DEFINITION AND CERTAIN CONVERGENCE PROPERTIES OF A TWO-SCALE METHOD FOR MONGE-AMPÈRE TYPE EQUATIONS

HEIKO KRÖNER

Abstract. The Monge-Ampère equation arises in the theory of optimal transport. When more complicated cost functions are involved in the optimal transportation problem, which are motivated e.g. from economics, the corresponding equation for the optimal transportation map becomes a Monge-Ampère type equation. Such Monge-Ampère type equations are a topic of current research from the viewpoint of mathematical analysis. From the numerical point of view there is a lot of current research for the Monge-Ampère equation itself and rarely for the more general Monge-Ampère type equation. Introducing the notion of discrete Q-convexity as well as specifically designed barrier functions this purely theoretical paper extends the very recently studied two-scale method approximation of the Monge-Ampère itself [23] to the more general Monge-Ampère type equation as it arises e.g. in [25] in the context of Sobolev regularity.

1. Introduction

The starting point and motivation on the very basic level for our paper is Monge's transportation problem which is formulated in [22]. Here we recall it by using its formulation from the introduction of [12]. Let $0 \leq f^-, f^+ \in L^1(\mathbb{R}^n)$ be probability densities with respect to the Lebesgue measure L^n on $\mathbb{R}^n$ and $c : \mathbb{R}^n \times \mathbb{R}^n \to [0, +\infty]$ a cost function. Then Monge's optimal transport problem consists in finding a mapping $G : \mathbb{R}^n \to \mathbb{R}^n$ which pushes $d\mu^+ = f^+ dL^n$ forward to $d\mu^- = f^- dL^n$ and which minimizes the expected transportation cost

$$\inf_{G_\# \mu^+ = \mu^-} \int_{\mathbb{R}^n} c(x, G(x)) d\mu^+(x) \tag{1.1}$$

where $G_\# \mu^+ = \mu^-$ means $\mu^-[Y] = \mu^+[G^{-1}(Y)]$ for each Borel set $Y \subset \mathbb{R}^n$. It is of interest under which conditions such a map G exists and, furthermore, under which conditions such a map has a certain classical or Sobolev regularity, for details concerning this we refer to [12] and [25] and the references to the literature therein. Under appropriate assumptions which are not stated here explicitly it turns out that the optimal transportation map u satisfies the following Monge-Ampère type equation

$$\begin{aligned} \det\left(D^2u - A(x, Du)\right) =& f \quad \text{in } \Omega \\ u =& g, \quad \text{on } \partial\Omega \end{aligned} \tag{1.2}$$

where $\Omega \subset \mathbb{R}^n$ is a bounded open set, $f > c_0$ where $c_0 > 0$ is a constant,

$$D^2u - A(x, Du) > 0 \tag{1.3}$$

in Ω and

$$\bar{\Omega} \times \mathbb{R}^n \ni (x,p) \mapsto A(x,p) \in \mathbb{R}^{n\times n} \tag{1.4}$$

is a C^∞-smooth matrix valued function and $f \in C^0(\bar{\Omega})$, $g \in C^0(\partial\Omega)$. Note that the assumed regularity for A, f and g is not minimal from the point of view of mathematical analysis but still quite low as well as challenging and interesting for a first approach with numerical analysis. In the next section we will present the most important examples of cost functions and derive in these special cases some properties for A which result from these special cases. While basically working with a general A in our paper we will for the sake of simplicity assume that A satisfies these latter assumptions, see Section 2.

Note that (1.2) reduces to the classical Monge-Ampère equation when $A = 0$.

As a survey and without claiming completeness we give the following list of references concerning approximation schemes for the Monge-Ampère equation [24, 9, 10, 3, 6, 7, 7, 15, 2, 21, 8, 13, 1]. We are not aware of any works about the finite element approximation error analysis for the Monge-Ampère type equation (1.2) with $A \neq 0$.

The first purpose of our paper is to adapt the two-scale method definition from [23] to a modified two-scale method for an approximation of the Monge-Ampère type equation (1.2). This is not completely straightforward and unique and we make an appropriate choice for the definition. Second and mainly we show convergence of the discrete solutions defined by our two-scale method to the solution of (1.2) when the two discrete parameters go to zero as well as their quotients satisfy certain bounds. For it we make certain (regularity) assumptions for the solution of (1.2), cf. Remark 7.1. Basically the convergence proof is achieved following the strategy from [23] by using suitable barriers and comparison principles. The main advance and crucial difference of our paper from [23] is that we design completely new and much more complicated barrier functions. Apart from the barriers themselves the arguments are much more involved since we have to handle the terms arising from A. This becomes especially obvious from the fact that we have only a so called $\mod O(h)$ uniqueness in the comparison principle on the discrete level which is still non-trivial. Furthermore, our convergence result for the convergence of the discrete solutions to the solution of the original problem is different since we require certain regularity assumptions for the solution (Remark 7.1).

Our paper is organized as follows. In section 2 we introduce the assumptions on the cost function and discuss the two most important cases. In section 3 we define our two-scale method. In section 4 we derive a discrete comparison principle and uniqueness for the discrete equation $\mod O(h)$. In section 5 we show existence of a discrete solution. In sections 6 and 7 we present some auxiliary facts. In section 8 we study the convergence properties of the discrete model.

2. Assumptions on the cost function and the setting in general

Here we first recall the setting from [25] and [12] concerning the general setup for the optimal transport problem. Then we specify the cost function and derive further properties for the matrix function A in these specific cases. These motivate further assumptions for the matrix function A (in addition to those from the above mentioned and below described general setup) which we will assume throughout the paper.

The general setting in [25] and [12] is motivated from the applications and the purpose to achieve certain regularity properties. We will present these assumptions in the following and will afterwards discuss the two most important examples of cost functions, especially it turns out that the corresponding matrices A for these examples are smooth. Let $X \subset \mathbb{R}^n$ be an open set and $u : X \mapsto \mathbb{R}$ be a c-convex function, i.e., u can be written as

$$u(x) = \max_{y \in \bar{Y}} \{-c(x,y) + \lambda_y\} \tag{2.1}$$

for some open set $Y \subset \mathbb{R}^n$ and $\lambda_y \in \mathbb{R}$ for all $y \in \bar{Y}$. We are going to assume that u is an Alexandrov solution of (1.2) inside some open set $\Omega \subset X$, i.e.,

$$|\partial^c u(E)| = \int_E f \quad \forall E \subset \Omega \quad \text{Borel}, \tag{2.2}$$

where

$$\partial^c u(E) := \bigcup_{x \in E} \partial^c u(x), \quad \partial^c u(x) := \{y \in \bar{Y} : u(x) = -c(x,y) + \lambda_y\} \tag{2.3}$$

and $|F|$ denotes the Lebesgue measure of a set F. For $y \in \bar{Y}$ we define the contact set

$$\Lambda_y := \{x \in X : u(x) = -c(x,y) + \lambda_y\}. \tag{2.4}$$

Let $O \subset\subset Y$ be an open neighborhood of $\partial^c u(\Omega)$. We define

$$|||c||| := \|c\|_{C^3(\bar{\Omega} \times \bar{O})} + \|D_{xxyy}c\|_{L^\infty(\bar{\Omega} \times \bar{O})} \tag{2.5}$$

and assume

(1) $|||c||| < \infty$
(2) For every $x \in \Omega$ and $p := -D_x c(x,y)$ with $y \in O$ it holds that

$$D_{p_l p_k} A_{ij}(x,p) \xi_i \xi_j \eta_k \eta_l \geq 0 \quad \forall \xi, \eta \in \mathbb{R}^n, \quad \xi \cdot \eta = 0 \tag{2.6}$$

where A is defined through c by

$$A_{ij}(x,p) := -D_{x_i x_j} c(x,y). \tag{2.7}$$

(3) For every $(x,y) \in \Omega \times O$ the maps $x \in \Omega \mapsto -d_y c(x,y)$ and $y \in O \mapsto -D_x c(x,y)$ are diffeomorphisms on their respective ranges.

Special choices for the cost function arise from the applications, for a motivation of such choices in an economical context we refer to [11]. Nevertheless, the two most relevant special cases for the cost function c are the following functions $c = c_1$ and $c = c_2$, cf. [12], for which we will derive the mapping A explicitly, namely

$$c_1(x,y) = \frac{1}{2}|x-y|^2 \quad \text{and} \quad c_2(x,y) = -\log|x-y|. \tag{2.8}$$

For $c = c_1$ we have

$$D_x c = x - y, \quad p = y - x \tag{2.9}$$

and hence

$$A_{ij}(x, y-x) = -I, \tag{2.10}$$

or, equivalently,

$$A_{ij}(x,\xi) = -I \quad \forall \xi. \tag{2.11}$$

For $c = c_2$ we have

$$\begin{aligned} D_x c =& -\frac{x-y}{|x-y|^2}, \\ D_{x_i x_j} c =& 2\frac{(x_i-y_i)(x_j-y_j)}{|x-y|^4} - \frac{\delta_{ij}}{|x-y|^2} \\ p =& -D_x c = \frac{x-y}{|x-y|^2} \end{aligned} \tag{2.12}$$

and hence

$$A_{ij}\left(x, \frac{x-y}{|x-y|^2}\right) = \frac{\delta_{ij}}{|x-y|^2} - 2\frac{(x_i-y_i)(x_j-y_j)}{|x-y|^4}, \tag{2.13}$$

or, equivalently,

$$A_{ij}(x,\xi) = |\xi|^2\delta_{ij} - 2\xi_i\xi_j \quad \forall \xi. \tag{2.14}$$

In these special cases the following assumption is valid.

Assumption 2.1. A is C^∞-smooth and in addition there holds

$$A(x,0) = 0 \quad \forall x \in \bar{\Omega} \quad \text{or} \quad A = -I. \tag{2.15}$$

Motivated by these two special cases and since we need such properties for technical reasons we will assume throughout the paper that Assumption 2.1 holds.

3. Definition of the two-scale method for Monge-Ampère type equations

In this section we adapt the definition of the two-scale method from [18] to the more general equation (1.2). Let $T_h = \{T_1, ..., T_N\}$, $h > 0$, be a shape-regular and quasi-uniform mesh consisting of closed simplices T_i, $i = 1, ..., N$, of diameter ch where here and in the following c denotes a generic constant which may vary from line to line. We furthermore denote

$$\Omega_h = \int \left(\bigcup_{i=1}^{N} T_i\right), \tag{3.1}$$

let N_h be the nodes of T_h and write $N_h^b = \{x_i \in N_h : x_i \in \partial\Omega_h\}$ for the boundary nodes and $N_h^0 = N_h \setminus N_h^b$ for the interior nodes. We furthermore assume that Ω is convex, that $N_h^b \subset \partial\Omega$ and denote the space of continuous functions on Ω_h, which are linear on T_i for every $i = 1, ..., N$, by V_h. We denote the set of $n \times n$ matrices of real numbers by $\mathbb{R}^{n\times n}$ and the subset of orthogonal matrices by $O(n)$, furthermore, we write elements $V \in \mathbb{R}^{n\times n}$ by $V = (v_j)_{j=1}^d$ where v_j are the columns of V with respect to the standard basis in $\mathbb{R}^n$.

We denote the unit sphere in $\mathbb{R}^n$ by S and for $\theta > 0$ we let S_θ be a finite subset of S with the property that

$$\forall\ v \in S \quad \exists\ v_\theta \in S_\theta : \quad |v - v_\theta| \leq \theta. \tag{3.2}$$

Especially, we may assign to an element $V = (v_j) \in O(n)$ a matrix $V = (v_j^\theta)$ where $v_j^\theta = (v_j)_\theta$ and denote the set of all such matrices by $O^\theta(n)$. Note that $O^\theta(n) \not\subset O(n)$ in general.

In addition to the meshsize h which will serve as the fine scale in the remaining part of the paper we introduce in the following a coarse scale $\delta > h$ as a second discrete parameter which will serve as step size in difference quotients defining discrete derivatives. For $x_i \in N_h^0$ let

$$\delta_i = \min\{\delta, \operatorname{dist}(x_i, \partial\Omega_h)\} \tag{3.3}$$

and note that $\delta_i \geq ch$ where c does not depend on h and that $B(x_i, \delta_i) \subset \Omega_h$. Here, $B(x_i, \delta_i)$ denotes the open ball of radius δ_i around x_i. For $w \in C^0(\overline{\Omega_h})$ we define the one-sided first order difference operator

$$\nabla_\delta w(x_i, v_j) = \frac{w(x_i + \delta_i v_j) - w(x_i)}{\delta_i} \tag{3.4}$$

and the centered second order difference operator

$$\nabla_\delta^2 w(x_i; v_j) = \frac{w(x_i + \delta_i v_j) - 2w(x_i) + w(x_i - \delta_i v_j)}{\delta_i^2} \tag{3.5}$$

for $x_i \in N_h^0$ and $v_j \in S_\theta$. Here we choose one-sidedness in the definition for the first order difference operator but remark that centered differences might also work. Altogether we have three discrete parameters which we will summarize as

$$\varepsilon = (h, \delta, \theta) \tag{3.6}$$

where $h, \delta, \theta > 0$ and $\delta > h$. To the two latter inequalities we will sometimes refer to by writing $\varepsilon > 0$. For the following we will fix ε for a while and will analyze the corresponding discrete model. Then later in a second step we will discuss the limit $\varepsilon \to 0$ and a necessary coupling between the parameters in order to achieve convergence of the solutions of the discrete equations to the solution of the original equation.

In the following definition we generalize the two-scale operator from [18].

Definition 3.1. For $x_i \in N_h^0$ we define for any $w_h \in V_h$

$$\begin{aligned} T_\varepsilon[w_h](x_i) := \min_{v^\theta \in O^\theta(n)} \Big(& \prod_{j=1}^{d} \big(\nabla_\delta^2 w(x_i, v_j^\theta) - (v_j^\theta)^T A(x_i, \nabla_\delta w(x_i, e_k)) v_j^\theta\big)^+ \\ & - \sum_{j=1}^{d} \big(\nabla_\delta^2 w(x_i, v_j^\theta) - (v_j^\theta)^T A(x_i, \nabla_\delta w(x_i, e_k)) v_j^\theta\big)^- \Big) \end{aligned} \tag{3.7}$$

where two remarks are in order concerning our notation. Firstly, we write $(\cdot)^+ = \max(\cdot, 0)$ and $(\cdot)^- = -\min(\cdot, 0)$ to denote the non-negative and non-positive part of $(\cdot)$, respectively. Secondly, we abbreviate

$$\nabla_\delta w(x_i, e_k) := (\nabla_\delta w(x_i, e_k))_{k=1}^d \tag{3.8}$$

where $w \in V_h$, $\delta > h$, $x_i \in N_h^0$ and $(e_k)_{k=1}^d$ denotes the canonical basis in $\mathbb{R}^d$ to simplify the notation in expression (3.7).

By using the discrete two-scale operator from Definition 3.1 we obtain the following discrete version of the Monge-Ampère type problem (1.2).

Definition 3.2. For a given (triple) $\varepsilon > 0$ a two-scale method solution of (1.2) is a function $u_\varepsilon \in V_h$ such that $u_\varepsilon(x_i) = g(x_i)$ for all $x_i \in N_h^b$ and

$$T_\varepsilon[u_\varepsilon](x_i) := f(x_i) \tag{3.9}$$

for all $x_i \in N_h^0$.

In view of the widely used convention in numerical analysis to denote discrete solutions with the subscript h, i.e. u_h, we will write in the following ocassionally u_h instead of u_ε.

4. Discrete Q-convexity, monotonicity and discrete comparison principle $\mod O(h)$

To simplify the notation we use the following conventions. Firstly, in the setting from Definition 3.1 we will abbreviate in the following

$$Q(x_i, v_j) = Q_w(x_i, v_j) = \nabla_\delta^2 w(x_i, v_j) - (v_j)^T A(x_i, \nabla_\delta w(x_i, e_k))v_j \tag{4.1}$$

so that (3.7) takes the form

$$T_\varepsilon[w_h](x_i) = \min_{v^\theta \in O^\theta(n)} \Big(\prod_{j=1}^d \left(Q(x_i, v_j^\theta)\right)^+ - \sum_{j=1}^d \left(Q(x_i, v_j^\theta)\right)^- \Big). \tag{4.2}$$

Secondly, when a variable ranges in a discrete set we sometimes emphasize this fact by adding a superscript to this variable which is linked to this discrete set, e.g. we write $v^\theta \in O^\theta(n)$ and $w_h \in V_h$ but equally $v \in O^\theta(n)$ and $w \in V_h$, respectively. Throughout this section we assume that (the triple) $\varepsilon > 0$ is fixed.

Definition 4.1. We say that $w_h \in V_h$ is discretely Q-convex if

$$Q(x_i; v_j) \geq 0 \quad \forall x_i \in N_h^0, \quad \forall v_j \in O^\theta(n). \tag{4.3}$$

Note, that discrete Q-convexity of w_h does not imply convexity of w_h in general.

Lemma 4.2. *If $w_h \in V_h$ satisfies*

$$T_\varepsilon[w_h](x_i) \geq 0 \quad \forall x_i \in N_h^0, \tag{4.4}$$

then w_h is discretely Q-convex and as a consequence

$$T_\varepsilon[w_h](x_i) = \min_{v \in O^\theta(n)} \prod_{j=1}^d Q(x_i, v_j). \tag{4.5}$$

Proof. We distinguish two cases depending on whether $T_\varepsilon[w_h](x_i) > 0$ or not. Let $v = (v_j)_{j=1}^d \in O^\theta(n)$ be a d-tuple that realizes the minimum in the definition of $T_\varepsilon[w_h](x_i)$ and note that

$$\prod_{j=1}^d Q(x_i; v_j)^+ \geq 0, \quad \sum_{j=1}^d Q(x_i; v_j)^- \geq 0. \tag{4.6}$$

(i) Assume that $T_\varepsilon[w_h](x_i) > 0$. The expression $T_\varepsilon[w_h](x_i)$ is defined as a product, cf. (4.2), so that its positivity implies the positivity of all its factors. These positive factors are differences of type $a - b$ of non-negative numbers a and b so that we also always necessarily have $a > 0$. This implies that each quantity $Q(x_i; v_j)^+$ is also positive and hence the sum-term in (4.2) vanishes.

(ii) Assume that $T_\varepsilon[w_h](x_i) = 0$. Using again the representation from (4.2) of this expression as a difference of a product and a sum we make the following conclusion. If this product is positive then by (i) the sum vanishes and hence the product and the sum vanish so that $Q(x_i; v_j) = 0$ and the claim follows as well. □

Note that Lemma 4.2 and Definition 4.1 make formally sense when in (3.7) and (4.2) the superscript θ is omitted and Lemma 4.2 is then even also true.

We need a definition in which we introduce a family of subspaces of V_h.

Definition 4.3. For $\Lambda, h > 0$ we define

$$V_h^\Lambda = \left\{ v_h \in V_h : \exists \eta \in C^\infty(\bar{\Omega}), I_h\eta = v_h, \|\eta\|_{C^2(\bar{\Omega})} \leq \Lambda \right\} \tag{4.7}$$

where I_h denotes the usual Lagrange interpolation operator.

In the course of the paper it will turn out that when fixing a sufficiently large $\Lambda > 0$ all considerations can be done (and will be done) for the sequence of discrete spaces $(V_h^\Lambda)_{h>0}$. Here, Λ will be chosen depending on the data of the problem, i.e. depending on A, g, f, Ω and the uniform (with respect to h) parameters of the triangulation. Interestingly, we only bound derivatives up to the *second* order in the definition of V_h^Λ. So arguments solely based on interpolation do not work hence they require at least bounds for the third derivative of the interpolating function.

Remark 4.4. For the sake of a simplier notation we will write in the following again V_h instead of V_h^Λ with Λ large and fixed. We will comment on Λ where necessary.

As a consequence of Remark 4.4 we have the following discrete version of the fact that the derivative of a differentiable function vanishes in interior extremal points. Let $v_h \in V_h$, $v_j \in S_\theta$ and $z \in N_h^0$ be a maximum (or a minimum) of v_h then

$$\nabla_\delta v_h(z, v_j) = O(h). \tag{4.8}$$

In the next lemma we show that T_ε is monotone $\mod O(h)$.

Lemma 4.5. *Let $u_h, w_h \in V_h$ be discretely Q-convex. If $u_h - w_h$ attains a maximum at an interior node $z \in N_h^0$ then*

$$T_\varepsilon[w_h] \geq T_\varepsilon[u_h] + O(h) \tag{4.9}$$

in Ω_h. Here, the constant hidden in the $O(h)$-notation depends on Λ.

Proof. If $u_h - w_h$ attains a maximum at $z \in N_h^0$ then

$$u_h(z) - w_h(z) \geq u_h(x_i) - w_h(x_i) \quad \forall x_i \in N_h. \tag{4.10}$$

Since u_h and w_h are piecewise linear this inequality can be generalized to

$$u_h(z) - w_h(z) \geq u_h(x) - w_h(x) \quad \forall x \in \Omega_h. \tag{4.11}$$

Especially, evaluating this for difference quotients gives in view of (3.5) that

$$\nabla_\delta^2 u_h(z, v_j) \leq \nabla_\delta^2 w_h(z, v_j) \quad \forall v_j \in S_\Theta. \tag{4.12}$$

It remains to show that

$$Q_{u_h}(z, v_j) \leq Q_{w_h}(z, v_j) + O(h) \tag{4.13}$$

which can be reduced by (4.12) to

$$(v_j)^T A(z, \nabla_\delta w_h(z, e_k)) v_j \leq (v_j)^T A(z, \nabla_\delta u_h(z, e_k)) v_j + O(h). \tag{4.14}$$

But this follows since

$$\nabla_\delta w_h(z, e_k) - \nabla_\delta u_h(z, e_k) = O(h), \tag{4.15}$$

cf. (4.8), and

$$|\nabla_\delta w_h(x_i, e_k)| \leq C, \tag{4.16}$$

cf. Definition 4.3, from the continuity of A. □

We will use the following notation.

Remark 4.6. Given two functions $f_1 = f_1(x,h)$ and $f_2 = f_2(x,h)$ where $x \in S$ ranges in a certain parameter set S as well as the discretization parameter $h > 0$ we write

$$f_1 \leq f_2 \text{ for all } x \in S \mod O(h) \tag{4.17}$$

if there exists a constant $C > 0$ which does not depend on h, f_1 or f_2 such that

$$f_1(x,h) \leq f_2(x,h) + Ch \text{ for all } x \in S \text{ and all } h > 0. \tag{4.18}$$

When the parameter set S is clear from the context we will not mention it explicitly.

We show the following discrete comparison principle, cf. Lemma 4.7. Note that the inequality in the lemma includes the error term $O(h)$. This linear error order is not trivial in the context of a nonlinear second order operator and the use of first order finite elements for the following reason. If one derives the inequality in the following lemma firstly via a well-known comparison principle on the continuous level and later on transfers this by using interpolation estimates to the discrete level then usually third derivatives appear. But according to Definition 4.3 third derivatives of the interpolating functions may be arbitrary large and hence there appears an error term which is of the size of the product of the third derivative of an artificial interpolating function and h and hence possibly large and especially larger than $O(h)$.

Lemma 4.7. *Let $u_h, w_h \in V_h$ with $u_h \leq w_h$ on the boundary $\partial\Omega_h$ be such that*

$$T_\varepsilon[u_h](x_i) \geq T_\varepsilon[w_h](x_i) > 0 \quad \forall x_i \in N_h^0. \tag{4.19}$$

Then we have $u_h \leq w_h$ in Ω_h $\mod O(h)$.

Proof. Since $u_h, w_h \in V_h$, it suffices to prove $u_h(x_i) \leq w_h(x_i)$ for all $x_i \in N_h^0$. In view of Lemma 4.2 we may write inequality (4.19) as

$$\min_{v \in O^\theta(n)} \prod_{j=1}^d Q_{u_h}(x_i, v_j) \geq \min_{v \in O^\theta(n)} \prod_{j=1}^d Q_{w_h}(x_i, v_j) > 0 \quad \forall x_i \in N_h^0. \tag{4.20}$$

Now we distinguish cases. For it we fix constants $C_1, C_2 > 0$ which depend only on the data of the problem, i.e. on A, f, Ω, and which will be specified later.

(i) Let us assume

$$\min_{v \in O^\theta(n)} \prod_{j=1}^d Q_{u_h}(x_i, v_j) - C_1 h > \min_{v \in O^\theta(n)} \prod_{j=1}^d Q_{w_h}(x_i, v_j) \quad \forall x_i \in N_h^0. \tag{4.21}$$

We argue by contradiction and assume that there is $x_k \in N_h^0$ such that

$$u_h(x_k) - w_h(x_k) = \max_{x_i \in N_h^0} u_h(x_i) - w_h(x_i) > 0. \tag{4.22}$$

Similarly, as in Lemma 4.5 we conclude that

$$Q_{u_h}(x_k, v_j) \leq Q_{w_h}(x_k, v_j) + C_3 h \quad \forall v_j \in S_\theta \tag{4.23}$$

where $C_3 > 0$ is a suitable constant which depends only on the data of the problem (and especially not on h). Taking the product on both sides and after this the infimum on the left-hand side of the equation yields

$$\min_{v \in O^\theta(n)} \prod_{j=1}^d Q_{u_h}(x_k, v_j) \leq \prod_{j=1}^d Q_{w_h}(x_k, \tilde{v}_j) + C_2 h \quad \forall \tilde{v} \in O^\theta(n) \tag{4.24}$$

where C_2 is a suitable constant which depends only on C_3 and the data of the problem. W.l.o.g. we may also take the infimum over all $\tilde{v} \in O^\theta(n)$ on the right-hand side of the inequality.

Combining this with nequality (4.21) we obtain a contradiction provided C_1 is sufficiently large compared to C_2. This finishes case (i).

(ii) Let us assume the other case, i.e. we have

$$\min_{v \in O^\theta(n)} \prod_{j=1}^{d} Q_{u_h}(x_i, v_j) - C_1 h \leq \min_{v \in O^\theta(n)} \prod_{j=1}^{d} Q_{w_h}(x_i, v_j) \quad \forall x_i \in N_h^0 \tag{4.25}$$

where now C_1 is fixed as it turned out to be necessary in case (i).

The strategy of the proof will now be as follows. We show that there are constants $h_0, C_4, C_5 > 0$ and an auxiliary function $q_h \in V_h$ which depend on the data of the problem such that

$$\begin{aligned} T_\varepsilon[u_h + \alpha q_h](x) > T_\varepsilon[w_h](x) + C_1 h \\ \forall 0 < h < h_0 \qquad \forall \alpha \geq C_4 h \text{ sufficiently small} \end{aligned} \tag{4.26}$$

and

$$\|q_h\|_{L^\infty(\Omega)} \leq C_5. \tag{4.27}$$

From this we conclude by using (i) that

$$u_h \leq w_h + C_4 C_5 h \tag{4.28}$$

and hence the claim.

We choose $\tilde{x} \in \mathbb{R}^n$ such that $\operatorname{dist}(\tilde{x}, \bar{\Omega}) \geq 1$, $\lambda > 0$ large and define the strictly convex function

$$q(x) = e^{\lambda |x - \tilde{x}|^2} - R \tag{4.29}$$

where $R = R(\lambda, \Omega)$ is so that $q \leq 0$ in Ω and especially in $\bar{\Omega}_h$. We now define

$$q_h = I_h q, \tag{4.30}$$

perform some relevant calculations on the level of q instead of q_h and translate them to q_h afterwards by using the standard interpolation estimate

$$\|q_h - q\|_{C^m(\Omega_h)} \leq c_{m,r} h^r \|q\|_{C^{m+r}(\bar{\Omega})}, \quad m, r \in \mathbb{N}, \tag{4.31}$$

where $c_{m,r}$ are suitable constants. We have

$$\begin{aligned} D_i q(x) =& 2\lambda (x_i - \tilde{x}_i) e^{\lambda |x - \tilde{x}|^2} \\ D_i D_j q(x) =& 2\lambda e^{\lambda |x - \tilde{x}|^2} \delta_{ij} + 4\lambda^2 e^{\lambda |x - \tilde{x}|^2} (x_i - \tilde{x}_i)(x_j - \tilde{x}_j). \end{aligned} \tag{4.32}$$

For $x \in N_h^0$ and $v \in O^\theta(n)$ we calculate

(4.33)
$$\begin{aligned}Q_{u_h+\alpha q}(x, v_r) =&\nabla_\delta^2 u_h(x, v_r) + \alpha \nabla_\delta^2 q(x, v_r)\\ &- v_r^T A(x, \nabla_\delta u_h(x, e_k) + \alpha \nabla_\delta q(x, e_k))v_r\\ =&\nabla_\delta^2 u_h(x, v_r) + \alpha\left(O(\delta) + D_{v_r} D_{v_r} q(x)\right)\\ &- v_r^T A(x, \nabla_\delta u_h(x, e_k) + \alpha D_k q(x) + \alpha O(\delta))v_r\\ =&\nabla_\delta^2 u_h(x, v_r) + \alpha O(\delta) + 2\alpha\lambda e^{\lambda|x-\tilde{x}|^2}\delta_{ij} v_{ri} v_{rj}\\ &+ 4\alpha\lambda^2 e^{\lambda|x-\tilde{x}|^2}(x_i - \tilde{x}_i)(x_j - \tilde{x}_j)v_{ri}v_{rj}\\ &- v_r^T A(x, \nabla_\delta u_h(x, e_k) + 2\alpha\lambda(x_k - \tilde{x}_k)e^{\lambda|x-\tilde{x}|^2} + \alpha O(\delta))v_r\end{aligned}$$

Here, the constant hidden in the $O(\delta)$-notation depends on q, more precisely, on its higher order derivatives. Since $T_\varepsilon[u_h] =: f > 0$ there is $\alpha_0 > 0$ such that

$$T_\varepsilon[u_h + \alpha q_h] > 0 \tag{4.34}$$

for all $\alpha \in (0, \alpha_0)$. Hence for these α we have

$$T_\varepsilon[u_h + \alpha q_h](x) = \min_{v \in O^\theta(d)} \prod_{j=1}^d Q_{u_h+\alpha q_h}(x, v_j) \tag{4.35}$$

and therefore by abbreviating $Q(\alpha, j) = Q_{u_h+\alpha q_h}(x, v_j)$ (where the arguments x and v_j are assumed to be implicitly clear from the context) for $\alpha \in (0, \alpha_0)$ and $j \in \{1, ..., d\}$ we write

$$\begin{aligned}&\frac{d}{d\alpha} T_\varepsilon[u_h + \alpha q_h](x)_{|\alpha=0}\\ &=\sum_{j=1}^d Q(0, 1)...Q(0, j-1)\frac{d}{d\alpha}Q(\alpha, j)_{|\alpha=0} Q(0, j+1)...Q(0, d).\end{aligned} \tag{4.36}$$

Note that the arguments x and v which appear here implicitly on the right-hand side are chosen obviously - namely as on the left-hand side of the equation as far as x is concerned; the matrix v is chosen so that in the point x the infimum is attained in the definition (4.34). In order to evaluate (4.36) we first calculate the derivative of the expression in (4.33). Observe that there holds for all $k \in \{1, ..., d\}$ that

$$\begin{aligned}\frac{d}{d\alpha}Q(\alpha, k) =&O(\delta) + 2\lambda e^{\lambda|x-\tilde{x}|^2}\delta_{ij} v_{ik} v_{jk}\\ &+ 4\lambda^2 e^{\lambda|x-\tilde{x}|^2}(x_i - \tilde{x}_i)(x_j - \tilde{x}_j)v_{ki}v_{kj}\\ &- v_k^T\left(\frac{\partial A}{\partial p_l}(x, \nabla_\delta u_h(x, e_k))2\lambda(x_l - \tilde{x}_l)e^{\lambda|x-\tilde{x}|^2} + O(\delta)\right)v_k.\end{aligned} \tag{4.37}$$

Assuming that θ is sufficiently small we have

$$(x_i - \tilde{x}_i)(x_j - \tilde{x}_j)v_{ki}v_{kj} \geq \frac{1}{2}|x - \tilde{x}|^2 \tag{4.38}$$

for all $k \in \{1, ..., d\} \setminus \{k_0\}$ and all $v \in O^\theta(d)$ where $k_0 = k_0(v) \in \{1, ..., d\}$ is suitable. Now having Λ fixed in the definition of V_h^Λ and choosing $0 < h_0 \leq 1$ (at the moment not further specified) we may assume that

$$\frac{d}{d\alpha}Q(\alpha, k)_{|\alpha=0} \geq \lambda^2 e^{\lambda|x-\tilde{x}|^2}|x - \tilde{x}|^2 > 0 \tag{4.39}$$

provided $\lambda > 0$ is sufficiently large. Furthermore, for $\alpha_0 = \alpha_0(\lambda) > 0$ sufficiently small the quantities $Q(\alpha, k)$ are uniformly with respect to k and with respect to $\alpha \in (-\alpha_0, \alpha_0)$ bounded by a positive constant from below. Hence we arrive at

$$\frac{d}{d\alpha} T_\varepsilon[u_h + \alpha q_h](x)_{|\alpha=0} \geq \mu_0 > 0 \tag{4.40}$$

with a suitable fixed $\mu_0 > 0$. An expansion of $T_\varepsilon[u_h + \alpha q_h](x)$ around 0 yields the existence of $\alpha \in (0, \alpha_0)$ such that

$$T_\varepsilon[u_h + \alpha q_h](x) > T_\varepsilon[u_h](x) + \frac{\alpha}{2}\mu_0. \tag{4.41}$$

Clearly, by assuming that h_0 is sufficiently small the above construction shows that we can realize property (4.26) with this specific α. Actually, we have in addition to choose $C_4 > 0$ but as long it is not too large it does not matter how we choose it. This finishes the proof. □

5. Existence of discrete solutions

We now prove uniqueness $\mod O(h)$ and existence of a discrete solution $u_\varepsilon \in V_h$ of (3.9). Here, the uniqueness $\mod O(h)$ means, that given two discrete solutions $u_\varepsilon^1, u_\varepsilon^2$ of (3.9) there holds $u_\varepsilon^i \leq u_\varepsilon^j \mod O(h)$ for all $i, j \in \{1, 2\}$.

Lemma 5.1. *There exists $u_\varepsilon \in V_h$ which satisfies the discrete Monge-Ampère type equation (3.9) and which is unique $\mod O(h)$. Furthermore, $\|u_\varepsilon\|_{L^\infty(\Omega)}$ does not depend on the parameter $\varepsilon = (h, \delta, \theta)$.*

Proof. Let us fix $\varepsilon > 0$. The uniqueness $\mod O(h)$ of a solution of (3.9) follows from Lemma 4.7. Hence it remains to show existence. For it we construct a special monotone sequence of discretely Q-convex subsolutions $\{u_\varepsilon^k\}_{k=0}^\infty$ of (3.9) from which we will select a subsequence which converges to the desired discrete solution of (3.9). The construction is by induction and works as follows.

(i) *Claim: There is $u_h^0 \in V_h$ such that $u_h^0 = I_h g$ on $\partial\Omega_h$ and*

$$T_\varepsilon[u_h^0](x_i) \geq f(x_i) \quad \forall x_i \in N_h^0. \tag{5.1}$$

Proof of the claim:

(a) We give a short proof in the case that there is $C^{2,\alpha}$-regularity of the solution available. Let us assume that there is $0 < \alpha < 1$ such that the problem (1.2) with right-hand side f replaced by $f + 1$ has a solution $u \in C^{2,\alpha}(\bar{\Omega})$. Setting $u_h^0 = I_h u$ and using the interpolation estimates from [23, Lemma 4.1] as well as the continuity of A yields the claim.

(b) In the general case we proceed without using (regularity of) the solution of (1.2). Let q denote the auxiliary function from (4.29) with arbitrary choice of R, e.g. $R = 0$. Let w be a smooth function in a ball $B_L(0)$, $L > 0$ large, let us say $2\bar{\Omega} \subset B_L(0)$, with

$$w(x_i) = g(x_i) - q(x_i), \quad x_i \in N_h^b. \tag{5.2}$$

Such a function can easily be obtained by fixing it in N_h^b and then extending it as smooth function to $B_L(0)$. But we would like to have that the size of $|Dw(x)|$ and $|D^2 w(x)|$ is of order $O(\lambda e^{\lambda|x-\tilde{x}|^2})$ and hence small compared to the order $O(\lambda^2 e^{\lambda|x-\tilde{x}|^2})$ which is the size of $D^2q(x)$. For it we define an artificial domain $\tilde{\Omega} \subset \mathbb{R}^d$ with smooth boundary $\partial\tilde{\Omega}$ passing through all elements of N_h^b, i.e. $N_h^b \subset \partial\tilde{\Omega}$. We extend $g-q$ from N_h^b to a function $b \in C^{1,\alpha}(\partial\tilde{\Omega})$ with

$$\|b\|_{C_x^{1,\alpha}(\partial\tilde{\Omega})} \leq \mu_x \lambda^2 e^{\lambda|x-\tilde{x}|^2} \tag{5.3}$$

for all $x \in \partial\tilde{\Omega}$ where we may and will choose here the constant

$$0 < \mu_x < \mu_0 \tag{5.4}$$

with $\mu_0 > 0$ small. Here, we denote

$$\|b\|_{C_x^{1,\alpha}(\partial\tilde{\Omega})} = |b(x)| + \sum_{i=1}^{d} \|D_i b\|_{C_x^{0,\alpha}(\partial\tilde{\Omega})} \tag{5.5}$$

where

$$\|D_i b\|_{C_x^{0,\alpha}(\partial\tilde{\Omega})} = |D_i b(x)| + \sup_{y\in\partial\tilde{\Omega}, y\neq x} \frac{|D_i b(x) - D_i b(y)|}{|x-y|^\alpha}, \quad x \in \partial\tilde{\Omega}. \tag{5.6}$$

To give derivatives (and their norms) of a function being defined on the hypersurface $\partial\tilde{\Omega}$ a sense we either consider these with respect to a fixed finite selection of local coordinate systems covering $\partial\tilde{\Omega}$ or with respect to an arbitrary but fixed extension to an open neighborhood of $\partial\tilde{\Omega}$ of the corresponding functions. In order to understand how the representation (5.3) is possible, we explain this for the most non-trivial case, i.e. on the level of the Hölder norm of the derivative. Given a small choice for $\mu_x > 0$, we estimate for $i \in \{1, ..., d\}$ and $x, y \in \partial\tilde{\Omega}$ that

$$\begin{aligned} \frac{|D_i b(x) - D_i b(y)|}{|x-y|^\alpha} =& \frac{|D_i b(x) - D_i b(y)|}{|x-y|^\alpha |x-y|^{1-\alpha}} |x-y|^{1-\alpha} \\ \approx& D^2 q(x) |x-y|^{1-\alpha} \\ \leq& D^2 q(x) \mu_x^{1-\alpha} \end{aligned} \tag{5.7}$$

if $|x-y| \leq \mu_x$. In the other case, i.e. when $|x-y| > \mu_x$, we estimate

$$\begin{aligned} \frac{|D_i b(x) - D_i b(y)|}{|x-y|^\alpha} \leq& \frac{|D_i b(x)| + |D_i b(y)|}{\mu_x^\alpha} \\ \leq& O\left(\frac{\lambda}{\mu_x} e^{\lambda|x-\tilde{x}|^2}\right). \end{aligned} \tag{5.8}$$

Then we solve the Dirichlet problem

$$\begin{aligned} \Delta w =& 0 \quad \text{in } \tilde{\Omega} \\ w =& b \quad \text{on } \partial\tilde{\Omega} \end{aligned} \tag{5.9}$$

and obtain by classical PDE-theory a solution $w \in C^{3,\alpha}\left(\overline{\tilde{\Omega}}\right)$ which satisfies the Schauder-estimate

$$\|w\|_{C^{3,\alpha}(\overline{\tilde{\Omega}})} \leq c\left(\|w\|_{C^0\left(\overline{\tilde{\Omega}}\right)} + \|b\|_{C^{1,\alpha}(\partial\tilde{\Omega})}\right). \tag{5.10}$$

Noting that

$$\|w\|_{C_x^0(\tilde{\Omega})} = O(q(x)) \tag{5.11}$$

we conclude that w satisfies the desired properties. Now we set

$$u^0 = w + q \tag{5.12}$$

and then

$$u_\varepsilon^0 := I_h u^0. \tag{5.13}$$

By construction u_ε^0 has the correct boundary values, i.e. $u_\varepsilon^0(x_i) = g(x_i)$ when $x_i \in N_h^b$ and it satisfies

$$T_\varepsilon[u_\varepsilon^0](x_i) \geq f(x_i) \tag{5.14}$$

for all $x_i \in N_h^0$ provided λ is large in view of the interpolation error estimates in [23, Lemma 4.1]. Hence we have constructed u_ε^0 as desired. Note that we proved here a little bit more than needed. In order to apply [23, Lemma 4.1] it suffices to have only the Schauder estimate (5.10) on the level of $C^{2,\alpha}$ available. Furthermore, we remark that the construction can be done so that the L^∞-norm of u_ε^0 can be estimated uniformly in h.

(ii) We follow a Perron construction from [23] and use induction. First we label all interior nodes, let us say, $N_h^0 = \{x_1, ..., x_m\}$, $m \in \mathbb{N}$. The induction begins with $u_h^0 \in V_h$ from (i). Let us assume we already have constructed $u_h^k \in V_h$ for some $k \in \mathbb{N}$ such that

$$\begin{aligned} u_h^k \geq& u_h^0 \\ u_h^k(x_i) =& I_h g(x_i), \quad x_i \in N_h^b, \\ T_\varepsilon[u_h^k](x_i) \geq& f(x_i), \quad x_i \in N_h^0. \end{aligned} \tag{5.15}$$

In order to construct $u_h^{k+1} \in V_h$ which satisfies

$$u_h^{k+1} \geq u_h^k \tag{5.16}$$

as well as the properties (5.15) with k replaced by $k+1$ we first define auxiliary functions $u_h^{k,i} \in V_h$, $i = 0, ..., m$. We set

$$u_h^{k,0} := u_h^k. \tag{5.17}$$

Assume that $u_h^{k,i-1} \in V_h$ is already defined, $i \geq 1$. In order to define $u_h^{k,i} \in V_h$ we increase (only) the value of $u_h^{k,i-1}(x_i)$ (eventually) until

$$T_\varepsilon[u_h^{k,i}](x_i) = f(x_i). \tag{5.18}$$

This defines $u_h^{k,i}$. The equality in (5.18) can indeed be achieved under this process which becomes clear when we look at Lemma 4.2 and (4.1). Noting that the centered second differences appearing in this definition of Q are decreasing with slope $\frac{c}{h^2}$, c a generic constant, with respect to the central value for all directions and that all other expressions therein change under this process at most by a rate of $\frac{c}{h}$ the equality in (5.18) can clearly be achieved for h sufficiently small. This process potentially increases the second centered differences at all the other nodes x_j, $j \neq i$ at a rate $\frac{c}{h^2}$ and changes lower order terms at most at a rate $\frac{c}{h}$. Hence

$$T_\varepsilon[u_h^{k,i}](x_j) \geq T_\varepsilon[u_h^{k,i-1}](x_j) \geq f(x_j) \quad \forall j \neq i. \tag{5.19}$$

We repeat this process with the remaining nodes x_j for $i < j \leq m$ and set

$$u_h^{k+1} := u_h^{k,m}. \tag{5.20}$$

Note that the 'sufficient smallness' of h can be chosen here uniformly. Clearly, u_h^{k+1} satisfies (5.15) and (5.16).

(iii) We derive an a priori L^∞-bound for the sequence $(u_h^k)_{k\in\mathbb{N}}$. The lower bound for this sequence follows from the remarks at the end of steps (i) and (ii). Recall that by (2.15) we have $A(x,\cdot)=0$ or $A=-I$. The upper bound is chosen as follows. We set $\tilde{b}_h=\max_{x_i\in N_h^b} g(x_i)\in V_h$. In the case $A(x,\cdot)=0$ we set $b_h=\tilde{b}_h$ and in the case $A=-I$ we set $b_h=\tilde{b}_h+c(\Omega)-(1-\frac{1}{4}\min f)I_h|x|^2$ where $c(\Omega)$ is a positive constant which depends on Ω. Clearly, by the comparison principle b_h is an upper barrier mod $O(h)$ for the sequence $(u_h^k)_{k\in\mathbb{N}}$ and we are finished, note that we assume here that h is small.

(iv) Since $(u_h^k(x_i))_{k=1}^\infty$ is monotone and bounded from above for all $x_i\in N_h^0$ it converges and we set

$$u_\varepsilon(x_i)=\lim_{k\to\infty}u_h^k(x_i)\quad\forall x_i\in N_h^0 \tag{5.21}$$

and extend u_ε without relabeling to $u_\varepsilon\in V_h$. Then we have $u_\varepsilon=I_hg$ on $\partial\Omega_h$ and

$$T_\varepsilon[u_\varepsilon](x_i)\geq f(x_i)\quad\forall x_i\in N_h^0. \tag{5.22}$$

We show that even equality holds in (5.22) and assume for it that the inequality in (5.22) is strict for a certain $x_i\in N_h^0$. Then we find arbitrary large k such that

$$T_\varepsilon[u_h^k](x_i)>f(x_i). \tag{5.23}$$

But then in the construction of u_h^{k+1} in step (ii) there was a certain 'space' for increasement which contradicts that $(u_h^k(x_i))_k$ is especially pointwisely a Cauchy sequence. Hence we have shown existence of u_ε as desired and we also have obtained an a priori L^∞-bound which is independent from the discretization parameters. □

Hereby, our analysis of the discrete model is finished. The remaining sections are concerned with the convergence analysis of the discrete solutions u_ε, i.e. they show that the discrete solutions u_ε of (3.9) converge to the solution u of (1.2) when $\varepsilon\to 0$ and the relative size of h and δ satisfies a certain relation provided the original equation satisfies appropriate properties. Recall that $\varepsilon=(h,\delta,\theta)$.

6. A special auxiliary function

We construct in the following lemma a special auxiliary function.

Lemma 6.1. *Let Ω be uniformly convex, $h_0>0$ and $1<E\leq c(\Omega,h_0)$ be sufficiently large within this range, $c(\Omega,h_0)$ a suitable constant which depends only on Ω and h_0 with*

$$c(\Omega,h_0)\to\infty \tag{6.1}$$

as $h_0\to 0$ (this relation becomes more explicit in the proof). There exists a $h_0=h_0(\Omega)$ such that for all $0<h<h_0$ the following holds. For each node $z\in N_h^0$ and $\delta>0$ with $\operatorname{dist}(z,\partial\Omega_h)\leq\delta$ *there exists a function $p_h\in V_h$ and $E'>E$ such that $T_\varepsilon[p_h](x_i)\geq E'$ for all $x_i\in N_h^0$, $p_h\leq 0$ on $\partial\Omega_h$ and*

$$|p_h(z)|\leq CE'\delta \tag{6.2}$$

with C depending on Ω.

Proof. Let $z\in N_h^0$ and $\delta>0$ be arbitrary. Let $\tilde{z}\in\partial\Omega$ be a nearest boundary point, i.e.

$$|z-\tilde{z}|=\operatorname{dist}(z,\partial\Omega). \tag{6.3}$$

Let

$$0 < \kappa_1(x) \leq ... \leq \kappa_n(x) \tag{6.4}$$

be the ordered-by-size n principal curvatures of $\partial\Omega$ in $x \in \partial\Omega$ with respect to the outer unit normal of $\partial\Omega$ in x (the convention is here as usual so that e.g. a unit sphere has principal curvatures equal to 1). In view of the uniform convexity of $\partial\Omega$ we have

$$\kappa := \min_{\partial\Omega} \kappa_1 > 0. \tag{6.5}$$

For the moment we fix a point $\tilde{x} \in \mathbb{R}^n$ and a large $\lambda > 0$ and we will adjust them later appropriately. We define

$$f(x) = e^{\lambda|x-\tilde{x}|^2} - e^{\lambda|\tilde{z}-\tilde{x}|^2}, \quad x \in \mathbb{R}^n. \tag{6.6}$$

The function f looks roughly spoken like a bowl, attains a global minimum in $\tilde{x}$, is rotationally symmetric around $\tilde{x}$. Furthermore, it is strictly monotone increasing and strictly convex along rays starting from $\tilde{x}$ where in addition this convexity in radial direction at a point $x \in \mathbb{R}^n$ can be quantified as being of size $O(\lambda^2 e^{\lambda|x-\tilde{x}|^2})$. Here, the constant hidden in the O-notation depends on Ω and $\tilde{x}$.

Let us now adjust $\tilde{x} \in \mathbb{R}^n$ where we assume w.l.o.g. that $\tilde{x} \notin \bar{\Omega}$ and choose $R > \frac{1}{\kappa}$ such that

$$\partial\Omega \cap \partial B_R(\tilde{x}) = \{\tilde{z}\} \quad \text{and} \quad \Omega \subset B_R(\tilde{x}). \tag{6.7}$$

Let

$$c_1 = \max_{\bar{\Omega}} |x - \tilde{x}|, \quad c_2 = \min_{\bar{\Omega}} |x - \tilde{x}| \tag{6.8}$$

then the second derivatives of f in $\bar{\Omega}$ are of size at least $O(\lambda^2 c_2^2 e^{\lambda c_2^2})$ and the first derivatives are of size at most $O(c_1 \lambda e^{\lambda c_1})$. We increase λ until $O(\lambda^2 c_2^2 e^{\lambda c_2^2})$ is large compared to E and $O(c_1 \lambda e^{\lambda c_1})$. Then we set

$$E' = \max\{O(\lambda^2 c_2^2 e^{\lambda c_2^2}), O(c_1 \lambda e^{\lambda c_1})\} \tag{6.9}$$

as well as

$$p_h = I_h(f - f(z)). \tag{6.10}$$

Now we may assume that E and E' are bounded by a constant which may become arbitrary large provided $h_0(\Omega)$ is correspondingly small so that for $0 < h < h_0$ the previous interpolation does not produce relevant errors. This finishes the proof of the lemma. Note that we tacitly introduced concrete but generic constants in order to replace the O-notation in the context of inequalities. □

7. An approximating problem

In the following remark we formulate some rather weak assumptions for equation (1.2) and its solution which allow us to construct an approximating smooth problem with a smooth solution (which is also the main purpose of this section).

Remark 7.1. (1) (Regularity) $u \in C^1(\bar{\Omega})$ is a viscosity solution of (1.2).

(2) (Comparison principle one sided around the solution) There is $\varepsilon_1 > 0$ such that the following holds. Given continuous $\tilde{f}_1, \tilde{f}_2, \tilde{g}_1, \tilde{g}_2$ with $f \leq \tilde{f}_2 \leq \tilde{f}_1 \leq f + \varepsilon_1$ in Ω and $g - \varepsilon_1 \leq \tilde{g}_1 \leq \tilde{g}_2 \leq g$ on $\partial\Omega$ and continuous viscosity solutions $\tilde{u}_1$ and $\tilde{u}_2$ of (1.2) with respect to the data $\tilde{f}_1, \tilde{g}_1$ and $\tilde{f}_2, \tilde{g}_2$,

respectively, then there holds a comparison principle in the usual sense, i.e. $\tilde{u}_1 \leq \tilde{u}_2$ in Ω.

Let $\Omega_n \supset \Omega$, $n \in \mathbb{N}$, be an approximation of Ω by smooth convex sets with respect to the Hausdorff distance d_H, i.e.

$$0 < \operatorname{dist}_H(\Omega, \Omega_n) \leq \delta_n \to 0. \tag{7.1}$$

Let $p \in \Omega$ be arbitrary and fixed. The family of rays

$$\{R_p = \{p + te : t \geq 0\} : e \in \mathbb{R}^n, \|e\| = 1\} \tag{7.2}$$

clearly defines a bijection

$$b_n : \partial\Omega \to \partial\Omega_n \tag{7.3}$$

by mapping $R_p \cap \partial\Omega$ to $R_p \cap \partial\Omega_n$. Let f_n and g_n be smooth functions in $\mathbb{R}^n$ approximating f and g, respectively, such that

$$f < f_n, \quad g_n < g, \tag{7.4}$$

$$|g(x) - g_n(b_n(x))| \leq \delta_n, \quad |f(x) - f_n(y)| \leq \delta_n \tag{7.5}$$

for all

$$x \in \partial\Omega, y \in [x, b_n(x)] = \{tx + (1-t)b_n(x) : 0 \leq t \leq 1\} \tag{7.6}$$

and

$$|f(x) - f_n(x)| \leq \delta_n, \quad x \in \bar{\Omega}. \tag{7.7}$$

We note that the above approximations can be obtained in a standard fashion and indicate that especially inequalites (7.4) can be achieved by first replacing f by $f + \frac{\delta_n}{2}$ and g by $g - \frac{\delta_n}{2}$ and then extending and mollifying these modified functions.

Let $u_n \in C^\infty(\bar{\Omega})$ be classical solutions of

$$\det\left(D^2 u_n - A(x, Du_n)\right) = f_n \quad \text{in } \Omega_n \tag{7.8}$$

and

$$u_n = g_n \quad \text{on } \partial\Omega_n, \tag{7.9}$$

cf. the useful exposition in the introduction of [11] for an overview of different assumptions and corresponding references leading to different regularities. Here, we mention especially the reference [19] mentioned on page 2 of [11] for the smooth case: smooth data imply smoothness of the solution. In the next lemma we estimate u_n on the boundary $\partial\Omega$ by constructing suitable barriers. We have the following plausible lemma which we will also prove rigorously in the following without using any a priori estimates for the solution. Our proof without using the last named type of estimates has the advantage that the lemma also holds when only the sufficient regularity without a priori estimates is available.

Lemma 7.2. *There holds*

$$|u_n - g| \to 0 \tag{7.10}$$

uniformly on $\partial\Omega$ as $n \to \infty$.

Proof. Let us fix $z \in \partial\Omega$ and evaluate g and u_n at z and compare them. Let y be the closest point to z in $\partial\Omega_n$, then $|z-y| \leq \delta_n$ and given $\delta > 0$ we have

$$|g(z) - g_m(y)| \leq |g(z) - g_m(z)| + |g_m(z) - g_m(y)| \leq \delta \tag{7.11}$$

provided m is sufficiently large and also $n = n(m)$ is sufficiently large. Let p be the (not discrete) barrier function from the proof of Lemma 6.1 associated with Ω_n, $z \in \Omega_n$, i.e.

$$p(x) = e^{\lambda|x-\tilde{x}|^2} - e^{\lambda|\tilde{z}-\tilde{x}|^2} \tag{7.12}$$

where $\tilde{x}, \tilde{z}$ are chosen accordingly to the proof of Lemma 6.1 and we may arrange it so that $\tilde{z}$ equals the above specified y, i.e. $y = \tilde{z}$. We define the function

$$b_m^- := p(x) + g_m(y) - C_0|x-y| \tag{7.13}$$

where $C_0 \geq \|g_m\|_{C^1(\bar{\Omega})}$. Clearly, $b_m^- \leq g_m$ in $\bar{\Omega}$ in view of $p \leq 0$ in $\bar{\Omega}$ and we also have that

$$T_\varepsilon[b_m^-] \geq f_m \quad \text{in } \Omega_n \tag{7.14}$$

provided λ is sufficiently large. Hence by the comparison principle we conclude that

$$b_m^- \leq u_m \quad \text{in } \Omega_n. \tag{7.15}$$

Evaluating this inequality in z and retranslation by using the definition of b_m^- leads to

$$g_m(y) - C_m\delta_n \leq u_m(z) \tag{7.16}$$

where $C_m > 0$ is a constant which may depend on m (but not on n). Similarly, using

$$b_m^+(x) := -p(x) + g_m(y) + C_0|x-y| \tag{7.17}$$

as upper barrier function for u_m which is Q-convex on the one-dimensional line

$$\bar{\Omega} \cap \{x_1 = 0\} \tag{7.18}$$

we conclude from the maximum principle in one variable that $b_m^+ \geq u_m$ in $\bar{\Omega}$. Similarly as before we then get

$$u_m(z) \leq g_m(y) + C_m\delta_z. \tag{7.19}$$

Putting this by using the triangle inequality together we conclude that

$$|g(z) - u_m(z)| \leq |g(z) - g_m(y)| + |g_m(y) - u_m(z)| \leq C_m\delta_n + \delta. \tag{7.20}$$

□

The following lemma gives the desired arbitrary good approximation of (1.2) by smooth problems (i.e. with smooth data) with smooth solutions.

Lemma 7.3. *Let f_n, g_n, Ω_n and u_n as before. Let $\varepsilon > 0$ then there is $n \in \mathbb{N}$ such that*

$$|u_n - u| \leq \varepsilon \tag{7.21}$$

in Ω.

Proof. Let $q \leq 0$ be the function from (4.29) and $\alpha, \beta > 0$ suitable constants which will be specified later. We consider the auxiliary function

$$u^- := u + \alpha q - \beta. \tag{7.22}$$

We observe that

$$u^- \leq u - \beta = g - \|g - g_n\|_{L^\infty(\partial\Omega)} \leq g_n \tag{7.23}$$

on $\partial\Omega$ for $\beta = \|g - g_n\|_{L^\infty(\partial\Omega)}$. Let $\phi \in C^2(\Omega)$ and $x_0 \in \Omega$ be a point where

$$u^- - \phi = u - (\phi - \alpha q + \beta) \tag{7.24}$$

attains a maximum. Abbreviating

$$w = \phi - \alpha q + \beta \in C^2(\Omega) \tag{7.25}$$

and using that u is a viscosity subsolution of (1.2) we conclude that

$$T[w] \geq f. \tag{7.26}$$

We would like to show that

$$T[\phi] \geq f_n \tag{7.27}$$

from which we deduce that u^- is a viscosity subsolution of the problem (7.8), (7.9). For it we evaluate $T[\phi]$ more explicitly. As a tool we use the following straightforward and general relation. For positive numbers $a_1, ..., a_n, \varepsilon$ holds when setting

$$\prod_{i=1}^n a_i = z > 0 \tag{7.28}$$

that

$$\prod_{i=1}^n (a_i + \varepsilon) \geq \prod_{i=1}^n a_i + \varepsilon^{n-1} \sum_{i=1}^n a_i \geq z + \varepsilon^{n-1} z^{\frac{1}{n}}. \tag{7.29}$$

For fixed $x \in \bar{\Omega}$ we let $a_1, ..., a_n$ be the eigenvalues of

$$D^2 w(x) - A(x, Dw(x)). \tag{7.30}$$

From the min-max characterization of eigenvalues (given by the Courant-Fisher-Weyl maximum principle) we conclude that the ordered by size eigenvalues $\lambda_1 \leq ... \leq \lambda_n$ of

$$D^2\phi(x) - A(x, D\phi(x)) \tag{7.31}$$

satisfy

$$\lambda_i \geq a_i + \varepsilon \tag{7.32}$$

for some $\varepsilon > 0$ provided λ in the definition of p is sufficiently large. Hence we have

$$T[\phi] \geq f + \varepsilon^{n-1} \min f^{\frac{1}{n}} \tag{7.33}$$

in view of our previous deliberation (7.29). Clearly, we can achieve that (7.27) holds. Since $u \in C^1(\bar{\Omega})$ we may assume w.l.o.g. in the previous argumentation that $\|\phi\|_{C^1(\bar{\Omega})} \leq c(\|u\|_{C^1(\bar{\Omega})})$. Hence the previous mechanism works for λ sufficiently large depending only on f, g and $\|u\|_{C^1(\bar{\Omega})}$ and independently from the choice of α. Hence we see that for n sufficiently large we may choose α sufficiently small and the claim follows since

$$u^- \leq u_n \leq u \tag{7.34}$$

and

$$u - u^- = -\alpha q + \beta \tag{7.35}$$

can be made small for large n. □

8. Convergence properties of the discrete solutions when the scales go to zero

Since u_ε is defined in the computational domain Ω_h and $\Omega_h \subset \Omega$, we extend u_ε to Ω as follows. Given $x \in \Omega \setminus \Omega_h$ we choose $z \in \partial\Omega_h$ as the nearest point in Ω_h to x which is unique because Ω_h is convex and let

$$u_\varepsilon(x) := u_\varepsilon(z) = I_h g(z) \quad \forall x \in \Omega \setminus \Omega_h. \tag{8.1}$$

In the following theorem we prove convergence of the discrete solutions to the solution of the original problem.

Theorem 8.1. *Let Ω be uniformly convex, $f, g \in C(\bar\Omega)$ and $f > 0$ in $\bar\Omega$. Let u be a solution of (1.2) satisfying the assumptions in Remark 7.1. The discrete solutions u_ε of (3.7) and (8.1) converge uniformly to u as $\varepsilon = (h, \delta, \theta) \to 0$ and $\frac{h}{\delta} \to 0$. Here, the constant Λ in the definition of the finite element space, cf. (4.7), depends on ε in the general case. If in addition the sequence of solutions u_n of the approximating problems as constructed in the previous section is uniformly bounded in C^3 then Λ can be chosen uniformly in ε.*

Proof. We first split the domain

$$\begin{aligned}\|u - u_\varepsilon\|_{L^\infty(\Omega)} \le& \|u - u_\varepsilon\|_{L^\infty(\Omega_h)} + \|u - u_\varepsilon\|_{L^\infty(\Omega\setminus\Omega_h)} \\ =& I_1 + I_2.\end{aligned} \tag{8.2}$$

Estimating the first term with the triangle inequality gives

$$I_1 \le \|u - u_n\|_{L^\infty(\Omega_h)} + \|u_n - I_h u_n\|_{L^\infty(\Omega_h)} + \|I_h u_n - u_\varepsilon\|_{L^\infty(\Omega_h)} \tag{8.3}$$

where u_n is the solution of the approximating problem from the previous section and n is assumed to be sufficiently large, and hence

$$\|u - u_n\|_{L^\infty(\Omega_h)} \tag{8.4}$$

can be assumed to be arbitrarily small. In view of the standard interpolation estimate

$$\|u_n - I_h u_n\|_{L^\infty(\Omega_h)} \le c\|u_n\|_{W^{2,\infty}(\Omega)} h^2 \tag{8.5}$$

we may assume that $h = h(n)$ is so small that the norm on the left-hand side is as small as desired as well as that $\Lambda = \Lambda(\varepsilon)$ is sufficiently large. From (8.1) we conclude that for all $x \in \Omega \setminus \Omega_h$ and corresponding $z = z(x) \in \partial\Omega_h$ we have

$$\begin{aligned}|u(x) - u_\varepsilon(x)| =& |u(x) - u_\varepsilon(z)| \\ \le& |u(x) - u(z)| + |u(z) - u_\varepsilon(z)|.\end{aligned} \tag{8.6}$$

Denoting the modulus of continuity of $u \in C(\bar\Omega)$ by τ we have

$$I_2 = \|u - u_\varepsilon\|_{L^\infty(\Omega\setminus\Omega_h)} \le \tau(\operatorname{dist}_H(\Omega, \Omega_h)) + \|u - u_\varepsilon\|_{L^\infty(\Omega_h)}. \tag{8.7}$$

Since $\operatorname{dist}_H(\Omega, \Omega_h) \to 0$ as $h \to 0$ the proof reduces to showing that

$$\|I_h u_n - u_\varepsilon\|_{L^\infty(\Omega_h)} \tag{8.8}$$

can be made arbitrarily small which will be shown in the remaining part of the proof. Note that instead of arguing with the modulus of continuity of u we could have also used the C^0-estimates which we derived in the proof of Lemma 7.3 and the modulus of continuity of the corresponding approximating u_n.

Recall that we have chosen and will choose for the following n sufficiently large. Furthermore, we will assume that $h = h(n)$ is chosen sufficiently small and $\Lambda = \Lambda(\varepsilon)$ sufficiently large.

We use the function $q_h = I_h q$ where

$$q(x) = e^{\lambda|x-\tilde{x}|^2} - R \tag{8.9}$$

with $\tilde{x}$ outside $\bar{\Omega}$ and $R > 0$ so that $q < 0$ in $\bar{\Omega}$. We define the discrete lower barrier as

$$b_\varepsilon^- = u_\varepsilon + \rho q_h \tag{8.10}$$

where $\rho > 0$ so that

$$b_\varepsilon^- \leq g_n \tag{8.11}$$

on $\partial\Omega_h$. W.l.o.g. let us assume that

$$T_\varepsilon[I_h u_n] \leq f_n + \|f - f_n\|_{L^\infty(\bar{\Omega})} + \frac{1}{n}. \tag{8.12}$$

Choosing $\lambda > 0$ sufficiently large we achieve that

$$T_\varepsilon[b_\varepsilon^-] \geq T_\varepsilon[I_h u_n] \tag{8.13}$$

and hence

$$b_\varepsilon^- \leq I_h u_n + O(h). \tag{8.14}$$

A similar argument with $b_\varepsilon^+ := u_\varepsilon - \rho q_h$, $\rho > 0$ suitable, results in $b_\varepsilon^+ \geq I_h u_n - O(h)$.

Clearly, this leads summarized to

$$|I_h u_n - u_\varepsilon| \leq -2\rho q_h + O(h). \tag{8.15}$$

Now, choosing ε (resp. h) small, Λ sufficiently large, and ρ, λ suitable (not depending on h or Λ) we get the desired convergence. This completes the proof. □

References

[1] G. Awanou, Convergence rate of a stable, monotone and consistent scheme for the Monge-Ampère equation, Symmetry, 8 (18), pp. 1-7, (2016).

[2] J.-D. Benàmou, F. Collino, J.-M. Mirebeau, Monotone and consistent discretization of the Monge-Ampère operator, arXiv:1409.6694, (2014).

[3] S. C. Brenner, T. Gudi, M. Neilan, L.-Y. Sung, C^0 penalty methods for the fully nonlinear Monge-Ampère equation, Math. Comp., 80 (276), pp. 1979-1995, (2011).

[4] H. Chen, G. Huang, X.-J. Wang, Convergence rate estimates for Aleksandrov's solution to the Monge-Ampère equation, SIAM J. Numer. Anal., 57, No. 1, 173-191 (2019).

[5] M. G. Crandall, H. Ishii, P.-L. Lions, User's guide to viscosity solutions of second order partial differential equations, Bull. Amer. Math. Soc., 27(1), pp. 1-67, 1992.

[6] E. J. Dean, R. Glowinski, An augmented Lagrangian approach to the numerical solution of the Dirichlet problem for the elliptic Monge-Ampère equation in two dimensions, Electron. Trans. Numer. Anal., 22, pp. 71-96, (2006).

[7] E. J. Dean, R. Glowinski, On the numerical solution of the elliptic Monge-Ampère equation in dimension two: A least-squares approach, Partial Differential Equations, Comput. Meths. Appl. Sci. 16, Springer, Dordrecht, pp. 43-63, (2008).

[8] X. Feng, M. Jensen, Convergent semi-Lagrangian methods for the Monge-Ampère equation on unstructured grids, arXiv:1602.04758v2, (2016).

[9] X. Feng, M. Neilan, Vanishing moment method and moment solutions for fully nonlinear second order partial differential equations, J. Sci. Comput., 38 (1), pp. 74-98, 2009.

[10] X. Feng, M. Neilan, Mixed finite element methods for the fully nonlinear Monge-Ampère euquation based on the vanishing moment method, SIAM J. Numer. Anal., 47(2), pp. 1226-1250, (2009).

[11] A. Figalli, Y.-H. Kim and R. J. McCann, When is multidimensional screening a convex program?, J. Econ. Theory., 146, No. 2, 454-478 (2011)

[12] A. Figalli, Y.-H. Kim and R. J. McCann, Hölder Continuity and Injectivity of Optimal Maps, Arch. Ration. Mech, Anal., 209, pp. 747-795, (2013)

[13] B. Froese, 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), pp. 1692-1714, (2012).

[14] D. Gilbarg, N. Trudinger, Elliptic partial differential equations of second order, volume 224 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, second edition, 1983.

[15] R. Glowinski, Numerical methods for fully nonlinear elliptic equations, Proceedings of the 6th International Congress on Industrial and Applied Mathematics, R. Jeltsch and G. Wanner, eds., ICIAM 07, Invited Lectures, pp. 155-192, (2009).

[16] H. Ishii, P.-L. Lions, Viscosity solutions of fully nonlinear second-order elliptic partial differential equations, J. Diff. Eqs. 83 (1), pp. 26-78, 1990.

[17] J. Jost, Partielle Differentialgleichungen, Springer-Verlag Berlin Heidelberg, 1998.

[18] W. Li and R. H. Nochetto, Optimal pointwise error estimates for two-scale methods for the Monge-Ampère equation, SIAM J. Numer. Anal., Vol 56, No. 3, pp. 915-1941 (2018)

[19] J. Liu, N. S. Trudinger and X.-J. Wang, Interior $C^{2,\alpha}$ regularity for potential functions in optimal transportation, Comm. Partial Differential Equations, 35(1), pp. 165-184, (2010)

[20] X. N. Ma, N. S. Trudinger and X.-J. Wang, Regularity of potential functions of the optimal transportation problem, Arch. Ration. Mech, Anal., 177, pp. 151-183, (2005)

[21] J.-M- Mirebeau, Discretization of the 3D Monge-Ampère operator, between wide stencils and power diagrams, arXiv:1503.00947, 2014.

[22] G. Monge, Mémoire sur la théorie des déblais et de remblais, Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année, pp. 666-704, (1781).

[23] R. H. Nochetto, D. Ntogkas and W. Zhang, Two-scale method for the Monge-Ampère equation: Convergence to the viscosity solution, Math. Comp., (2018)

[24] V. I. Oliker, L. D. Prussner, On the numerical solution of the equation $(\partial^2 z/\partial z^2)(\partial^2 z/\partial y^2) - (\partial^2 z/\partial x \partial y)^2 = f$ and its discretizations, I. Numer. Math., 54(3), pp. 271-293, (1988).

[25] G. D. Phillippis and A. Figalli, Sobolev regularity for Monge-Ampère type equations, SIAM J. Math. Anal., Vol. 45, No. 3, pp. 1812-1824, (2013)

Fachbereich Mathematik, Universität Hamburg, Bundesstrasse 55, 20146 Hamburg, Germany

E-mail address: heiko.kroener@uni-hamburg.de