cambridge, 2007.
26. Grégoire Loeper and Francesca Rapetti, *Numerical solution of the Monge-Ampère equation by a Newton's algorithm*, C. R. Math. Acad. Sci. Paris **340** (2005), no. 4, 319–324.
27. Adam M. Oberman, *Wide stencil finite difference schemes for the elliptic Monge-Ampère equation and functions of the eigenvalues of the Hessian*, Discrete Contin. Dyn. Syst. Ser. B **10** (2008), no. 1, 221–238.
28. V. I. Oliker and L. D. Prussner, *On the numerical solution of the equation $(\partial^2 z / \partial x^2)(\partial^2 z / \partial y^2) - ((\partial^2 z / \partial x \partial y))^2 = f$ and its discretizations. I*, Numer. Math. **54** (1988), no. 3, 271–293.
29. Neil S. Trudinger and Xu-Jia Wang, *Boundary regularity for the Monge-Ampère and affine maximal surface equations*, Ann. of Math. (2) **167** (2008), no. 3, 993–1028.

DEPARTMENT OF MATHEMATICAL SCIENCES, NORTHERN ILLINOIS UNIVERSITY, DEKALB, IL, 60115

E-mail address: `awanou@math.niu.edu`

URL: `http://www.math.niu.edu/~awanou`