Acta Numerica (2017), pp. 137–303
doi:10.1017/S0962492917000071
© Cambridge University Press, 2017
Printed in the United Kingdom

Numerical analysis of strongly nonlinear PDEs

Michael Neilan
Department of Mathematics,
University of Pittsburgh, Pittsburgh, PA, USA
E-mail: neilan@pitt.edu

Abner J. Salgado
Department of Mathematics,
The University of Tennessee, Knoxville, TN, USA
E-mail: asalgad1@utk.edu

Wujun Zhang
Department of Mathematics,
Rutgers University, Piscataway, NJ, USA
E-mail: wujun@math.rutgers.edu

We review the construction and analysis of numerical methods for strongly nonlinear PDEs, with an emphasis on convex and non-convex fully nonlinear equations and the convergence to viscosity solutions. We begin by describing a fundamental result in this area which states that stable, consistent and monotone schemes converge as the discretization parameter tends to zero. We review methodologies to construct finite difference, finite element and semi-Lagrangian schemes that satisfy these criteria, and, in addition, discuss some rather novel tools that have paved the way to derive rates of convergence within this framework.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

CONTENTS

1 Introduction 138
2 Elements of the theory of strongly nonlinear elliptic PDEs 149
3 Monotonicity in numerical methods 180
4 Finite element methods for elliptic problems in non-divergence form 210
5 Discretizations of convex second-order elliptic equations 229
6 The Monge–Ampère equation 250
7 Discretizations of non-convex second-order elliptic equations 274
8 Outlook 295
References 296

1. Introduction

Все счастливые семьи похожи
друг на друга, каждая
несчастливая семья
несчастлива по-своему.

Leo Tolstoy

The quote above from Leo Tolstoy (1998), which roughly translates to 'All happy families resemble each other, but each unhappy one is unhappy in its own way', was used by Iserles (2009, Section 8.1) to point out that the numerical approximation of partial differential equations substantially differs from that of ordinary differential equations. The same quote was used in the preface of Holden and Risebro (2015) to say that the theory for nonlinear equations is very different to that for linear problems, and that each nonlinearity needs to be treated in its own way. For these reasons we feel compelled to begin our discussion with the same quote, since we have chosen the unhappiest of all possible families for numerical approximation: fully nonlinear equations.

The goal of this paper is to summarize recent advances and trends in the numerical approximation and theory of strongly second-order nonlinear PDEs with an emphasis on fully nonlinear convex and non-convex PDEs. Such PDEs appear in diverse applications such as weather and climate modelling, stochastic optimal control, determining the initial shape of the universe, optimal reflector design, differential geometry, optimal transport, mathematical finance, image processing and mesh generation. Despite their importance in these application areas, and in contrast to the PDE and

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

solution theory, the study of numerical methods for fully nonlinear problems is still an emerging field in numerical analysis. The reasons for the delayed development are plentiful. As well as the strong nonlinearity, the fundamental difficulties in computing solutions of fully nonlinear problems are the lack of regularity of solutions, the conditional uniqueness of solutions and, most importantly, the notion of viscosity solutions. Similar to weak solutions for PDEs in divergence form, the viscosity solution concept relaxes the pointwise meaning of the PDE, and while doing so, broadens the class of admissible functions in which to seek a solution. However, unlike the weak solution framework, the definition of viscosity solutions is not based on variational principles, but rather comparison principles. While viscosity solution theory and the PDE theory of fully nonlinear problems have made incredible progress during the last 25 years, numerical results for such problems have been slow to catch up due to the pointwise nature and non-variational structure found in the theory.

A breakthrough occurred in Barles and Souganidis (1991), which, roughly speaking, asserts that a consistent, stable and *monotone* numerical method (or general approximation scheme) converges to the viscosity solution as the discretization or regularization parameter tends to zero. The first two conditions in this framework – consistency and stability – are expected; they are the cornerstone of any convergence theory of numerical PDEs and are recognized as the basis of the Lax–Richtmyer equivalence theorem. While arguably less well known, monotonicity of numerical methods is also a long-established area of study; for example, its importance in the context of linear finite difference schemes was realized (at least) 80 years ago (see *e.g.* Gerschgorin 1930, Motzkin and Wasow 1953). On the other hand, the construction of numerical methods that satisfy all three criteria, at least for fully nonlinear problems, is not immediately obvious.

Around the same time, and complementary to the Barles–Souganidis framework, Kuo and Trudinger (1990, 1992) gave a methodology to construct consistent, stable and monotone finite difference schemes for uniformly elliptic fully nonlinear operators. In addition, they showed that such discrete schemes satisfy properties found in viscosity solution theory, including Alexandrov–Bakelman–Pucci (ABP) maximum principles, Harnack inequalities and Hölder estimates. While these results provided a fairly practical guide to computing viscosity solutions, and the theory paved the way for future advances, the fundamental issue of convergence *rates* was explicitly stated as an open and elusive problem.

For the next 15 years, progress in numerical fully nonlinear second-order PDEs was relatively limited and mostly constrained to the theory and convergence rates for *convex* PDEs, in particular the Hamilton–Jacobi–Bellman equation. In this direction, Krylov (1997) introduced the groundbreaking idea of 'shaking the coefficients' to obtain sufficiently smooth subsolutions,

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

which, along with comparison principles, yield rates of convergence even for degenerate problems. These techniques were later refined under various scenarios and assumptions of the problem and discretization (*e.g.* Barles and Jakobsen 2005, Barles and Jakobsen 2002, Krylov 2000), although convexity of the equation always played an essential role in the analysis.

The last ten years have seen an explosion of results for numerical nonlinear PDEs, including a variety of discretization types and convergence results. These include the construction of relatively simple and practical wide-stencil finite difference schemes tailored to specific PDEs (Oberman 2008, Froese and Oberman 2011, Benamou, Froese and Oberman 2014, Froese 2016), and the emergence of Galerkin methods for fully nonlinear problems (Feng and Neilan 2011, Feng and Neilan 2009, Feng and Lewis 2014, Dean and Glowinski 2006*a*, Dean and Glowinski 2006*b*). With regard to rates of convergence, Caffarelli and Souganidis (2008), using intricate regularity results, derived algebraic rates of convergence for finite difference approximations for non-convex PDEs with constant coefficients, and these results were quickly extended by Krylov (2015) and Turanova (2015) to problems with variable coefficients and lower-order terms. On the Galerkin front, Nochetto and Zhang (2017*a*) extended the Kuo–Trudinger theory to finite element methods and derived ABP maximum principles for linear elliptic problems. These results were soon extended in several directions, including rates of convergence for a discrete Monge–Ampère equation (Nochetto and Zhang 2017*b*) and wide-stencil finite difference schemes (Nochetto, Ntogakas and Zhang 2017) and the construction and analysis of finite element methods for non-convex fully nonlinear problems (Salgado and Zhang 2016).

The intention of this survey is to summarize the 25 years of development of fully nonlinear numerical PDEs. Let us describe the organization and the problems we consider in the paper. After setting the notation and stating some instances of fully nonlinear problems, we review some of the basic theory and analysis of elliptic PDEs in Section 2. Here different notions of solutions are introduced and the underlying properties of the solutions and operators are discussed. Besides being of independent interest, these fundamental results motivate both the construction and analysis of the numerical methods. We develop a general framework to compute second-order elliptic problems in Section 3. Basic properties of numerical methods are given, namely consistency, stability and monotonicity, which will lay the groundwork for future developments in the paper. Special attention will be on discrete ABP maximum principles in both the finite difference and finite element setting. Section 4 concerns finite element approximations for linear problems in non-divergence form and with non-smooth coefficients. In Section 5 we combine the ideas of the previous sections and consider finite element and finite difference approximations for fully nonlinear convex PDEs. Alongside the construction and convergence of the schemes, this

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

section will focus on the rates of convergence and the techniques to obtain these results. In Section 6 these results are extended to a particular convex PDE, the Monge–Ampère equation. In Section 7 we discuss recent results of the numerical approximations for fully nonlinear non-convex PDEs. Finally, in Section 8, we give our concluding remarks and state some open problems.

Before starting our discussion let us first briefly outline the derivation of some fully nonlinear PDEs that we focus on in the paper and illustrate their connection with some applications and other areas of mathematics.

1.1. Convex PDEs

This section obtains an instance of the Hamilton–Jacobi–Bellman (HJB) equation, a prototypical fully nonlinear second-order *convex* PDE, and shows how such problems arise in stochastic optimal control problems. In addition, we show below that *every* uniformly elliptic, convex operator with bounded gradient is an implicit HJB problem.

Following Øksendal (2003, Chapter 11) and Fleming and Soner (2006), we consider a stochastic process X_τ governed by the differential equation

$$\begin{cases} \mathrm{d}X_\tau = \ell(X_\tau, \tau, \alpha_\tau)\,\mathrm{d}\tau + \sigma(X_\tau, \tau, \alpha_\tau)\,\mathrm{d}W_\tau \quad \tau > 0, \\ X_0 = x_0 \in \mathbb{R}^d. \end{cases} \tag{1.1}$$

Here, W_τ is a Brownian motion of dimension d, σ is a $d \times d$ matrix-valued function, $\alpha_\tau \in \mathcal{A}$ is a (Markov) control and $\mathcal{A}$ is the control space. Problem (1.1) describes a dynamical system driven by additive white noise with diffusion coefficient (or volatility) σ and non-stochastic drift ℓ. Under appropriate smoothness and growth conditions on ℓ and σ, and for a fixed $\alpha(\cdot) \in \mathcal{A}$, there is a path-wise unique solution to (1.1). Associated with problem (1.1) is the family of stochastic processes that satisfy (1.1) but with initial time $t \geq 0$:

$$\begin{cases} \mathrm{d}X^{x,t}_\tau = \ell(X^{x,t}_\tau, \tau, \alpha_\tau)\,\mathrm{d}\tau + \sigma(X^{x,t}_\tau, \tau, \alpha_\tau)\,\mathrm{d}W_\tau \quad \tau > t, \\ X^{x,t}_t = x \in \mathbb{R}^d. \end{cases} \tag{1.2}$$

Now let $\Omega \subset \mathbb{R}^d$ be an open bounded domain, and set $Q = \Omega \times (0, T)$ for some $T \in (0, \infty]$. Let $\widehat{T}$ denote the first exit time for the process $X^{x,t}_\tau$ satisfying (1.2), that is,

$$\widehat{T} = \widehat{T}^{x,t} = \inf\{\tau > t; (X^{x,t}_\tau, \tau) \notin Q\}.$$

We then define the performance function (or cost functional)

$$J(x, t, \alpha) = \mathbb{E}_{x,t}\left\{ \int_t^{\widehat{T}} f(X^{x,t}_\tau, \tau, \alpha_\tau)\,\mathrm{d}\tau + \chi_{\widehat{T}<\infty}\psi(X^{x,t}_{\widehat{T}}, \widehat{T}) \right\}$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

over all $\alpha \in \mathcal{A}$ under the constraint (1.2). Here, $f : \mathbb{R}^d \times \mathbb{R} \times \mathcal{A} \to \mathbb{R}$ is the profit rate function, $\psi : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ is the bequest function, $\chi_{\widehat{T}<\infty}$ is the indicator function, which equals one if $\widehat{T} < \infty$ and zero otherwise, and $\mathbb{E}_{x,t}$ represents the expectation conditional on $X_t^{x,t} = x$. We then consider the problem of finding the value function $u : \bar{Q} \to \mathbb{R}$ and optimal control $\alpha^* \in \mathcal{A}$ such that

$$u(x,t) = \sup_{\alpha\in\mathcal{A}} J(x,t,\alpha) = J(x,t,\alpha^*).$$

Let us now give a heuristic derivation of the Hamilton–Jacobi–Bellman equation based on Bellman's principle of optimality and Itô's lemma. First, Bellman's principle of optimality states that (Bellman 2010):

Whatever the initial state $\{X_s^{x,t}\}_{s<\tau}$ and initial decision $\{\alpha_s\}_{s<\tau}$ are, the remaining controls $\{\alpha_s\}_{s>\tau}$ must constitute an optimal policy with regard to the state $X_\tau^{x,t}$ resulting from the initial decision.

More precisely, this principle equates to

$$u(x,t) = \sup_{\alpha\in\mathcal{A}} \mathbb{E}_{x,t}\left\{u(X_{t+\delta t}^{x,t}, t+\delta t) + \int_t^{t+\delta t} f(X_\tau^{x,t},\tau,\alpha_\tau)\,\mathrm{d}\tau\right\}. \tag{1.3}$$

Next, recalling that $X_t^{x,t} = x$, by Itô's lemma we obtain

$$\begin{aligned} u(X_{t+\delta t}^{x,t}, t+\delta t) = u(x,t) &+ \int_t^{t+\delta t} \frac{\partial u}{\partial t}\,\mathrm{d}\tau + \int_t^{t+\delta t} Du\cdot\,\mathrm{d}X_\tau^{x,t} \\ &+ \frac{1}{2}\int_t^{t+\delta t} \mathrm{d}X_\tau^{x,t}\cdot D^2u\,\mathrm{d}X_\tau^{x,t}. \end{aligned}$$

Since the state variable $X_\tau^{x,t}$ is governed by the stochastic equation (1.2), and by using the identity $\mathrm{d}W_\tau \otimes \mathrm{d}W_\tau = I\,\mathrm{d}\tau$, we obtain

$$\begin{aligned} &u(X_{t+\delta t}^{x,t}, t+\delta t) - u(x,t) \\ &\quad = \int_t^{t+\delta t}\left(\frac{\partial u}{\partial t} + Du\cdot\ell + \frac{1}{2}(\sigma\sigma^\intercal) : D^2u\right)\mathrm{d}\tau + \int_t^{t+\delta t}(\sigma^\intercal Du)\cdot\,\mathrm{d}W_\tau. \end{aligned}$$

Therefore, from principle of optimality (1.3),

$$\sup_{\alpha\in\mathcal{A}} \mathbb{E}_{x,t}\left\{u(X_{t+\delta t}^{x,t}, t+\delta t) - u(x,t) + \int_t^{t+\delta t} f(X_\tau^{x,t},\tau,\alpha)\,\mathrm{d}\tau\right\} = 0,$$

and Itô's formula above, we obtain

$$\sup_{\alpha\in\mathcal{A}} \mathbb{E}_{x,t}\left\{\int_t^{t+\delta t}\left(\frac{\partial u}{\partial t} + Du\cdot\ell + \frac{1}{2}(\sigma\sigma^\intercal) : D^2u + f\right)\mathrm{d}\tau\right\} = 0,$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

where we used that

$$\mathbb{E}_{x,t}\left\{\int_t^{t+\delta t}(\sigma^{\mathsf{T}}Du)\cdot \mathrm{d}W_\tau\right\}=0.$$

Dividing by δt and formally taking the limit $\delta t\to 0$, we then obtain a deterministic equation, namely the Hamilton–Jacobi–Bellman equation

$$\frac{\partial u}{\partial t}(x,t)+\sup_{\alpha\in\mathcal{A}}(\mathcal{L}^\alpha u(x,t)+f(x,t,\alpha))=0,$$

where

$$\mathcal{L}^\alpha u(x,t)=Du(x,t)\cdot\ell(x,t,\alpha)+\frac{1}{2}\sigma(x,t,\alpha)\sigma(x,t,\alpha)^{\mathsf{T}}:D^2u(x,t).$$

To derive the equation, we assumed that the value function $u(x,t)$ has continuous second-order derivatives. However, the PDE theory reveals that the solution of the HJB equation in general does not satisfy this regularity assumption. To justify the derivation above, the concept of viscosity solutions was introduced in Crandall and Lions (1983) for first-order Hamilton–Jacobi equations and later generalized to second-order Hamilton–Jacobi–Bellman equations (Lions 1982, 1983). Viscosity solutions – an essential concept in the PDE theory – will also play an important role in this paper.

1.2. Non-convex PDEs

The Isaacs equation, a prototypical non-convex PDE, describes zero sum stochastic games with two players. Each player has one control and they have opposite objectives. The first player chooses a control to maximize the expected payoff, whereas the second player chooses a control to minimize it. Stochastic game theory has wide applications in engineering and mathematical finance. Here, we follow the ideas in Fleming and Souganidis (1989) to derive the equation.

The dynamics of the stochastic differential game we investigate are given by the controlled stochastic differential equation

$$\begin{cases}\mathrm{d}X_\tau^{x,t}=\ell(X_\tau^{x,t},\tau,\alpha_\tau,\beta_\tau)\,\mathrm{d}\tau+\sigma(X_\tau^{x,t},\tau,\alpha_\tau,\beta_\tau)\,\mathrm{d}W_\tau \quad \tau>t,\\ X_t^{x,t}=x\in\mathbb{R}^d,\end{cases}\tag{1.4}$$

and the expected payoff is

$$J(x,t,\alpha,\beta)=\mathbb{E}_{x,t}\left\{\int_t^{\widehat{T}}f(X_\tau^{x,t},\tau,\alpha_\tau,\beta_\tau)\,\mathrm{d}\tau+\chi_{\widehat{T}<\infty}\psi(X_{\widehat{T}}^{x,t},\widehat{T})\right\},$$

where $X_\tau^{x,t}$ satisfies the differential equation (1.4) for $\tau>t$ and has initial condition $X_t^{x,t}=x$.

The Isaacs equation can be derived in a similar fashion to the HJB equation. If, at time t, the first player chooses a control α to maximize the

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

expected payoff J, and the second player, based on the decision of the first player, chooses the control β to minimize it, then we set

$$u^+(x,t) = \inf_{\beta\in\mathcal{B}} \sup_{\alpha\in\mathcal{A}} \mathbb{E}_{x,t}\{J(x,t,\alpha,\beta)\}$$

and call this the upper value function. On the other hand, if the second player makes the decision first and the first player reacts accordingly, then we set the lower value function as

$$u^-(x,t) = \sup_{\alpha\in\mathcal{A}} \inf_{\beta\in\mathcal{B}} \mathbb{E}_{x,t}\{J(x,t,\alpha,\beta)\}.$$

By the principle of optimality we have

$$u^+(x,t) = \inf_{\beta\in\mathcal{B}} \sup_{\alpha\in\mathcal{A}} \mathbb{E}_{x,t}\left\{u^+(X^{x,t}_{t+\delta t}, t+\delta t) + \int_t^{t+\delta t} f(X^{x,t}_\tau, \tau, \alpha, \beta)\,\mathrm{d}\tau\right\}.$$

Using a derivation similar to the HJB equation, we obtain that for all $(x,t) \in Q$ the upper value function must satisfy

$$\frac{\partial u^+}{\partial t}(x,t) + H^+(x,t,Du^+,D^2u^+) = 0,$$

where

$$H^+(x,t,\mathbf{p},M) = \inf_{\beta\in\mathcal{B}} \sup_{\alpha\in\mathcal{A}} \left\{\frac{1}{2}\sigma\sigma^\intercal : M + \ell(x,t,\alpha,\beta)\cdot\mathbf{p} + f(x,t,\alpha,\beta)\right\}.$$

Similarly we obtain, for $(x,t) \in Q$,

$$\frac{\partial u^-}{\partial t}(x,t) + H^-(x,t,Du^-,D^2u^-) = 0,$$

with

$$H^-(x,t,\mathbf{p},M) = \sup_{\alpha\in\mathcal{A}} \inf_{\beta\in\mathcal{B}} \left\{\frac{1}{2}\sigma\sigma^\intercal : M + \ell(x,t,\alpha,\beta)\cdot\mathbf{p} + f(x,t,\alpha,\beta)\right\}.$$

We finally comment that if the Isaacs' condition holds, that is, we have that for all $x,t,\mathbf{p},M$,

$$H^+(x,t,\mathbf{p},M) = H^-(x,t,\mathbf{p},M),$$

then we can conclude that $u^+(x,t) = u^-(x,t)$ for all $(x,t) \in Q$. In this case, we say that an optimal policy exists.

1.3. Characterizations of elliptic PDEs

Let us show that there is no loss in generality in confining our considerations to these two equations by following the construction proposed by Evans (1980, Lemma 2.2). Let $\Omega \subset \mathbb{R}^d$ be a bounded domain and let $F : \Omega \times \mathbb{S}^d \to \mathbb{R}$ be continuously differentiable, non-decreasing with respect to its

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

second argument and with bounded gradient. Here, $\mathbb{S}^d$ denotes the space of symmetric $d \times d$ matrices. We comment that, as we will see below (see Definition 2.10), these conditions guarantee that the operator F is elliptic. We have the following representation: for every $x \in \Omega$ and $M \in \mathbb{S}^d$, we have

$$F(x,M) = \inf_{\beta\in\mathbb{S}^d} \sup_{\alpha\in\mathbb{S}^d} \left[\int_0^1 \frac{\partial F}{\partial M}(x,(1-t)\beta + t\alpha) : (M-\beta) + F(x,\beta) \right]. \quad (1.5)$$

To see this, let IS denote the right-hand side of (1.5), and note that by setting $\beta = M$ we obtain

$$\begin{aligned} IS &\leq \sup_{\alpha\in\mathbb{S}^d} \left[\int_0^1 \frac{\partial F}{\partial M}(x,(1-t)M + t\alpha) : (M-M) + F(x,M) \right] \\ &= F(x,M). \end{aligned}$$

On the other hand, setting $\alpha = M$ we obtain that

$$\begin{aligned} &\int_0^1 \frac{\partial F}{\partial M}(x,(1-t)\beta + tM) : (M-\beta) + F(x,\beta) \\ &\qquad \leq \sup_{\alpha\in\mathbb{S}^d} \left[\int_0^1 \frac{\partial F}{\partial M}(x,(1-t)\beta + t\alpha) : (M-\beta) + F(x,\beta) \right]. \end{aligned}$$

Under the given assumptions on F, the left-hand side of this inequality can be rewritten as

$$\begin{aligned} &\int_0^1 \frac{\partial F}{\partial M}(x,(1-t)\beta + tM) : (M-\beta) + F(x,\beta) \\ &\quad = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} F(x,(1-t)\beta + tM) + F(x,\beta) = F(x,M) - F(x,\beta) + F(x,\beta), \end{aligned}$$

and consequently

$$F(x,M) \leq \sup_{\alpha\in\mathbb{S}^d} \left[\int_0^1 \frac{\partial F}{\partial M}(x,(1-t)\beta + t\alpha) : (M-\beta) + F(x,\beta) \right],$$

or $F(x,M) \leq IS$.

With representation (1.5) at hand we define

$$\begin{aligned} A^{\alpha,\beta}(x) &= \int_0^1 \frac{\partial F}{\partial M}(x,(1-t)\beta + t\alpha), \\ f^{\alpha,\beta}(x) &= A^{\alpha,\beta}(x) : \beta - F(x,\beta), \\ \mathcal{A} &= \mathcal{B} = \mathbb{S}^d, \end{aligned}$$

and note that, since $F(x,\cdot)$ is non-decreasing, $A^{\alpha,\beta}(x) \geq 0$ for all $\alpha \in \mathcal{A}$, $\beta \in \mathcal{B}$ and $x \in \Omega$. In conclusion, $F(x,\cdot)$ can be represented as the inf-sup

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

of a family of affine maps, that is,

$$F(x, M) = \inf_{\beta \in \mathcal{B}} \sup_{\alpha \in \mathcal{A}} \big[A^{\alpha,\beta}(x) : M - f^{\alpha,\beta}(x) \big].$$

If we assume in addition that F is convex in its second argument, then we have

$$F(x, M) - F(x, \alpha) \geq \frac{\partial F}{\partial M}(x, \alpha) : (M - \alpha), \quad \text{for all } \alpha \in \mathbb{S}^d.$$

Setting $\beta = \alpha$ in (1.5) yields

$$F(x, M) \leq \sup_{\alpha \in \mathbb{S}^d} \left[\frac{\partial F}{\partial M}(x, \alpha) : (M - \alpha) + F(x, \alpha) \right],$$

so that

$$F(x, M) = \sup_{\alpha \in \mathcal{A}} \big[A^{\alpha}(x) : M - f^{\alpha}(x) \big],$$

with $\mathcal{A} = \mathbb{S}^d$ and

$$A^{\alpha}(x) = \frac{\partial F}{\partial M}(x, \alpha) \geq 0, \quad f^{\alpha}(x) = \frac{\partial F}{\partial M}(x, \alpha) : \alpha - F(x, \alpha).$$

We conclude by remarking that more general types of dependence can also be reduced to a similar inf-sup form; see, for instance, Kuo and Trudinger (1992, Section 2.1), Katsoulakis (1995) and Koike and Święch (2013).

Nomenclature

A_z	adjacent set of node z
$\{\!\{\cdot\}\!\}$	average operator over faces F
B_r	ball with radius r centred at the origin
$\delta_{y,h}$	centred difference operator in the direction y with step h
$\partial\Omega$	boundary of Ω
$\delta^-_{y,h}$	backward difference operator in the direction y with step h
$A \Subset B$	the set A is compactly contained in B
C	a generic positive constant
$\mathcal{C}^-(v)$	lower contact set of v
$C^k(\bar{\Omega})$	space of k-times continuously differentiable functions defined in $\bar{\Omega}$, $k \in \mathbb{N}$
conv A	convex hull of the set A
$C_0^\infty(\Omega)$	infinitely differentiable functions whose support is compactly contained in Ω
$\mathcal{C}_h^-(v_h)$	discrete lower contact set of v_h

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

$\det M$	determinant of the matrix M
I_h	abstract canonical interpolant
d	spatial dimension
$\Gamma_h(u_h)$	discrete convex envelope of u_h
$\mathrm{dist}(x, A)$	distance between the point x and the set A
$D\cdot$	divergence operator
Ω	domain in d-dimensional space $\mathbb{R}^d$
Ω_h^B	abstract discrete boundary domain
Ω_h	abstract discrete domain
Ω_h^I	abstract discrete interior domain
$\cdot$	inner/dot product in $\mathbb{R}^d$
X_h	abstract approximation space with parameter $h > 0$
K	elements in a mesh $\mathcal{T}_h$
$\lvert\mathbf{q}\rvert_M$	norm in $\mathbb{R}^d$ induced by the matrix $M > 0$, that is, $(\mathbf{q} \cdot M\mathbf{q})^{1/2}$
$\Gamma(u)$	convex envelope of function u
F	faces in a mesh $\mathcal{T}_h$
$\mathcal{F}_h$	set of faces
$\mathcal{F}_h^I$	interior faces
$\delta_{y,h}^+$	forward difference operator in the direction y with step h
Δ_h	finite element Laplacian
X_h^l	finite element space of piecewise linear continuous functions
X_h^{cd}	H^2-conforming finite element space
X_h^{dg}	discontinuous finite element space
X_h^{cg}	H^1-conforming finite element space with polynomials of degree $k \in \mathbb{N}$
$X_{0,h}^l$	functions in X_h^l that vanish at the boundary
$\lvert\cdot\rvert$	ℓ^2-norm/Fröbenius norm of a matrix
D	gradient operator
$\{\phi_i\}$	normalized hat basis functions of X_h^l
$\{\tilde{\phi}_i\}$	canonical hat basis functions of X_h^l
D^2u	Hessian of the function u

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

$C^{k,\alpha}(\bar{\Omega})$	space of functions in $C^k(\bar{\Omega})$ whose kth derivative is Hölder-continuous with exponent $\alpha \in (0,1]$
I_h^{ep}	elliptic projection onto X_h^l
I_h^{fd}	canonical nodal interpolant onto X_h^{fd}
I_h^{fe}	canonical nodal interpolant onto X_h^l
u_τ^-	inf-convolution of the function u
$\int_\Omega v$	integral of v over the domain Ω
$[[\cdot]]$	jump operator on faces F
$L^p(\Omega)$	space of Lebesgue-integrable functions with exponent $p \in [1,\infty]$
$\lvert\cdot\rvert_\infty$	ℓ^∞-norm
$LSC(\Omega)$	space of lower semi-continuous functions
$\mathbb{S}^d$	space of symmetric $d \times d$ matrices
$\lvert U\rvert$	measure of the set U
$v^- = -\min\{v,0\}$	negative part of v
X_h^{fd}	space of nodal functions
$\mathbb{SO}^d$	special orthogonal group
$\partial v/\partial u$	partial derivative of the function v with respect to its argument u
$\{\omega_i\}_{i=1}^N$	arbitrary (not necessarily simplicial) partition of $\bar{\Omega}$
ω_z	patch of vertex z
$\mathbb{P}_k$	space of polynomials of degree $k \geq 0$
$v^+ = \max\{v,0\}$	positive part of v
$\delta_{y,h}^2$	second difference operator in the direction y with step h
$W^{s,p}(\Omega)$	Sobolev spaces of functions whose kth derivatives belong to $L^p(\Omega)$, $s \geq 0$
$W_0^{s,p}(\Omega)$	closure of $C_0^\infty(\Omega)$ in $W^{s,p}(\Omega)$, $s \geq 0$
$H_0^2(\Omega)$	Sobolev space $W_0^{s,2}(\Omega)$
$H^s(\Omega)$	Sobolev space $W^{s,2}(\Omega)$
$W^{-s,p'}(\Omega)$	dual of $W_0^{s,p}(\Omega)$ $s \geq 0$
u_h	discrete solution of PDEs
$\mathbb{S}_+^d$	symmetric non-negative definite $d \times d$ matrices
S	stencil
m	stencil size

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

$\partial v(x)$	subdifferential of the function v at the point x
u_τ^+	sup-convolution of the function u
$\bar{\partial} u(x)$	superdifferential of u at x
$D_T \cdot$	tangential divergence operator
D_T	tangential gradient operator
Δ_T	tangential Laplacian
$\operatorname{tr} M$	trace of the matrix M
$\mathbb{S}_1^d$	set of matrices M in $\mathbb{S}_+^d$ for which $\operatorname{tr} M = 1$
$\mathbb{D}_1^d$	set of diagonal positive definite matrices with unit trace
Tv_h	translates of the function v_h
$M^\intercal$	transpose of the matrix M
$\mathcal{T}_h$	triangulation
$USC(\Omega)$	space of upper semi-continuous functions
z	vertices/nodes of mesh $\mathcal{T}_h$
$VMO(\Omega)$	space of functions with vanishing mean oscillation

2. Elements of the theory of strongly nonlinear elliptic PDEs

In order to get an idea of how to discretize strongly nonlinear partial differential equations, we must first understand their underlying structure and the main ideas that are at the basis of their theory and analysis. In this introductory section we collect all the relevant information that will later serve as a guide in the construction and analysis of numerical schemes. We will describe the fundamental properties that define an elliptic equation, even in the case of strong nonlinearities, and thereby define various suitable notions of solutions and their properties. We will provide existence and non-existence results, as well as a review of the available regularity results. While it is not our intention to provide a thorough exposition of the theory, which can be found in textbooks such as those of Evans (2010), Han and Lin (2011), Gilbarg and Trudinger (2001), Krylov (1996, 2008) and Maugeri, Palagachev and Softova (2000), we believe it is fundamental to understand this if one wishes to provide a rigorous analysis of approximation schemes.

2.1. Two defining consequences of ellipticity

We begin our description by providing two fundamental properties that lie at the heart of much of the theory for elliptic PDEs, namely energy considerations and comparison principles. We will see how these give rise to various concepts of solutions and how much of the existence and regularity theory stems from these two simple ideas.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

To make matters precise, we set $\Omega \subset \mathbb{R}^d$ with $d \geq 1$ to be a bounded domain with Lipschitz boundary. If further smoothness of the domain becomes necessary we will specify this at every stage. For simplicity, and because these ideas are better motivated in this case, let us consider *the Laplacian* which, for a function $u \in C^2(\Omega)$, is defined by

$$\Delta u = \sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2}.$$

The first fundamental property that can be observed for this operator is a *maximum principle*.

Theorem 2.1 (maximum principle). Let $u \in C^2(\Omega) \cap C(\bar{\Omega})$ be such that $\Delta u \geq 0$. Then

$$\sup_{x \in \bar{\Omega}} u(x) = \sup_{x \in \partial\Omega} u(x).$$

While we will not provide a detailed proof here, we wish to provide some intuition into this fact. Namely, if we assume that a strict global maximum is attained at an interior point $z \in \Omega$, then elementary considerations from calculus will give us that

$$Du(z) = 0, \quad D^2u(z) < 0,$$

where Du denotes the gradient and D^2u the Hessian of u, respectively. Since it can be easily seen that $\Delta u = \operatorname{tr} D^2u$, a contradiction ensues.

An important consequence of this result is the following *comparison principle*.

Corollary 2.2 (comparison principle). Let $u, v \in C^2(\Omega) \cap C(\bar{\Omega})$ be such that $u \leq v$ in $\partial\Omega$ and $\Delta u \geq \Delta v$ in Ω. Then $u \leq v$ in Ω.

This result easily follows by setting $w = u - v$ and observing that $w \leq 0$ on $\partial\Omega$ and $\Delta w \geq 0$ in Ω. An application of the maximum principle then allows us to conclude the result.

The comparison principle is one of the fundamental properties of an elliptic operator, namely that an ordering on the boundary and a (reverse) ordering of the operators implies an order of the underlying functions. Throughout this survey the application of a similar principle will be a recurring feature.

Having understood comparison principles we now proceed to describe *energy* considerations. Consider the equation

$$\Delta u = f \tag{2.1}$$

in Ω and multiply it by a sufficiently smooth function φ that vanishes on

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

$\partial\Omega$. An application of Green's identity reveals that

$$\int_\Omega Du \cdot D\varphi = -\int_\Omega f\varphi, \tag{2.2}$$

which we immediately recognize as the Euler–Lagrange equation for the minimization of the *energy* functional

$$J(v) = \int_\Omega \left(\frac{1}{2}|Dv|^2 + fv\right)$$

subject to the condition $v = u$ on $\partial\Omega$.

It is important to notice that, as opposed to (2.1), identity (2.2) only requires the existence of square-integrable first derivatives. Note also that the second variation of J is non-negative:

$$\partial^2 J(v)[w_1, w_2] = \int_\Omega Dw_1 \cdot Dw_2, \quad \partial^2 J(v)[w, w] = \int_\Omega |Dw|^2 \geq 0, \tag{2.3}$$

which shows its positivity in a certain sense.

On the basis of this observation, we now introduce our first definition of ellipticity (Gilbarg and Trudinger 2001, Chapter 3). In order to do so, in what follows we let $\mathbb{S}^d$ denote the space of symmetric $d \times d$ matrices. We endow $\mathbb{S}^d$ with the usual partial order

$$M, N \in \mathbb{S}^d : \quad M \leq N \iff \boldsymbol{\xi} \cdot M\boldsymbol{\xi} \leq \boldsymbol{\xi} \cdot N\boldsymbol{\xi}, \quad \text{for all } \boldsymbol{\xi} \in \mathbb{R}^d.$$

We denote the identity matrix by $I \in \mathbb{S}^d$.

Definition 2.3 (elliptic operator in divergence form). Let $A : \Omega \to \mathbb{S}^d$. We say that the operator

$$Lu(x) = -D \cdot (A(x)Du(x)) \tag{2.4}$$

is *elliptic* at $x \in \Omega$ if $0 < \lambda(x)I \leq A(x) \leq \Lambda(x)I$, *strictly* elliptic if $\lambda(x) \geq \lambda_0 > 0$ for all $x \in \Omega$, and *uniformly* elliptic if $\Lambda(x)/\lambda(x)$ is bounded in Ω.

Example 2.4 (lower-order terms). The concept of elliptic operators in divergence form can be extended to operators having lower-order terms. For instance, consider the operator

$$\tilde{L}u(x) = -D \cdot (A(x)Du(x) + \mathbf{b}(x)u(x)) + \mathbf{c}(x) \cdot Du(x) + e(x)u(x),$$

where A is as in Definition 2.3, and the functions $\mathbf{b}, \mathbf{c} : \Omega \to \mathbb{R}^d$ and $e : \Omega \to \mathbb{R}$ are assumed to be measurable.

Example 2.5 (quasi-linear operators). Given a differentiable vector-valued function $\Omega \times \mathbb{R} \times \mathbb{R}^d \ni (x, z, \mathbf{p}) \mapsto \mathbf{a}(x, z, \mathbf{p}) \in \mathbb{R}^d$ and a scalar function $\Omega \times \mathbb{R} \times \mathbb{R}^d \ni (x, z, \mathbf{p}) \mapsto b(x, z, \mathbf{p}) \in \mathbb{R}$, we consider the *quasi-linear* operator

$$Qu(x) = -D \cdot \mathbf{a}(x, u, Du) + b(x, u, Du)$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

defined for $u \in C^2(\Omega)$. We say that this operator is *variational* if it is the Euler–Lagrange operator of the energy functional

$$\int_\Omega E(x, u, Du),$$

that is, $\mathbf{a}(x, z, \mathbf{p}) = D_{\mathbf{p}} E(x, z, \mathbf{p})$ and $b(x, z, \mathbf{p}) = D_z E(x, z, \mathbf{p})$. Following Definition 2.3 we realize that the ellipticity of Q is equivalent to the strict convexity of E with respect to the $\mathbf{p}$ variables. This immediately hints at the fact that tools from calculus of variations will be essential in the study of equations with this type of operator. Examples of quasi-linear operators can be given by choosing appropriate energies E. For instance, setting

$$E = E(\mathbf{p}) = (1 + |\mathbf{p}|^2)^{s/2}$$

(Gilbarg and Trudinger 2001, Chapter 10) for $s > 1$, we obtain a family of uniformly elliptic quasi-linear operators.

On the other hand, many problems cannot be cast in this form. The prototypical example is that given by the operator

$$\mathcal{L}u(x) = A(x) : D^2 u(x), \tag{2.5}$$

where, for $M, N \in \mathbb{S}^d$, $M : N$ denotes the Fröbenius inner product:

$$M : N = \sum_{i,j=1}^{d} M_{i,j} N_{i,j}.$$

The Fröbenius norm of a matrix M will be denoted by $|M| := \sqrt{M : M}$.

If A is sufficiently smooth, the operator (2.5) can be recast in divergence form and $-\mathcal{L}$ can be understood as an elliptic operator in the sense of Definition 2.3. However, this is not always possible, and consequently we must extend the notion of ellipticity to non-divergence-form operators (Gilbarg and Trudinger 2001, Chapter 3).

Definition 2.6 (non-divergence elliptic operator). We say that the operator (2.5) is *elliptic* at $x \in \Omega$ if $0 < \lambda(x) I \leq A(x) \leq \Lambda(x) I$, *strictly* elliptic if $\lambda(x) \geq \lambda_0 > 0$ for all $x \in \Omega$, and *uniformly* elliptic if $\Lambda(x)/\lambda(x)$ is bounded in Ω.

The reader is encouraged to verify that, for an operator that is strictly elliptic in the sense of Definition 2.6, variants of Theorem 2.1 and Corollary 2.2 hold.

Example 2.7 (lower-order terms). As in the divergence-form case, the notion of elliptic operators extend to those with lower-order terms, for example,

$$\tilde{\mathcal{L}}u(x) = A(x) : D^2 u(x) + \mathbf{b}(x) \cdot Du(x) + c(x) u(x),$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

where A is as in Definition 2.6, and the lower-order coefficients $\mathbf{b}$ and c are (vector- and scalar-valued) functions defined on Ω.

Example 2.8 (linear operators in divergence form). Consider the operator L of Definition 2.3 and assume that the coefficient matrix A is differentiable. One can then rewrite $Lu(x)$ as

$$Lu(x) = -A(x) : D^2u(x) - (D \cdot A(x)) \cdot Du(x),$$

where the divergence operator acts on A column-wise. Consequently, the operator $-L$ is of the form $\tilde{\mathcal{L}}$ of Example 2.7.

Example 2.9 (quasi-linear operators in non-divergence form). For a function $u \in C^2(\Omega)$ we define the *quasi-linear* operator

$$\mathcal{Q}u(x) = A(x, u, Du) : D^2u(x) + b(x, u, Du)$$

where $\Omega \times \mathbb{R} \times \mathbb{R}^d \ni (x, z, \mathbf{p}) \mapsto A(x, z, \mathbf{p}) \in \mathbb{S}^d$ and $\Omega \times \mathbb{R} \times \mathbb{R}^d \ni (x, z, \mathbf{p}) \mapsto b(x, z, \mathbf{p}) \in \mathbb{R}$. We say that $\mathcal{Q}$ is elliptic at the function u if $A(x, u, Du)$ satisfies the positivity conditions of Definition 2.6.

The previous definitions and examples entailed linear and quasi-linear operators. While, by linearization, one could extend Definitions 2.3 and 2.6 to more general nonlinear problems, we shall instead give a general definition of ellipticity, one that preserves the fundamental concept of comparison (Crandall, Ishii and Lions 1992, Caffarelli and Cabré 1995).

Definition 2.10 (elliptic operator). Let $F \in C(\Omega \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{S}^d)$. We say that F is *elliptic* in Ω if F satisfies the following *monotonicity condition*. If $r, s \in \mathbb{R}$ and $M, N \in \mathbb{S}^d$ with $r \geq s$ and $M \leq N$ then

$$F(x, r, \mathbf{p}, M) \leq F(x, s, \mathbf{p}, N).$$

We will say, moreover, that F is *uniformly elliptic* if there are constants $0 < \lambda \leq \Lambda$ such that for all $M \in \mathbb{S}^d$ and $r \geq s$ we have

$$\lambda|N| \leq F(x, r, \mathbf{p}, M + N) - F(x, s, \mathbf{p}, M) \leq \Lambda|N|, \quad \text{for all } N \geq 0.$$

Example 2.11 (linear and quasi-linear equations). As a first example, let us show that the linear operator in non-divergence form $\mathcal{L}$ of (2.5) is elliptic in the sense of Definition 2.10. By doing so and following the considerations of the examples previously given, we see that all the other cases also fit into this framework. Define

$$F(x, r, \mathbf{p}, M) = \operatorname{tr}(A(x)M).$$

Since, for symmetric matrices, $A : B = \operatorname{tr}(AB)$, we see that $\mathcal{L}u(x) = F(x, u(x), Du(x), D^2u(x))$. Moreover, the positivity of A implies the monotonicity of F.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

We now present several examples of *fully nonlinear* equations that fit into Definition 2.10.

Example 2.12 (Hamilton–Jacobi–Bellman operator). Let $\mathcal{A}$ be any compact set and assume that for every $\alpha \in \mathcal{A}$ we are given a uniformly elliptic linear operator

$$\mathcal{L}^\alpha u(x) = A^\alpha(x) : D^2 u(x)$$

and a function $f^\alpha \in C(\Omega)$. Define

$$F(x, r, \mathbf{p}, M) = \sup_{\alpha \in \mathcal{A}} \big[\operatorname{tr}(A^\alpha(x) M) - f^\alpha(x) \big].$$

Note immediately that

$$F(x, u(x), Du(x), D^2 u(x)) = \sup_{\alpha \in \mathcal{A}} \big[\mathcal{L}^\alpha u(x) - f^\alpha(x) \big].$$

Moreover, since for every $\alpha \in \mathcal{A}$, $x \in \Omega$ and $M, N \in \mathbb{S}^d$ we have that $A^\alpha(x) M \leq A^\alpha(x) N$ whenever $M \leq N$, we immediately conclude that the operator F is monotone and thus elliptic in the sense of Definition 2.10. A similar argument shows that F is uniformly elliptic whenever the family of linear operators $\{\mathcal{L}^\alpha\}_{\alpha \in \mathcal{A}}$ is uniformly elliptic. More importantly, we note that the function F is convex with respect to M.

Example 2.13 (Isaacs operator). The previous example can be generalized as follows. Assume now that we have two index sets $\mathcal{A}$ and $\mathcal{B}$ and, for each $(\alpha, \beta) \in \mathcal{A} \times \mathcal{B}$, we have a uniformly elliptic linear operator

$$\mathcal{L}^{\alpha,\beta} u(x) = A^{\alpha,\beta}(x) : D^2 u(x).$$

Define

$$F(x, r, \mathbf{p}, M) = \inf_{\beta \in \mathcal{B}} \sup_{\alpha \in \mathcal{A}} \big[\operatorname{tr}(A^{\alpha,\beta}(x) M) - f^{\alpha,\beta}(x) \big],$$

and note that

$$F(x, u(x), Du(x), D^2 u(x)) = \inf_{\beta \in \mathcal{B}} \sup_{\alpha \in \mathcal{A}} \big[\mathcal{L}^{\alpha,\beta} u(x) - f^{\alpha,\beta}(x) \big].$$

One more time, the uniform ellipticity of the operators $\mathcal{L}^{\alpha,\beta}$ yields the uniform ellipticity of F. Note that F is neither convex nor concave with respect to M.

Example 2.14 (Monge–Ampère operator). As a final example, consider the operator

$$F(x, r, \mathbf{p}, M) = \det M - f(x).$$

Note that, in general, this operator does not satisfy the monotonicity condition of Definition 2.10. However, if we restrict it to *positive definite* matrices,

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

then this operator is elliptic. Consequently, for a positive f and a *strictly convex* function $u \in C^2(\Omega)$, we define the Monge–Ampère operator as

$$F(x, u(x), Du(x), D^2u(x)) = \det D^2u(x) - f(x).$$

With Definition 2.10 at hand, we turn our attention to boundary value problems for elliptic operators. In other words, for an elliptic operator F we consider the following problem: find $u : \bar{\Omega} \to \mathbb{R}$ such that

$$F(x, u, Du, D^2u) = 0 \ \text{ in } \Omega, \quad u = g \ \text{ on } \partial\Omega. \tag{2.6}$$

The sense in which (2.6) is satisfied will give rise to the various existing concepts of solutions.

2.2. Classical solutions

The first, and obvious, notion of solution is when identity (2.6) is understood in a pointwise sense. This gives rise to so-called classical solutions.

Definition 2.15 (classical solution). Let $F \in C(\Omega \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{S}^d)$. Then the function $u \in C^2(\Omega) \cap C(\bar{\Omega})$ is said to be a *classical* solution of (2.6) if this identity holds for every $x \in \bar{\Omega}$.

An immediate consequence of ellipticity is that classical solutions are unique.

Theorem 2.16 (uniqueness). Let F be elliptic in the sense of Definition 2.10. If F is strictly decreasing in the r variable or uniformly elliptic, then problem (2.6) cannot have more than one classical solution.

Proof. Let us prove this result under the assumption that the map F is strictly decreasing in the r variable. The remaining case can be found, for instance, in Gilbarg and Trudinger (2001, Corollary 17.2). Assume that u and v are classical solutions to (2.6) and set $w = u - v$. Note that $w = 0$ on $\partial\Omega$, and that if w attains a (positive) maximum at $x_0 \in \Omega$, then $Dw(x_0) = 0$ and $D^2w(x_0) \leq 0$. Therefore, if $u(x_0) > v(x_0)$, ellipticity and the fact that the map is strictly decreasing imply

$$0 = F(x_0, u(x_0), Du(x_0), D^2u(x_0)) < F(x_0, v(x_0), Dv(x_0), D^2v(x_0)),$$

which is a contradiction. Similarly, the function w cannot attain a negative minimum in Ω, and consequently $w \equiv 0$. □

As an example, we mention that Theorem 2.16 holds for the operator $\tilde{\mathcal{L}}$ of Example 2.7 whenever the zero order coefficient $c \leq 0$.

In the linear case of Example 2.11, the Dirichlet problem (2.6) reads as follows: find $u \in C^2(\Omega) \cap C(\bar{\Omega})$ such that

$$\mathcal{L}u = f \ \text{ in } \Omega, \quad u = g \ \text{ on } \partial\Omega. \tag{2.7}$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

In this case, the existence of classical solutions is guaranteed by what is known as Schauder estimates which, simply put, boil down to freezing the coefficients and a continuity argument; see Gilbarg and Trudinger (2001, Theorems 6.13–6.14).

Theorem 2.17 (existence). Let Ω satisfy an exterior sphere condition at every boundary point. Assume the operator (2.5) is strictly elliptic in the sense of Definition 2.6. If $g \in C(\partial\Omega)$ and, for some $\alpha \in (0,1)$, f and the coefficients of $\mathcal{L}$ are bounded and belong to $C^{\alpha}(\Omega)$, then the Dirichlet problem (2.7) has a unique classical solution $u \in C^{2,\alpha}(\Omega) \cap C(\bar{\Omega})$. If, in addition, we assume that $\partial\Omega \in C^{2,\alpha}$, that f and the coefficients of $\mathcal{L}$ belong to $C^{\alpha}(\bar{\Omega})$; and $g \in C^{2,\alpha}(\bar{\Omega})$, then $u \in C^{2,\alpha}(\bar{\Omega})$ and

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

where the constant C is independent of u, f and g.

At this point, the reader may wonder if Hölder continuity is indeed necessary for these results. Example 2.23 below will show us that this is the case.

If more regularity is assumed on the domain and problem data, it can be shown that the (unique) classical solution is also more regular (Gilbarg and Trudinger 2001, Theorem 6.19).

Theorem 2.18 (regularity). In the setting of Theorem 2.17, assume additionally that, for some $k \geq 0$ we have $\partial\Omega \in C^{k+2,\alpha}$, $g \in C^{k+2,\alpha}(\bar{\Omega})$ and that f and the coefficients of $\mathcal{L}$ belong to $C^{k,\alpha}(\bar{\Omega})$. Then $u \in C^{k+2,\alpha}(\bar{\Omega})$.

While Theorems 2.17 and 2.18 provide a satisfactory and conclusive answer for a linear operator with smooth coefficients, it does not cover rough coefficients or nonlinear problems, in which a classical solution may not exist. This is why we must depart from classical solutions and consider *weakened* or *generalized* concepts of solutions.

2.3. Weak (variational) solutions

We now turn our attention to the case of divergence-form operators as in Definition 2.3 and consider, for this particular operator, the Dirichlet problem (2.6). The natural solution concept in this case is called weak solution, and follows from an integration-by-parts argument, and an integral identity similar to (2.2).

Definition 2.19 (weak solutions of linear equations). A function $u \in H^1(\Omega)$ is said to be a *weak* solution of the Dirichlet problem

$$Lu = f \ \text{ in } \Omega, \quad u = g \ \text{ on } \partial\Omega \tag{2.8}$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

if $u - g \in H^1_0(\Omega)$ and, for every $\varphi \in H^1_0(\Omega)$, we have

$$\int_\Omega (ADu) \cdot D\varphi = \int_\Omega f\varphi. \tag{2.9}$$

Existence and uniqueness follow from the classical Lax–Milgram lemma or, more generally, from so-called inf-sup conditions.

Theorem 2.20 (existence and uniqueness). Let Ω be bounded, and let L be uniformly elliptic in the sense of Definition 2.3 and such that its coefficients belong to $L^\infty(\Omega)$. If $f \in H^{-1}(\Omega)$ and $g \in H^1(\Omega)$, then problem (2.8) has a unique weak solution $u \in H^1(\Omega)$.

Again, under additional smoothness assumptions on the domain and problem data, one can assert further differentiability of the solution. This is the content of the following result (Gilbarg and Trudinger 2001, Grisvard 2011).

Theorem 2.21 (regularity). Assume, in addition to the conditions of Theorem 2.20, that $\partial\Omega \in C^2$ or that Ω is convex. If $A \in C^{0,1}(\bar\Omega, \mathbb{S}^d)$, $f \in L^2(\Omega)$ and $g \in H^2(\Omega)$, then we have $u \in H^2(\Omega) \cap H^1(\Omega)$ and

$$\|u\|_{H^2(\Omega)} \le C\big(\|u\|_{L^2(\Omega)} + \|f\|_{L^2(\Omega)} + \|g\|_{H^2(\Omega)}\big),$$

where the constant C is independent of u, f and g.

We also wish to mention the remarkable result by E. De Giorgi concerning the Hölder regularity of weak solutions.

Theorem 2.22 (De Giorgi I). Let $u \in H^1(\Omega)$ be a weak solution to (2.9) with $g = 0$, $f \in L^q(\Omega)$ with $q > d/2$. Then there exists $\alpha \in (0,1)$ for which $u \in C^\alpha_{loc}(\Omega)$. If, in addition, $q > d$ and $A \in C^\beta(\bar\Omega, \mathbb{S}^d)$, with $\beta = 1 - d/q$, then $u \in C^{1,\beta}_{loc}(\Omega)$.

In light of the second part of the previous result, it is natural to ask whether $f \in L^\infty(\Omega)$ with appropriate assumptions on the boundary data g and the coefficients of L would yield that $Du \in C^1_{loc}(\Omega)$. The following example shows that this is generally false (Han and Lin 2011, Section 3.4).

Example 2.23 (second derivatives are not continuous). For $R < 1$ let $\Omega = \{x \in \mathbb{R}^d : |x| < R\}$ and consider

$$f(x) = \begin{cases} 0 & x = 0, \\ \dfrac{x_2^2 - x_1^2}{2|x|^2}\left(\dfrac{d+2}{\sqrt{-\ln|x|}} + \dfrac{1}{2(-\ln|x|)^{3/2}}\right) & x \neq 0. \end{cases}$$

Note that $f \in C(\bar\Omega)$ and that the function

$$u(x) = (x_1^2 - x_2^2)\sqrt{-\ln|x|} \in C(\bar\Omega) \cap C^\infty(\bar\Omega \setminus \{0\})$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

satisfies $\Delta u = f$ with boundary conditions

$$g = \sqrt{-\ln R}(x_1^2 - x_2^2).$$

However, this function cannot be a classical solution since

$$\lim_{|x|\to 0} \frac{\partial^2 u(x)}{\partial x_1^2} = \infty$$

so that $u \notin C^2(\Omega)$. In fact, although the problem has a weak solution, it does not have a classical one. This example also shows that, in the classical solution theory given in Theorem 2.17, mere continuity of the data is not sufficient, thus justifying the need for Hölder continuity.

Let us now focus our attention on the quasi-linear operator Q of Example 2.5 and consider the Dirichlet problem

$$Qu = f \ \text{ in } \Omega, \quad u = g \ \text{ on } \partial\Omega. \tag{2.10}$$

The definition of weak solution is as follows.

Definition 2.24 (weak solutions of quasi-linear equations). A function $u \in W^{1,p}(\Omega)$ $(1 < p < \infty)$ is called a *weak* solution of (2.10) if $u - g \in W_0^{1,p}(\Omega)$ and

$$\int_\Omega (\mathbf{a}(x, u, Du) \cdot Dv + b(x, u, Du)v) = \int_\Omega fv$$

for all $v \in W_0^{1,p}(\Omega)$.

Note that the equation that defines weak solutions to (2.10) are the Euler–Lagrange equations of the functional

$$I(u) = \int_\Omega (E(x, u, Du) - fu)$$

over the set of functions $v \in W^{1,p}(\Omega)$ such that $u - g \in W_0^{1,p}(\Omega)$. Consequently, the existence of weak solutions is tightly bound with the calculus of variations.

Theorem 2.25 (existence and uniqueness). Assume that there is a $p \in (1, \infty)$ for which the function E satisfies the coercivity condition: there are constants $C_1 > 0$, $C_2 \geq 0$ such that, for every $x \in \Omega$, $z \in \mathbb{R}$ and $\mathbf{p} \in \mathbb{R}^d$, we have

$$E(x, z, \mathbf{p}) \geq C_1|\mathbf{p}|^p - C_2.$$

Assume, in addition, that E is convex in the $\mathbf{p}$ variable. Then, for $f \in L^{p'}(\Omega)$, the functional I has a minimizer $u \in W^{1,p}(\Omega)$ such that $u - g \in W_0^{1,p}(\Omega)$. Finally, if E does not depend on z and is uniformly convex, then this minimizer is unique.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

With this theorem at hand it can be readily shown that, in this setting, the (unique) minimizer u of I is a weak solution of (2.10) in the sense of Definition 2.24.

We can also establish, under additional assumptions on E, further differentiability of minimizers. To shorten the exposition we confine ourselves to the case where E is independent of x and z, it is coercive with $p = 2$, satisfies the *growth condition*

$$|D_{\mathbf{p}}E(\mathbf{p})| \leq C(|\mathbf{p}| + 1), \quad \text{for all } \mathbf{p} \in \mathbb{R}^d, \tag{2.11}$$

and

$$|D^2E(\mathbf{p})| \leq C, \quad \text{for all } \mathbf{p} \in \mathbb{R}^d. \tag{2.12}$$

With these additional assumptions we have the following regularity result (Evans 2010, Theorem 8.3.1).

Theorem 2.26 (regularity). In the setting of Theorem 2.25 assume, in addition, that E depends only on $\mathbf{p}$ and satisfies (2.11) and (2.12). If $g = 0$, $f \in L^2(\Omega)$ and $\partial\Omega \in C^2$, we have that $u \in H^2(\Omega)$ with the estimate

$$\|u\|_{H^2(\Omega)} \leq C\|f\|_{L^2(\Omega)}.$$

What is more interesting and remarkable is that the results of De Giorgi presented in Theorem 2.22 can be extended to this case as well.

Theorem 2.27 (De Giorgi II). Let $u \in W^{1,p}(\Omega)$ be a minimizer of I. If E satisfies the growth and monotonicity conditions

$$|D_{\mathbf{p}}E(x, z, \mathbf{p})| \leq C_1(1 + |\mathbf{p}|^{p-1}), \quad D_{\mathbf{p}}E(x, z, \mathbf{p}).\mathbf{p} \geq C_2|\mathbf{p}|^p - C_3,$$

then there exists $\alpha \in (0, 1)$ for which $u \in C^{\alpha}_{loc}(\Omega)$.

Under suitable assumptions, local Hölder continuity of the gradients of the minimizers can also be established. For further regularity results for quasilinear problems the reader is referred, for instance, to Mingione (2006).

While, in this setting, we have a sufficiently rich theory, it only applies to divergence-form operators. Below, in Section 2.5, we will describe the right generalization of the notion of solutions for more general problems.

2.4. Strong solutions

We now describe a solution concept that, in a sense, lies in between classical and weak solutions, and that can also be applied to non-divergence-form operators such as (2.5) and that of Example 2.9. These solutions are called strong.

Definition 2.28 (strong solutions). The function $u \in W^{2,p}(\Omega)$ $(1 < p < \infty)$ is a *strong* solution of the boundary value problem (2.6) if the

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

equation and boundary conditions hold almost everywhere in Ω and $\partial\Omega$, respectively.

We immediately remark that every classical solution is a strong solution. Moreover, an integration-by-parts and density argument shows that a sufficiently regular weak solution (see Theorems 2.21 and 2.26) is also a strong solution. Therefore, strong solutions for the divergence-form equations (2.8) and (2.10) can be obtained from regularity considerations.

Let us now turn our attention to the non-divergence-form problem (2.7) and study the existence of strong solutions. In this case we have the following result.

Theorem 2.29 (existence). Let Ω be a $C^{1,1}$ domain and the coefficients of the operator L belong to $C(\bar{\Omega})$. If $f \in L^p(\Omega)$ and $g \in W^{2,p}(\Omega)$ $(1 < p < \infty)$, then the Dirichlet problem (2.7) has a unique strong solution $u \in W^{2,p}(\Omega)$, and moreover

$$\|u\|_{W^{2,p}(\Omega)} \leq C\big(\|f\|_{L^p(\Omega)} + \|g\|_{W^{2,p}(\Omega)}\big),$$

where C is independent of u, f and g but depends on $\|A\|_{C(\bar{\Omega},\mathbb{S}^d)}$, the dimension d and the exponent p.

We must comment on the technique of proof for this result. First, for $A = I$ and $p = 2$, this follows from the regularity result of Theorem 2.21. An interpolation result, in conjunction with the celebrated Calderón–Zygmund decomposition technique (Calderón and Zygmund 1952), yields the result for any p. Using the continuity of A, the result can be extended to a general $\mathcal{L}$.

Remark 2.30 (Hölder regularity). Let us briefly describe the results of Krylov and Safonov; see Gilbarg and Trudinger (2001, Section 9.8) and Theorem 2.65 below. To do so, we assume that $f \in L^d(\Omega)$, $g \in C^\beta(\bar{\Omega})$ for some $\beta \in (0,1)$ and $\partial\Omega$ satisfies a uniform exterior cone condition. Then, given $\omega \Subset \Omega$, there is a constant $\alpha \in (0,1)$ such that

$$|u|_{C^\alpha(\omega)} \leq C.$$

The constants α and C depend, in particular, on the dimension d and the ratio Λ/λ that defines the ellipticity of $\mathcal{L}$. A natural question to ask is whether a similar estimate for the gradient Du (possibly under stricter smoothness assumptions) is possible. A result by Nirenberg (see Theorem 2.57) showed that this is the case for $d = 2$. For higher dimensions, however, this turns out to be false. Safonov (1987) showed that in $B_1 \subset \mathbb{R}^3$, the unit ball, for every $\alpha \in (0,1]$ there are:

- a bounded function $v \in C^\infty(\mathbb{R}^3 \setminus \{0\})$,
- a constant $\nu \in (0,1)$,

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

- a family $\{A_\varepsilon\}_{\varepsilon>0} \subset C^\infty(\bar{B}_1, \mathbb{S}^d)$ such that the associated non-divergence operators $\mathcal{L}_\varepsilon$ are uniformly elliptic with $\Lambda/\lambda = 1/\nu^2$.

With these objects at hand, he showed that the solution to the problem

$$\mathcal{L}_\varepsilon u_\varepsilon = 0 \ \text{ in } B_1, \quad u_\varepsilon = v \ \text{ on } \partial B_1,$$

satisfies $u_\varepsilon \in C^\infty(\bar{B}_1)$, $\|u_\varepsilon\|_{L^\infty(B_1)} = 1$ but

$$\lim_{\varepsilon \downarrow 0} |u_\varepsilon|_{C^\alpha(B_{1/2})} = \infty.$$

From this it immediately follows that Hölder estimates on the derivatives are not possible.

Let us point out now that, in Theorem 2.29, the assumption that $A \in C(\bar{\Omega}, \mathbb{S}^d)$ cannot generally be weakened. The following example is due to Pucci.

Example 2.31 (non-uniqueness). Let us show, following Maugeri *et al.* (2000, Section 1.1), that for $d \geq 3$ there is a bounded measurable matrix A such that problem (2.7) with $f = 0$ and $g = 0$ has more than one strong solution in $H^2(\Omega) \cap H_0^1(\Omega)$. Let Ω be the unit ball of $\mathbb{R}^d$, and define

$$A(x) = I + b\frac{xx^\intercal}{|x|^2}, \quad b = \frac{d-2+\lambda}{1-\lambda}, \quad \max\{2 - d/2, 0\} < \lambda < 1.$$

Obviously

$$|\boldsymbol{\xi}|^2 \leq \boldsymbol{\xi} \cdot A\boldsymbol{\xi} \leq (1+b)|\boldsymbol{\xi}|^2,$$

so that A is bounded and the associated operators $\mathcal{L}$ are uniformly elliptic. Define $u(x) = |x|^\lambda - 1$ and note that

$$D^2u(x) = \lambda(\lambda-2)|x|^{\lambda-4}xx^\intercal + \lambda|x|^{\lambda-2}I,$$

which, since $\lambda > 2 - d/2$, shows that $u \in H^2(\Omega) \cap H_0^1(\Omega)$. Moreover, due to the choice of b, we have

$$\mathcal{L}u(x) = A : D^2u(x) = \lambda|x|^{\lambda-2}\big[(1+b)\lambda + d - b - 2\big] = 0.$$

Since this will be important in subsequent developments, we now focus on conditions weaker than continuity that allow for the existence and uniqueness of a strong solution for (2.7).

2.4.1. The Cordes condition

Since, as Example 2.31 shows, mere boundedness of the coefficients in the operator of (2.7) does not suffice to ensure uniqueness of strong solutions, here we study the so-called Cordes condition for linear operators in non-divergence form. The idea behind it and the theory that follows is to reformulate the operator in such a way that the result is 'close' to one in

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

divergence form, in particular, the Poisson equation. This reformulation allows us to apply classical tools in functional analysis to study the existence, uniqueness and *a priori* estimates for problem (2.5).

To motivate and derive this condition, consider the following problem: given $x \in \Omega$, find $\gamma(x) \in \mathbb{R}$ that minimizes the quadratic function

$$\tau \mapsto |\tau A(x) - I|^2.$$

Simple arguments show that the minimum is attained at

$$\gamma(x) = \frac{\operatorname{tr} A(x)}{|A(x)|^2} \quad \text{and} \quad |\gamma(x)A(x) - I|^2 = d - \frac{(\operatorname{tr} A(x))^2}{|A(x)|^2}. \tag{2.13}$$

In particular, this simple calculation shows that

$$\begin{aligned} |\gamma(x)\mathcal{L}v(x) - \Delta v(x)|^2 &= |(\gamma(x)A(x) - I) : D^2 v(x)|^2 \\ &\leq \left(d - \frac{\operatorname{tr} A(x)^2}{|A(x)|^2}\right)|D^2 v(x)|^2. \end{aligned} \tag{2.14}$$

The Cordes condition ensures that the multiplicative constant on the right-hand side of (2.14) is less than one.

Definition 2.32 (Cordes condition). A positive definite matrix $A \in L^\infty(\Omega, \mathbb{S}^d)$ satisfies the *Cordes condition* provided that there exists an $\epsilon \in (0, 1]$ such that

$$\frac{|A|^2}{(\operatorname{tr} A)^2} \leq \frac{1}{d - 1 + \epsilon} \quad \text{a.e. } \Omega. \tag{2.15}$$

Note that the Cordes condition ensures that there exists $\gamma > 0$ such that, for all $v \in H^2(\Omega)$, we have

$$\|\gamma\mathcal{L}v - \Delta v\|_{L^2(\Omega)} \leq \sqrt{1 - \epsilon}\|D^2 v\|_{L^2(\Omega)}. \tag{2.16}$$

Remark 2.33 (the Cordes condition in spectral terms). Since, by assumption, for a.e. $x \in \Omega$ we have that $A(x) \in \mathbb{S}^d$ and that it is positive definite, it is diagonalizable and all its eigenvalues $\{\lambda_i\}_{i=1}^d = \{\lambda_i(x)\}_{i=1}^d$ are positive. Using the well-known identities

$$|A|^2 = \sum_{i=1}^d \lambda_i^2, \quad \operatorname{tr} A = \sum_{i=1}^d \lambda_i \quad \text{and} \quad \left(\sum_{i=1}^d \lambda_i\right)^2 \leq d \sum_{i=1}^d \lambda_i^2,$$

condition (2.15) can be recast in terms of the eigenvalues of A as follows:

$$\frac{1}{d} \leq \frac{\sum_{i=1}^d \lambda_i^2}{\left(\sum_{i=1}^d \lambda_i\right)^2} \leq \frac{1}{d - 1 + \epsilon} \quad \text{a.e. } \Omega.$$

In other words, (2.15) is an anisotropy condition on A that becomes more stringent in higher dimensions.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

The considerations in Remark 2.4.1 show that the Cordes condition is always satisfied in two dimensions with $\epsilon = \inf_{x\in\Omega} 2\lambda_1\lambda_2/(\lambda_1^2+\lambda_2^2) \in (0,1]$. On the other hand, there exist symmetric positive definite matrices in three dimensions (and higher) that do not satisfy (2.15).

Example 2.34 (three dimensions). Consider the matrix

$$A = \begin{pmatrix} 1 & 0 & c \\ 0 & 1 & b \\ c & b & 4 \end{pmatrix}$$

with $b^2 + c^2 < 4$ so that $\det(A) = 4 - (b^2 + c^2) > 0$. Sylvester's criterion ensures that the matrix is positive definite, and a straightforward calculation shows that

$$\frac{|A|^2}{(\operatorname{tr} A)^2} = \frac{18 + 2(b^2+c^2)}{6^2} \geq \frac{1}{2}.$$

Thus A does not satisfy the Cordes condition.

Example 2.35 (the example of Pucci). As another example consider the matrix of Example 2.31 for $d = 3$. Note first of all that in this case we have

$$\frac{1}{2} < \lambda < 1, \quad b = \frac{3 - 2 + \lambda}{1 - \lambda} > 3.$$

Simple calculations then yield

$$|A|^2 = d + 2b + b^2,$$
$$(\operatorname{tr} A)^2 = (d+b)^2,$$

and therefore

$$|A|^2 - \frac{1}{d-1}(\operatorname{tr} A)^2 = d + 2b + b^2 - \frac{(d+b)^2}{d-1} > 0$$

Thus one concludes that A does not satisfy the Cordes condition.

The Cordes condition is a key assumption to establish the well-posedness of the elliptic problem (2.5) with discontinuous coefficients. Another crucial ingredient is the Miranda–Talenti estimate, which is summarized in the next lemma.

Lemma 2.36 (Miranda–Talenti estimate). Let $\Omega \subset \mathbb{R}^d$ be a bounded convex domain. Then, for any $v \in H^2(\Omega) \cap H_0^1(\Omega)$,

$$|v|_{H^2(\Omega)} \leq \|\Delta v\|_{L^2(\Omega)}. \tag{2.17}$$

While this result can be understood as a regularity estimate in the spirit of Theorem 2.21, we remark that it can be obtained without appealing to this theory; we refer the reader to Maugeri *et al.* (2000, Lemma 1.2.2) for

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

a proof. Moreover, while the aforementioned regularity results yield that, for functions in $H^2(\Omega) \cap H_0^1(\Omega)$, the norm $v \mapsto \|\Delta v\|_{L^2(\Omega)}$ is equivalent to the $H^2(\Omega)$-norm, the important feature of estimate (2.17) is that the equivalence constant is exactly one on convex domains.

Remark 2.37 (polygonal domains). In two dimensions the Miranda–Talenti estimate (2.17) holds for a polygonal domain Ω *without the convexity assumption*. Indeed, assuming that $u \in C^\infty(\bar{\Omega})$, we have

$$|D^2 u|^2 = |\Delta u|^2 + 2(|\partial_{12} u|^2 - \partial_{11} u \partial_{22} u),$$

where we explicitly used that we are in two dimensions. In addition, integration by parts and some algebraic manipulations show that

$$\int_\Omega (|\partial_{12} u|^2 - \partial_{11} u \partial_{22} u) = \int_{\partial\Omega} (-\partial_{\tau\tau} u \partial_n u + \partial_{n\tau} u \partial_\tau u)$$

(see Ciarlet 2002, equation (1.2.9)), where $\boldsymbol{\tau}$ is the unit tangential vector along the boundary $\partial\Omega$, ∂_τ is the derivative in its direction and ∂_n denotes the normal derivative. Now, if $u = 0$ on $\partial\Omega$ then we have that $\partial_\tau u = 0$, so that the second term on the right-hand side of this expression vanishes. If, in addition, $\partial\Omega$ is polygonal, this also implies that $\partial_{\tau\tau} u = 0$, which allows us to obtain (2.17). By density, the same result holds for every $u \in H^2(\Omega) \cap H_0^1(\Omega)$.

Identities (2.13) and (2.17) motivate the introduction of the bilinear form

$$a(\cdot,\cdot) : (H^2(\Omega) \cap H_0^1(\Omega))^2 \ni (v, w) \mapsto a(v, w) = \int_\Omega \gamma \mathcal{L} v \Delta w \in \mathbb{R}.$$

The properties of a are as follows.

Lemma 2.38 (properties of a). Assume that the coefficient A of the operator $\mathcal{L}$ satisfies the Cordes condition (2.15). If Ω is convex, then the bilinear form a is bounded and coercive on $H^2(\Omega) \cap H_0^1(\Omega)$.

Proof. Since γ is bounded, the continuity immediately follows.

If the Cordes condition (2.15) is satisfied and Ω is convex, then from the Miranda–Talenti estimate (2.17) and the Cauchy–Schwarz inequality we obtain

$$\begin{aligned} a(v, v) &= \|\Delta v\|_{L^2(\Omega)}^2 + \int_\Omega (\gamma \mathcal{L} v - \Delta v) \Delta v \\ &\geq \|\Delta v\|_{L^2(\Omega)}^2 - \sqrt{1-\epsilon} |v|_{H^2(\Omega)} \|\Delta v\|_{L^2(\Omega)} \\ &\geq (1 - \sqrt{1-\epsilon}) \|\Delta v\|_{L^2(\Omega)}^2. \end{aligned} \tag{2.18}$$

In conclusion, a is coercive on $H^2(\Omega) \cap H_0^1(\Omega)$. □

The coercivity estimate of Lemma 2.38 allows us to show the existence and uniqueness of strong solutions under the Cordes condition.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Theorem 2.39 (existence and uniqueness). Assume that the coefficient A of the operator $\mathcal{L}$ satisfies the Cordes condition (2.15). If Ω is convex, then the Dirichlet problem (2.7) with $f \in L^2(\Omega)$ and $g = 0$ has a unique strong solution $u \in H^2(\Omega) \cap H_0^1(\Omega)$. Moreover, we have

$$\|u\|_{H^2(\Omega)} \leq C \frac{\|\gamma\|_{L^\infty(\Omega)}}{1 - \sqrt{1-\epsilon}} \|f\|_{L^2(\Omega)},$$

where the constant C is independent of u and f.

Proof. From Lemma 2.38 and the Lax–Milgram lemma, there exists a unique $u \in H^2(\Omega) \cap H_0^1(\Omega)$ satisfying

$$a(u, v) = \int_\Omega \gamma f \Delta v, \quad \text{for all } v \in H^2(\Omega) \cap H_0^1(\Omega).$$

Since $u \in H^2(\Omega)$ and the Laplace operator $\Delta : H^2(\Omega) \cap H_0^1(\Omega) \to L^2(\Omega)$ is surjective on convex domains, standard arguments show that u satisfies $\mathcal{L}u = f$ almost everywhere, that is, it is a strong solution to the elliptic problem (2.7). The coercivity condition (2.18) also implies the *a priori* estimate

$$\|u\|_{H^2(\Omega)} \leq C\|\Delta u\|_{L^2(\Omega)} \leq C \frac{\|\gamma\|_{L^\infty(\Omega)}}{1 - \sqrt{1-\epsilon}} \|f\|_{L^2(\Omega)}. \qquad \square$$

Remark 2.40 (inf-sup conditions). Since the Laplace operator is surjective from $H^2(\Omega) \cap H_0^1(\Omega)$ to $L^2(\Omega)$ on convex domains, the above arguments show that the inf-sup condition

$$\sup_{w \in L^2(\Omega)\setminus\{0\}} \frac{\int_\Omega \gamma \mathcal{L} v w}{\|w\|_{L^2(\Omega)}} \geq (1 - \sqrt{1-\epsilon}) \|\Delta v\|_{L^2(\Omega)}, \quad \text{for all } v \in H^2(\Omega) \cap H_0^1(\Omega)$$

is satisfied. One can then appeal to the Babuška–Brezzi theorem to deduce the existence of strong solutions to (2.5). To our knowledge the use of this inf-sup condition for the numerical approximation has yet to be investigated.

Remark 2.41 (the case $p \neq 2$). It is possible to show (Maugeri *et al.* 2000, Theorem 1.2.3) that, if Ω is convex and A satisfies the Cordes condition, there are $1 < p_l < 2 < p_r < \infty$ such that if $p \in (p_l, p_r)$, $f \in L^p(\Omega)$ and $g = 0$, then problem (2.7) has a unique strong solution $u \in W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)$. We also have an *a priori* estimate in which the constant now depends on p.

Remark 2.42 (strong solutions under other conditions). It is possible to obtain the existence and uniqueness of strong solutions for problem (2.7) under other assumptions. Let us discuss two of them.

- Assuming that $A \in W^{1,d}(\Omega, \mathbb{S}^d)$ one can write the operator L in non-divergence form and extend the theory of weak solutions, described in

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Section 2.3, to coefficients in this class. What is remarkable is that, in Example 2.31, for every $\varepsilon > 0$ the parameter λ can be chosen so that $A \in W^{1,d-\varepsilon}(\Omega, \mathbb{S}^d)$, thus showing that $A \in W^{1,d-\varepsilon}(\Omega, \mathbb{S}^d)$ is not sufficient for uniqueness. On the other hand, if $A \in W^{1,d+\varepsilon}(\Omega, \mathbb{S}^d)$, for some $\varepsilon > 0$, then $A \in C^{0,\alpha}(\bar{\Omega}, \mathbb{S}^d)$ and, consequently, the classical Schauder theory applies (see Theorem 2.17).

- In essence, the case of uniformly continuous coefficients boils down to realizing that, locally, their oscillation in the $L^\infty(\Omega)$-norm is small, and so they can be considered a constant. These ideas have been extended (see Maugeri *et al.* 2000, Chapter 2) to the case of a coefficient $A \in VMO(\Omega, \mathbb{S}^d)$, thus showing that this is a sufficient condition to obtain strong solutions. Since $W^{1,d}(\Omega)$ is a proper subset of $VMO(\Omega)$ this result truly extends the case of Sobolev coefficients detailed above.

2.5. Viscosity solutions

At this point we wish to introduce one final notion of solution, the one that will be suited to the study of fully nonlinear equations. This is that of a viscosity solution. The reader may recall that the notion of weak solutions, introduced in Section 2.3, was based on an integration-by-parts argument (2.2) and the positivity (2.3) of the resulting operators. While this proved sufficient for linear and quasi-linear operators in divergence form, different arguments are necessary for fully nonlinear operators such as those of Examples 2.12–2.14. The fundamental property that will be used to define solutions in this case, as in Corollary 2.2, will be a comparison principle.

2.5.1. Definition and first properties

Let us begin by motivating the definition following Katzourakis (2015). Let F be an elliptic operator in the sense of Definition 2.10 and let $u \in C^2(\Omega)$ be a classical solution to

$$F(x, u, Du, D^2u) = 0 \quad \text{in } \Omega. \tag{2.19}$$

Let $x_0 \in \Omega$ and assume that there is a smooth function $\varphi \in C^2(\Omega)$ that can *touch from above* the graph of u at x_0. More precisely, we assume that

$$u(x) \leq \varphi(x) \quad \text{for all } x \in \Omega, \quad u(x_0) = \varphi(x_0).$$

These conditions imply that the function $u - \varphi$ has a local maximum at x_0 and, consequently,

$$D(u - \varphi)(x_0) = 0, \quad D^2(u - \varphi)(x_0) \leq 0.$$

Since the operator F is assumed to be elliptic we obtain

$$\begin{aligned} 0 &= F(x_0, u(x_0), Du(x_0), D^2u(x_0)) = F(x_0, \varphi(x_0), D\varphi(x_0), D^2u(x_0)) \\ &\leq F(x_0, \varphi(x_0), D\varphi(x_0), D^2\varphi(x_0)). \end{aligned}$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Similar considerations will give us that if $\psi \in C^2(\Omega)$ *touches from below* the graph of u at x_0, we would obtain

$$F(x_0, \psi(x_0), D\psi(x_0), D^2\psi(x_0)) \leq 0.$$

Finally we note that it is possible to reach the same conclusions if we replace the equality in (2.19) with a corresponding inequality. These considerations motivate the following definition.

Definition 2.43 (viscosity solution). Let F be elliptic in the sense of Definition 2.10. We say that the function $u \in C(\Omega)$ is:

(a) a *viscosity subsolution* of (2.19) if, whenever $x_0 \in \Omega$, $\varphi \in C^2(\Omega)$ and $u - \varphi$ has a local maximum at x_0 we have that

$$F(x_0, \varphi(x_0), D\varphi(x_0), D^2\varphi(x_0)) \geq 0,$$

(b) a *viscosity supersolution* of (2.19) if, whenever $x_0 \in \Omega$, $\varphi \in C^2(\Omega)$ and $u - \varphi$ has a local minimum at x_0 we have that

$$F(x_0, \varphi(x_0), D\varphi(x_0), D^2\varphi(x_0)) \leq 0,$$

(c) a *viscosity solution* if it is a sub- and supersolution.

Remark 2.44 (viscosity solutions). Several remarks must immediately be made about Definition 2.43.

- While the motivation provided assumed that the function u is smooth, the definition only requires its continuity.
- By approximation and continuity of F, it is sufficient to verify the condition for quadratic polynomials $\varphi \in \mathbb{P}_2$; see Caffarelli and Cabré (1995, Proposition 2.4).
- If $u \in C^2(\Omega)$ is a classical solution then it is a viscosity solution. This follows from the ellipticity of F. Moreover, sufficiently smooth viscosity solutions are also classical; see Caffarelli and Cabré (1995, Lemma 2.5) and Katzourakis (2015, Theorem 2.11).
- This definition talks only about solutions to equation (2.19), not the boundary value problem (2.6). More details on this issue will be provided below; see Definition 2.48 and Section 2.5.3.
- The definition assumes that the candidate solution *can* be touched from above (below). At points where this is not possible there is nothing to verify, and the function automatically satisfies the equation at these points.
- For the divergence-form operators L and $\tilde{L}$ of Definition 2.3 and Example 2.4, respectively, it is known (Ishii 1995) that the concepts of

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

weak solution, in the sense of Definition 2.19, and viscosity solutions coincide.

- We will not provide a historical account of the origin and development of this definition. The interested reader can consult the classical reference by Crandall *et al.* (1992).

A remarkable property of viscosity solutions is their *stability*, which is detailed in the following two results. For a proof of the first one we refer to Caffarelli and Cabré (1995, Proposition 2.8), Katzourakis (2015, Theorem 3.2) or Crandall *et al.* (1992, Section 6). For the second one, we refer to Caffarelli and Cabré (1995, Proposition 2.7) or Katzourakis (2015, Theorem 3.12).

Theorem 2.45 (limits and viscosity solutions). Let $\{F_k\}_{k\in\mathbb{N}}$ be a sequence of uniformly elliptic operators in the sense of Definition 2.10, and let $\{u_k\}_{k\in\mathbb{N}} \subset C(\Omega)$, for each k, be viscosity subsolutions to the equations

$$F_k(x, u_k, Du_k, D^2u_k) = 0.$$

If, as $k \to \infty$, $F_k \to F$ uniformly on compact subsets of $\Omega \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{S}^d$ and $u_k \to u$ uniformly in compact subsets of Ω, then u is a viscosity subsolution of

$$F(x, u, Du, D^2u) = 0.$$

Theorem 2.46 (suprema of subsolutions). Let $\mathcal{U} \subset C(\Omega)$ be a set of viscosity subsolutions of (2.19). For $x \in \Omega$ define

$$\bar{u}(x) = \sup\{u(x) : u \in \mathcal{U}\}.$$

Suppose that $\bar{u} \in C(\Omega)$ and is bounded. Then $\bar{u}$ is a viscosity subsolution of (2.19).

2.5.2. Existence and uniqueness

Let us now turn our attention to the existence and uniqueness of viscosity solutions to (2.6). To do so we must specify in which sense the boundary conditions are being understood. We begin by introducing the notion of semi-continuity.

Definition 2.47 (semi-continuity). We say the function $u \in LSC(\Omega)$ (is *lower semi-continuous*) if, for all $x \in \Omega$,

$$u(x) \leq \liminf_{y\to x} u(y).$$

On the other hand, we say $u \in USC(\Omega)$ (is *upper semi-continuous*) if $-u \in LSC(\Omega)$.

Note that, in Definition 2.43 and the discussion that followed, nothing would have changed if we had only required subsolutions and supersolutions

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

to be upper and lower semi-continuous, respectively. With this definition at hand, we may define viscosity solutions to the Dirichlet problem.

Definition 2.48 (solution to the Dirichlet problem). Let F be elliptic in the sense of Definition 2.10 and $g \in C(\partial\Omega)$.

(a) The function $u_\star \in USC(\bar{\Omega})$ is a viscosity subsolution to (2.6) if it is a viscosity subsolution to the equation (2.19) and $u_\star(x) \leq g(x)$ for all $x \in \partial\Omega$.

(b) The function $u^\star \in LSC(\bar{\Omega})$ is a viscosity supersolution to (2.6) if it is a viscosity supersolution to (2.19) and $u^\star(x) \geq g(x)$ for all $x \in \partial\Omega$.

(c) The function $u \in C(\bar{\Omega})$ is a viscosity solution to (2.6) if it is a sub- and supersolution.

Note that this definition requires the boundary values to be attained in the *classical sense*. Different boundary conditions might require a different interpretation, and we will briefly comment on this below.

We now turn our attention to the existence of solutions and the so-called Perron's method. Simply put, this method provides existence under the assumption that the problem cannot have more than one solution. While uniqueness of classical solutions is immediate, as shown in Theorem 2.16, in this more general setting we need one additional condition.

Definition 2.49 (comparison). We say that the Dirichlet problem (2.6) satisfies a comparison principle if, whenever $u_\star \in USC(\Omega)$ and $u^\star \in LSC(\Omega)$ are sub- and supersolutions, respectively, we have

$$u_\star \leq u^\star \quad \text{in } \Omega.$$

Note that from Definition 2.49, it immediately follows that (2.6) cannot have more than one solution. Indeed, if u and v are two viscosity solutions, then, in particular, u is a subsolution and v a supersolution; consequently, $u \leq v$. An analogous reasoning yields the reverse inequality.

With these two conditions at hand, we proceed to show existence of solutions. For a proof, we refer the reader, for instance, to Crandall *et al.* (1992, Theorem 4.1) and Katzourakis (2015, Theorem 5.3).

Theorem 2.50 (Perron's method). Let F be elliptic in the sense of Definition 2.10 and $g \in C(\partial\Omega)$. Assume that the Dirichlet problem (2.6) satisfies a comparison principle in the sense of Definition 2.49. If there exist a subsolution $u_\star$ and a supersolution $u^\star$ to (2.6) that satisfy the boundary condition, then

$$u(x) = \sup\big\{v(x) : u_\star \leq v \leq u^\star \text{ and } v \text{ is a subsolution}\big\}$$

defines a viscosity solution to (2.6).

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Note that, while Theorem 2.50 provides a somewhat explicit construction of the unique solution to (2.6), one still needs to verify the existence of sub- and supersolutions that satisfy the boundary condition in a *classical sense*. This must be done on a case-by-case basis, and we refer the reader to Crandall *et al.* (1992, Example 4.6) and Katzourakis (2015, Application 5.9) for two examples where these are constructed.

It remains to understand which operators satisfy the comparison principle of Definition 2.49. Loosely speaking, similar ideas to those presented in Theorem 2.16 should yield uniqueness of viscosity solutions. However, the arguments presented there cannot be applied directly since we are dealing with functions that are merely continuous, and additional structural conditions must be imposed. This is due to the subtle fact – which may have escaped the reader – that Definition 2.10 is *too general*. By this we mean that, for instance, first-order and parabolic equations fit into this definition. For this reason, many authors say that an operator is *degenerate elliptic* if it only satisfies the monotonicity condition with respect to the M variable. This is in contrast with uniform ellipticity, which precludes these two degenerate cases.

For the sake of illustration, let us concentrate our efforts on finding a comparison principle for uniformly elliptic equations. We begin by showing, following Kawohl and Kutev (2007, Example 1), that uniform ellipticity is not enough to ensure a comparison principle.

Example 2.51 (lack of comparison). Consider the Dirichlet problem

$$u'' + 18x(u')^4 = 0 \ \text{ in } (-1,1), \quad u(-1) = -b,\ u(1) = b,$$

with $b > 1$. Clearly, the equation is uniformly elliptic. It is easy to check that the functions

$$u_\star(x) = \begin{cases} \sqrt[3]{x} - 1 + b & x \in [0,1], \\ \sqrt[3]{x} + 1 - b & x \in [-1,0), \end{cases} \qquad u^\star(x) = \begin{cases} \sqrt[3]{x} - 1 + b & x \in (0,1], \\ \sqrt[3]{x} + 1 - b & x \in [-1,0], \end{cases}$$

which differ only at the origin, are viscosity sub- and supersolutions, respectively, and that they satisfy the boundary values. Note, however, that

$$\max_{x\in(-1,1)} \{u_\star(x) - u^\star(x)\} = u_\star(0) - u^\star(0) = 2b - 2 > 0.$$

While, to our knowledge, necessary and sufficient conditions for the existence of a comparison principle for a general elliptic operator are not known, there are several sufficient conditions. We collect these in the following result.

Theorem 2.52 (existence of comparison principle). If the Dirichlet problem (2.6) satisfies any of the structural conditions given below, then it satisfies a comparison principle in the sense of Definition 2.49.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

(a) (Katzourakis 2015, Theorem 6.1) The dependence with respect to x is decoupled, that is, the equation reads

$$F(u, Du, D^2u) = f,$$

with $F \in C(\mathbb{R}, \mathbb{R}^d, \mathbb{S}^d)$, $f \in C(\bar{\Omega})$. The operator F is elliptic, and satisfies, for some $\gamma > 0$,

$$F(r, \mathbf{p}, M) \geq F(s, \mathbf{p}, M) + \gamma(s - r), \quad \text{for all } r \leq s.$$

(b) The operator F is elliptic, independent of the r and $\mathbf{p}$ variables, and satisfies, for some $\lambda > 0$,

$$F(x, M + tI) \geq F(x, M) + \lambda t, \quad \text{for all } t \geq 0.$$

(c) (Trudinger 1988) The operator F is uniformly elliptic, Lipschitz-continuous in $\mathbf{p}$ and the following continuity assumption holds:

$$|F(x, r, \mathbf{p}, M) - F(y, r, \mathbf{p}, M)| \leq \mu_2 |x - y|^{1/2} |M| + \omega(|x - y|),$$

for all $x, y \in \Omega$, r and $\mathbf{p}$ in a suitable ball and $\omega(a) \to 0$ as $a \downarrow 0$. Additionally, one must assume that sub- and supersolutions belong to $C^{0,1}(\Omega)$.

(d) (Koike 2004) The dependence with respect to the r variable is decoupled, that is, the equation reads

$$\nu u + F(x, Du, D^2u) = 0,$$

with either $\nu < 0$ and F elliptic, or $\nu \leq 0$, F uniformly elliptic and Lipschitz in the $\mathbf{p}$ variable, for $\mathbf{p} \in \mathbb{R}^d$.

(e) (Silvestre 2015) The operator F is independent of x and $\mathbf{p}$ and is strictly decreasing in r, that is, whenever $r > s$,

$$F(r, M) < F(s, M), \quad \text{for all } M \in \mathbb{S}^d.$$

Other conditions can be found in the literature.

2.5.3. Other boundary conditions

So far, for all notions of solutions, we have only discussed the Dirichlet problem (see the second equation in (2.6)). Moreover, for viscosity solutions we have assumed that the boundary conditions are attained in a classical sense. Let us here consider other types of boundary conditions as well as generalized notions for them. Consider

$$F(x, u, Du, D^2u) = 0 \ \text{ in } \Omega, \quad B(x, u, Du) = 0 \ \text{ on } \partial\Omega, \tag{2.20}$$

where, as before, the map F is elliptic, but its domain of definition on the x variable is now $\bar{\Omega}$. The function $B : \partial\Omega \times \mathbb{R} \times \mathbb{R}^d$ is assumed to be non-increasing in its second argument, that is,

$$r \geq s \Rightarrow B(x, r, \mathbf{p}) \leq B(x, s, \mathbf{p}), \quad \text{for all } x \in \partial\Omega, \ \mathbf{p} \in \mathbb{R}^d.$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

The Dirichlet problem obviously falls into this description with $B(x, r, \mathbf{p}) = g(x) - r$, but others are also admissible. For instance, let $\mathbf{n}(x)$ denote the outer normal to $\partial\Omega$ at x and let $\boldsymbol{\nu} : \partial\Omega \to \mathbb{R}^d$ be such that, for all $x \in \partial\Omega$, we have $\boldsymbol{\nu}(x) \cdot \mathbf{n}(x) > 0$. The boundary condition

$$B(x, \mathbf{p}) = \boldsymbol{\nu}(x) \cdot \mathbf{p} - g(x)$$

gives rise to the so-called *oblique derivative* problem; if $\boldsymbol{\nu} = \mathbf{n}$, this is the Neumann problem. A nonlinear example is the capillarity condition

$$B(x, r, \mathbf{p}) = \mathbf{n} \cdot \mathbf{p} - g(x, r)\sqrt{1 + |\mathbf{p}|^2}.$$

At the beginning of Section 2.5.1 the introduction of viscosity solution was motivated by the assumption that the function $u - \varphi$ had a local maximum (minimum) at $x_0 \in \Omega$. When dealing with boundary conditions, we must now allow for $x \in \partial\Omega$. At these points the relations that led to the definition of viscosity solution no longer hold, and a modification is necessary. It turns out that the correct notion is as follows.

Definition 2.53 (viscosity subsolution). With the functions F and B as above, we say that $u \in USC(\bar{\Omega})$ is a *viscosity subsolution* to (2.20) if it is a viscosity subsolution to (2.19) and, whenever there is a $\varphi \in C^2(\mathbb{R}^d)$ that touches the graph of u from above at $x_0 \in \partial\Omega$, then either

$$B(x_0, \varphi(x_0), D\varphi(x_0)) \geq 0 \quad \text{or} \quad F(x_0, \varphi(x_0), D\varphi(x_0), D^2\varphi(x_0)) \geq 0.$$

In an analogous manner we can consider supersolutions and, as before, a solution to (2.20) is a function $u \in C(\bar{\Omega})$ that is both a sub- and supersolution.

It is important to realize that boundary conditions in the viscosity sense, in general, are *not* equivalent to those in the classical sense. The reason for this, once more, is that Definition 2.10 is rather general and, for instance, allows us to consider first-order equations for which Dirichlet conditions cannot be imposed on the whole boundary. It is therefore natural to ask when a boundary condition in the viscosity sense is attained classically. Let us briefly elaborate on this issue for the Dirichlet problem (2.6). We begin with the definition of a barrier.

Definition 2.54 (barrier). We say that (2.6) has *barriers* at $x_0 \in \partial\Omega$ if there exist two continuous functions $\bar{u}$, $\underline{u}$ that are super- and subsolutions to (2.6), respectively, and that satisfy $\bar{u}(x_0) = \underline{u}(x_0) = g(x_0)$.

Proposition 2.55 (viscosity versus classical). Let u be a viscosity solution to (2.6) in the sense of Definition 2.53. If barriers exist at $x_0 \in \partial\Omega$, then $u(x_0) = g(x_0)$.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

In other words, classical and viscosity conditions coincide at points where it is possible to construct a barrier. We conclude this discussion by providing a sufficient condition for the existence of barriers.

Proposition 2.56 (existence of barriers). Let Ω be such that it has a tangent ball from outside at every point of $\partial\Omega$. If F is uniformly elliptic, Lipschitz with respect to all its variables and, for every $x \in \bar{\Omega}$, we have $F(x, 0, \mathbf{0}, 0) = 0$, then barriers exist at every point $x_0 \in \partial\Omega$.

2.5.4. Regularity

To finalize the presentation on viscosity solutions, we elaborate on their regularity. This is important not only because these results will serve as a guide to establishing rates of convergence for numerical schemes, but also many of the ideas and techniques that we present here have a discrete analogue that will be detailed in subsequent sections.

We begin with a result by Nirenberg (1953) that shows that in two dimensions, essentially, all solutions to elliptic equations are locally $C^{2,\alpha}$.

Theorem 2.57 (regularity in two dimensions). Let $d = 2$. Assume that F is uniformly elliptic in the sense of Definition 2.10 and that it has bounded first derivatives with respect to all its arguments. If u is a solution to (2.19), then for every $\omega \Subset \Omega$ there are $C > 0$, $\alpha \in (0, 1)$ that depend only on the ellipticity constants of F, the bounds on its first derivatives and the distance between ω and $\partial\Omega$, for which

$$\|u\|_{C^{2,\alpha}(\omega)} \leq C\|u\|_{L^\infty(\Omega)}.$$

It is remarkable that this result was obtained long before the development of the theory of viscosity solutions.

To obtain global regularity or results in more dimensions we begin by introducing several notions of a more or less geometric nature. Recall that a function $u : \Omega \to \mathbb{R}$ is convex if

$$u(\alpha x + (1-\alpha)y) \leq \alpha u(x) + (1-\alpha)u(y), \quad \text{for all } x, y \in \Omega,\ \alpha \in [0, 1].$$

For a convex function we define its subdifferential as follows.

Definition 2.58 (subdifferential). Let $u \in C(\Omega)$. The subdifferential of u at the point $x \in \Omega$ is

$$\partial u(x) = \left\{\mathbf{p} \in \mathbb{R}^d : u(y) - u(x) \geq \mathbf{p} \cdot (y - x)\ \forall y \in \Omega\right\}.$$

It is well known (Ekeland and Témam 1999) that if u is convex then $\partial u(x) \neq \emptyset$, and that if u is differentiable at x then $\partial u(x) = \{Du(x)\}$. Given a function u, we can always construct the largest convex function lying below u. This gives rise to the *convex envelope*. In what follows we will only need this concept for the negative part of a function, so we define the convex envelope in this restricted setting.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

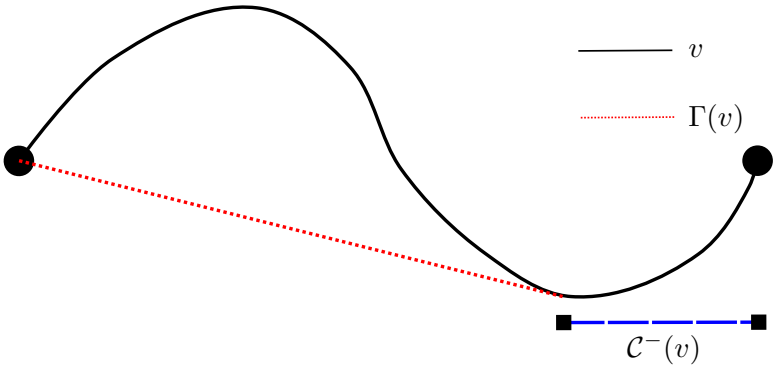


Figure 2.1. Convex envelope (dotted red) and contact set (dashed blue).

Definition 2.59 (convex envelope and contact set). Let B_r be a ball such that $\Omega \subset B_r$, and let $v \in C(\Omega)$ with $v \geq 0$ on $\partial\Omega$. Extend v^- by zero to $B_r \setminus \Omega$. The *convex envelope* of v is defined, for $x \in B_r$, by

$$\Gamma(v)(x) = \sup\big\{L(x) : L(z) \leq -v^-(z)\ \forall z \in B_r,\ L \in \mathbb{P}_1\big\}.$$

The points at which these two functions coincide are called *contact points*:

$$\mathcal{C}^-(v) = \big\{x \in B_r : v(x) = \Gamma(v)(x)\big\}.$$

An illustration of the convex envelope of a function and its contact set is given in Figure 2.1. From the figure it is intuitively clear that, for fixed values of v on the boundary, how deep the graph of v can go depends only on the values of v at $\mathcal{C}^-(v)$. The formalization of this observation is the so-called Alexandrov estimate.

Theorem 2.60 (Alexandrov estimate). Let $v \in C(\bar{B}_r)$ with $v \geq 0$ on ∂B_r. If $\Gamma(v) \in C^{1,1}(B_r)$, then

$$\sup_{B_r} v^- \leq Cr|\partial\Gamma(v)(\mathcal{C}^-(v))|^{1/d}.$$

In other words, there is a set $A \subset B_r$ that satisfies $|B_r \setminus A| = 0$ and for which we have

$$\sup_{B_r} v^- \leq Cr\left(\int_{A\cap\mathcal{C}^-(v)} \det D^2\Gamma(v)\right)^{1/d},$$

where the constant C depends only on d.

Proof. For completeness, we sketch the proof for $v \in C^2(B_r)$, since in this case $\partial\Gamma(v)$ is single-valued on $\mathcal{C}^-(v)$.

Let $M = \sup_{B_r} v^-/2r$ and assume that $B_M \subset \partial\Gamma(v)(\mathcal{C}^-(v))$. If that is the case,

$$M^d \leq C|\partial\Gamma(v)(\mathcal{C}^-(v))|,$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

for a constant that depends only on the dimension d. This shows the first estimate. On the other hand, a simple change of variables yields

$$|\partial\Gamma(v)(\mathcal{C}^-(v))| = \int_{\partial\Gamma(v)(\mathcal{C}^-(v))} = \int_{\mathcal{C}^-(v)} \det D^2\Gamma(v),$$

so that the second statement follows from the first one.

We now show the inclusion $B_M \subset \partial\Gamma(v)(\mathcal{C}^-(v))$. Let $z \in B_r$ be a point where $\sup_{B_r} v^-$ is attained. For $\mathbf{a} \in B_M$, define the affine function

$$L(x) = -\sup_{B_r} v^- + \mathbf{a} \cdot (x - z)$$

and note that $L(z) = -\sup_{B_r} v^-$ and, for all $x \in B_r$,

$$L(x) \leq -\sup_{B_r} v^- + |\mathbf{a}|\,|x - z| < -\sup_{B_r} v^- + 2Mr = 0.$$

Since $Dv(z) = 0$ there is a $x_1 \in B_r$ such that $v(x_1) < L(x_1) < 0$. In addition, we have that $v(x) \geq 0 > L(x)$ for $x \in \partial B_r$. This shows that, if $\bar{x} \in B_r$ is a point where $v - L$ attains its minimum, then $v(\bar{x}) < L(\bar{x}) \leq 0$ and $Dv(\bar{x}) = DL(\bar{x}) = \mathbf{a}$.

Define $\tilde{L}(x) = L(x) + v(\bar{x}) - L(\bar{x})$ and note that $v(\bar{x}) = \tilde{L}(\bar{x})$, $Dv(\bar{x}) = D\tilde{L}(\bar{x}) = \mathbf{a}$ and, for every $x \in B_r$ $v(x) \geq \tilde{L}(x)$. In other words, $\tilde{L}$ is a supporting hyperplane for v. This shows that $\bar{x} \in \mathcal{C}^-(v)$ and that $\mathbf{a} \in \partial\Gamma(v)(\bar{x})$, that is, $B_M \subset \partial\Gamma(v)(\mathcal{C}^-(v))$. □

Note that in Theorem 2.60 only the contact set is relevant. This is due to the fact that if for $x_0 \in B_r$ we have $\Gamma(v)(x_0) < v(x_0)$, then locally $\Gamma(v)$ is affine, and thus $D^2\Gamma(v) = 0$.

With this estimate at hand we proceed to obtain the fundamental *a priori* estimate for viscosity solutions, the so-called Alexandrov–Bakelman–Pucci estimate. We begin by providing some motivation for this result. To do so, assume that $u \in C^2(\bar{B}_r)$ with $u \geq 0$ on ∂B_r satisfies $\mathcal{L}u \leq f$ in B_r. In this setting we have that, for $x \in \mathcal{C}^-(u)$, $D^2u(x) \geq 0$, and consequently $A(x) : D^2u(x) \geq 0$ as well. This, in particular, implies that $f(x) \geq 0$. Let $D = \inf_{x\in\bar{B}_r} \det A(x) > 0$ and observe that

$$\det A(x)D^2u(x) = \det A(x) \det D^2u(x) \geq D \det D^2u(x) \geq 0.$$

Let $\sigma(A(x)D^2u(x)) = \{\mu_i(x)\}_{i=1}^d$. Then an application of the arithmetic–geometric inequality reveals that

$$\begin{aligned}
&\det A(x)D^2u(x) \\
&\quad = \prod_{i=1}^d \mu_i(x) = \left(\prod_{i=1}^d \mu_i(x)^{1/d}\right)^d \leq \left(\frac{1}{d}\sum_{i=1}^d \mu_i(x)\right)^d \\
&\quad = \left(\frac{1}{d}\operatorname{tr} A(x)D^2u(x)\right)^d = \left(\frac{1}{d}A(x) : D^2u(x)\right)^d \leq \left(\frac{1}{d}f(x)\right)^d,
\end{aligned}$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

where in the last step we used that $\mathcal{L}u \leq f$. Theorem 2.60 then yields that

$$\sup_{B_r} u^- \leq Cr\left(\int_{\mathcal{C}^-(u)} (f^+)^d\right)^{1/d},$$

for a constant that depends only on d and D.

While the considerations presented assumed that we were working with a linear equation, we essentially used that the matrix was uniformly positive definite and bounded, that is, the operator $\mathcal{L}$ is elliptic. A similar conclusion can be drawn from the fact that an operator is elliptic in the sense of Definition 2.10. We begin by observing (Caffarelli and Cabré 1995, Lemma 2.2) that F is uniformly elliptic if and only if

$$F(x, r, \mathbf{p}, M + N) - F(x, r, \mathbf{p}, M) \leq \Lambda|N^+| - \lambda|N^-|,$$

where $N = N^+ - N^-$ with $N^+, N^- \geq 0$ and $N^+N^- = 0$. Now, if u is a sufficiently smooth subsolution of (2.19), from the observation above we have

$$\begin{aligned} f(x) &:= -F(x, 0, \mathbf{0}, 0) \leq F(x, u(x), Du(x), D^2u(x)) - F(x, 0, \mathbf{0}, 0) \\ &\leq \Lambda \sum_{\lambda_i(u)>0} \lambda_i(u) + \lambda \sum_{\lambda_i(u)<0} \lambda_i(u), \end{aligned}$$

where $\sigma(D^2u(x)) = \{\lambda_i(u)\}_{i=1}^d$. Similarly, for a supersolution we have

$$f(x) \geq \Lambda \sum_{\lambda_i(u)<0} \lambda_i(u) + \lambda \sum_{\lambda_i(u)>0} \lambda_i(u).$$

This motivates the following definitions which, in a sense, describe the class of all possible viscosity solutions to uniformly elliptic equations.

Definition 2.61 (class $\mathcal{S}$). Let the operator F be uniformly elliptic in the sense of Definition 2.10 and let $f = -F(\cdot, 0, \mathbf{0}, 0)$. We say that $u \in \underline{\mathcal{S}}(\lambda, \Lambda, f)$ if $u \in C(\Omega)$ and the inequality

$$f(x) \leq \Lambda \sum_{\lambda_i(u)>0} \lambda_i(u) + \lambda \sum_{\lambda_i(u)<0} \lambda_i(u)$$

holds in the viscosity sense. Similarly, we say that $u \in \overline{\mathcal{S}}(\lambda, \Lambda, f)$ if $u \in C(\Omega)$ and

$$f(x) \geq \Lambda \sum_{\lambda_i(u)<0} \lambda_i(u) + \lambda \sum_{\lambda_i(u)>0} \lambda_i(u)$$

in the viscosity sense. Finally $\mathcal{S}(\lambda, \Lambda, f) = \underline{\mathcal{S}}(\lambda, \Lambda, f) \cap \overline{\mathcal{S}}(\lambda, \Lambda, f)$.

With this notation at hand we present the Alexandrov–Bakelman–Pucci (ABP) estimate.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Theorem 2.62 (ABP estimate). Let $u \in \overline{\mathcal{S}}(\lambda, \Lambda, f)$ in Ω with $u \geq 0$ on $\partial\Omega$ and assume that f is continuous and bounded in Ω. Then

$$\sup_{\Omega} u^- \leq Cr\left(\int_{\mathcal{C}^-(u)} (f^+)^d\right)^{1/d},$$

where the constant C depends only on d, λ and Λ, and r is such that $\Omega \subset B_{r/2}$, and we have extended u by zero outside Ω.

Note that, as in the Alexandrov estimate, only the contact set $\mathcal{C}^-(u)$ is relevant in this estimate. Note also that we obtain control of the L^∞-norm of u in terms of the L^d-norm of the data f. While Theorem 2.62 is a sort of stability estimate, it is also useful in establishing regularity of solutions. To do so, we begin with the Harnack inequality of Krylov and Safonov; see Safonov (1980). In what follows, we let Q_l denote a cube with sides parallel to the coordinate axes and of length l.

Theorem 2.63 (Harnack inequality). Let $u \in \mathcal{S}(\lambda, \Lambda, f)$ in Q_1 with $f \in C(Q_1) \cap L^\infty(Q_1)$. If $u \geq 0$ in Q_1, then

$$\sup_{Q_{1/2}} u \leq C\Big(\inf_{Q_1} u + \|f\|_{L^d(Q_1)}\Big),$$

where the constant C depends only on d, λ and Λ.

Since this will be useful later on, let us now show how one can obtain interior Hölder continuity of functions in $\mathcal{S}(\lambda, \Lambda, f)$ from a Harnack inequality. We begin with a technical result, commonly referred to as an iteration lemma; see Han and Lin (2011, Lemma 3.4) and Gilbarg and Trudinger (2001, Lemma 8.23).

Lemma 2.64 (iteration). Let $\varphi : (0, R] \to \mathbb{R}$ be non-decreasing. Assume that for some $A > 0$, $B \geq 0$ and $\alpha > \beta$ we have

$$\varphi(\rho) \leq A\left[\left(\frac{\rho}{r}\right)^\alpha + \varepsilon\right]\varphi(r) + Br^\beta$$

whenever $0 < \rho \leq r \leq R$. Then, for every $\gamma \in (\beta, \alpha)$ there exists ε_0 such that if $\varepsilon < \varepsilon_0$ then

$$\varphi(r) \leq C\left[\frac{\varphi(R_0)}{R_0} r^\gamma + Br^\beta\right], \quad \text{for all } r \in [0, R_0),$$

for some fixed constant C.

With this result at hand we obtain local Hölder continuity.

Theorem 2.65 (local Hölder regularity). Let $u \in \mathcal{S}(\lambda, \Lambda, f)$ in Q_1. Then there exists $\alpha \in (0, 1)$ for which $u \in C^\alpha(\bar{Q}_{1/2})$ and

$$\|u\|_{C^\alpha(\bar{Q}_{1/2})} \leq C\big(\|u\|_{L^\infty(Q_1)} + \|f\|_{L^d(Q_1)}\big),$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

where the constant C is independent of u and f.

Proof. The proof is rather standard, so we merely sketch it. Let $m_r = \inf_{Q_r} u$, $M_r = \sup_{Q_r} u$ and $\varpi_r = M_r - m_r$. Applying the Harnack inequality of Theorem 2.63 to the non-negative function $u - m_1$ yields

$$M_{1/2} - m_1 \leq C(m_{1/2} - m_1 + \|f\|_{L^d(Q_1)}).$$

Since the function $M_1 - u \geq 0$, we can once more apply the Harnack inequality to obtain

$$M_1 - m_{1/2} \leq C(M_1 - M_{1/2} + \|f\|_{L^d(Q_1)}).$$

Adding these two inequalities yields

$$\varpi_{1/2} \leq \mu\varpi_1 + 2\|f\|_{L^d(Q_1)},$$

where $\mu = (C-1)/(C+1) \in (0,1)$.

A similar argument for the functions $u_r(y) = u(ry)/r^2$ and $f_r(y) = f(ry)$ with $y \in Q_1$ reveals that

$$\varpi_r \leq \mu\varpi_{r/2} + 2r\|f\|_{L^d(Q_1)}.$$

An application of the iteration Lemma 2.64 with $\varphi(r) = \varpi_r$ immediately yields Hölder continuity and the estimate. □

In a similar fashion, we can establish smoothness up to the boundary; see Caffarelli and Cabré (1995, Proposition 4.14).

Theorem 2.66 (global regularity). Let Ω be sufficiently smooth and $u \in \mathcal{S}(\lambda, \Lambda, f) \cap C(\bar{\Omega})$ with $f \in C(\Omega)$. Let $g = u_{|\partial\Omega}$ and ϱ be a modulus of continuity of g. Then there is a modulus of continuity ϱ^* of u in $\bar{\Omega}$ which depends only on $\lambda, \Lambda, \varrho, \|f\|_{L^d(\Omega)}$ and $\|g\|_{L^\infty(\Omega)}$.

We now focus on Hölder estimates for first and second derivatives. To simplify the presentation, in problem (2.6), the PDE takes the form

$$F(x, D^2u) = f \quad \text{in } \Omega. \tag{2.21}$$

To quantify the smoothness of F with respect to the x variable, we introduce the function

$$\beta(x) = \sup_{M \in \mathbb{S}^d} \frac{|F(M,x) - F(M,0)|}{|M| + 1}.$$

The local regularity is as follows.

Theorem 2.67 (local $C^{2,\alpha}$-regularity). Assume that F is uniformly elliptic in the sense of Definition 2.10, $\beta, f \in C^\alpha(B_1)$ and that there is a constant $\bar{\alpha} \in (0,1)$ such that for any $M \in \mathbb{S}^d$ with $F(0,M) = 0$ and $w_0 \in C(\partial B_1)$ there exists $w \in C^2(B_1) \cap C(\bar{B}_1) \cap C^{2,\bar{\alpha}}(B_{1/2})$ which satisfies

$$F(0, D^2w + M) = 0 \ \text{ in } B_1, \quad w = w_0 \ \text{ on } \partial B_1, \tag{2.22}$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

with an *a priori* estimate. If u is a viscosity solution of (2.21) then $u \in C^{2,\alpha}(\bar{B}_{1/2})$ for some $\alpha \in (0,1)$ with an *a priori* estimate.

To apply this theorem, one must verify that solutions to (2.22) have $C^{2,\bar{\alpha}}$-estimates. For an F that is convex, the Evans–Krylov theorem (Caffarelli and Silvestre 2010*a*) provides such an estimate.

Theorem 2.68 (Evans–Krylov). Let F be convex and depend only on M. If u is a viscosity solution of $F(D^2u) = 0$ in B_1, then

$$\|u\|_{C^{2,\bar{\alpha}}(\bar{B}_{1/2})} \leq C\big(\|u\|_{L^\infty(B_1)} + |F(0)|\big),$$

for some constants $\bar{\alpha} \in (0,1)$ and C that depend only on the dimension d and the ellipticity of F.

We mention that Theorem 2.67 can be applied to the Hamilton–Jacobi–Bellman operators of Example 2.12. With the aid of the so-called method of continuity, this allows us to show that solutions to the Dirichlet problem (2.6) for this class of operators are classical.

On the other hand, it is natural to ask if the convexity of F is essential for this result. To understand this, we begin by providing a $C^{1,\alpha}$-estimate without convexity assumptions.

Theorem 2.69 ($C^{1,\alpha}$-regularity). Let u be a viscosity solution of

$$F(D^2u) = 0 \ \text{ in } B_1.$$

Then

$$\|u\|_{C^{1,\bar{\alpha}}(\bar{B}_{1/2})} \leq C\big(\|u\|_{L^\infty(B_1)} + |F(0)|\big),$$

where the constants $\bar{\alpha} \in (0,1)$ and C depend only on the dimension and ellipticity of F.

A similar argument to Theorem 2.67 then allows us to conclude that, in this setting and under similar assumptions, solutions to (2.6) are locally $C^{1,\alpha}$ with an *a priori* estimate. However, there exists a series of counterexamples (Nadirashvili and Vlăduţ 2007, Nadirashvili and Vlăduţ 2008, Nadirashvili, Tkachev and Vlăduţ 2012) showing that, in general, the convexity assumption on F cannot be removed.

Example 2.70 (non-classical viscosity solution). Let $d = 5$. Define

$$P_5(x) = x_1^3 + \frac{3}{2}x_1(x_3^2 + x_4^2 - 2x_5^2 - 2x_2^2) + \frac{3\sqrt{3}}{2}(x_2x_3^2 - x_2x_4^2 + 2x_3x_4x_5),$$

and, for $\delta \in [0,1)$,

$$w(x) = P_5(x)/|x|^{1+\delta} \in C^{1,1-\delta}(\bar{B}_1) \setminus C^2(B_1).$$

There exists an Isaacs operator F that depends only on M and is Lipschitz,

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

such that w is a viscosity solution of

$$F(D^2 w) = 0 \ \text{ in } B_1, \quad w = P_5 \ \text{ on } \partial B_1.$$

The existence of non-classical solutions in dimensions $3 \leq d < 5$ is an open problem.

We conclude by providing global regularity results in the general case; see Caffarelli and Souganidis (2008), Caffarelli and Silvestre (2010*b*) and Turanova (2015).

Theorem 2.71 (global regularity). Let Ω be sufficiently regular and let u be a viscosity solution to (2.6) with F of the form (2.21) being Lipschitz and uniformly elliptic. If $f \in C^{0,1}(\Omega)$ and, for some $\gamma \in (0,1]$, $g \in C^{1,\gamma}(\partial\Omega)$, then $u \in C^{1,\alpha}(\Omega) \cap C^{0,1}(\bar{\Omega})$ with

$$\|u\|_{C^{1,\alpha}(\Omega)} \leq C\big(\|f\|_{L^\infty(\Omega)} + \|g\|_{C^{1,\gamma}(\partial\Omega)}\big),$$

where $\alpha \in (0,1)$ and C depend only on d, λ, Λ and the smoothness of F.

3. Monotonicity in numerical methods

In this section we review some basic properties of numerical methods and state sufficient conditions to ensure that discrete approximations converge to the solutions of the underlying PDE. The underlying theme of this section is that as the notion of solution to the PDE becomes weaker, additional conditions of the numerical approximation are required to guarantee convergence. For example, for linear differential equations, the well-known Lax–Richtmyer equivalence theorem shows that any consistent scheme is convergent if and only if it is stable; these results extend to mildly nonlinear problems as well. However, in the fully nonlinear regime, consistency and stability are no longer sufficient in general. Rather, additional monotonicity conditions are required, that essentially mimic the comparison principles discussed in the previous section.

We note that while the content of this section deals with finite difference and finite element methods, the main ideas extend to other discretization techniques as well.

3.1. Stability, consistency and monotonicity implies convergence

As before, let Ω be an open subset of $\mathbb{R}^d$ with Lipschitz boundary $\partial\Omega$. We consider numerical approximations of elliptic problems of the most general form (2.20), which we recall for convenience:

$$F(x, u, Du, D^2 u) = 0 \ \text{ in } \Omega, \quad B(x, u, Du) = 0 \ \text{ on } \partial\Omega. \tag{3.1}$$

Here $F \in C(\Omega \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{S}^d)$ is locally bounded and elliptic in the sense of Definition 2.10.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

We further assume that $B \in C(\Omega \times \mathbb{R} \times \mathbb{R}^d)$ is non-increasing in its second argument. For the moment, we consider viscosity solutions that satisfy the boundary conditions only in a viscosity sense; see Definition 2.53. In this case, and to simplify the presentation, we define the operator

$$\mathsf{F}(x, r, \mathbf{p}, M) = \begin{cases} F(x, r, \mathbf{p}, M) & \text{if } x \in \Omega, \\ B(x, r, \mathbf{p}) & \text{if } x \in \partial\Omega, \end{cases}$$

so that (3.1) becomes

$$\mathsf{F}[u] := \mathsf{F}(x, u, Du, D^2u) = 0 \ \text{ in } \bar{\Omega}. \tag{3.2}$$

Note that, since B is non-increasing in its second argument and F is elliptic, the operator F is elliptic in the sense of Definition 2.10. We further see that, according to Definition 2.53, $u \in C(\bar{\Omega})$ is a viscosity solution to (3.1) (equivalently (3.2)) if it is a viscosity solution to

$$F(x, u, Du, D^2u) = 0 \ \text{ in } \Omega,$$

and

$$\begin{aligned} \max\{F(x, u, Du, D^2u), B(x, u, Du)\} &\geq 0 \ \text{ on } \partial\Omega, \\ \min\{F(x, u, Du, D^2u), B(x, u, Du)\} &\leq 0 \ \text{ on } \partial\Omega \end{aligned}$$

in the viscosity sense.

We now consider approximation schemes. Find $u_h \in X_h$ satisfying

$$\mathsf{F}_h[u_h](z) = 0 \ \text{ in } \bar{\Omega}_h, \tag{3.3}$$

where F_h is a locally bounded operator which we may think of as an approximation to F, and X_h is some finite-dimensional space. The operator is parametrized by $h > 0$, which we may view as a discretization parameter, or in some cases, a regularization parameter. The discrete domain $\overline{\Omega}_h$ is an approximation to $\overline{\Omega}$ with the property $\lim_{h\to 0^+} \bar{\Omega}_h = \bar{\Omega}$; namely, for all $z_0 \in \bar{\Omega}$, there exists a sequence $\{z_h\}_{h>0} \subset \bar{\Omega}_h$ such that $\lim_{h\to 0^+} z_h = z_0$.

We now address the well-posedness of (3.3), and the sufficient structure conditions on F_h to ensure that the discrete solutions to (3.3) (if they exist) converge. As a first step we state the fundamental notions of consistency and stability.

Definition 3.1 (consistency). The discrete problem (3.3) is said to be *consistent* with (3.2) if there exists an operator $I_h : C(\bar{\Omega}) \to X_h$ such that I_h converges uniformly to the identity operator as $h \to 0^+$, and for all sequences $\{z_h\}_{h>0}$ with $z_h \in \bar{\Omega}_h$ and $z_h \to z_0 \in \bar{\Omega}$ and $\phi \in C^2(\bar{\Omega})$,

$$\lim_{h\to 0^+} \mathsf{F}_h[I_h\phi](z_h) = \mathsf{F}[\phi](z_0).$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Remark 3.2 (envelopes). If the operators are not continuous, then the notion of consistency is changed to

$$\limsup_{h\to 0^+} \mathsf{F}_h[I_h\phi](z_h) \le \mathsf{F}^*[\phi](z_0),$$
$$\liminf_{h\to 0^+} \mathsf{F}_h[I_h\phi](z_h) \ge \mathsf{F}_*[\phi](z_0),$$

where F^* (resp. F_*) denote the upper (resp. lower) semi-continuous envelope of F; see Definition 2.47.

Definition 3.3 (stability). We say that problem (3.3) is *stable* if, for all $h > 0$, there exists a solution $u_h \in X_h$ to (3.3), and moreover, if $w_h \in X_h$ satisfies $\mathsf{F}_h[w_h] = \epsilon_h$, then $\|u_h - w_h\|_{L^\infty(\bar\Omega_h)} \le C\|\epsilon_h\|_{L^\infty(\bar\Omega_h)}$, with $C > 0$ independent of h.

The following theorem states the well-known result that, for linear problems with classical solutions, consistent and stable schemes converge.

Theorem 3.4 (Lax–Richtmyer). Suppose that F_h is an affine operator, and that there exists a classical solution $u \in C^2(\bar\Omega)$ satisfying (3.2). Suppose further that problem (3.3) is stable and that the operator in Definition 3.1 satisfies the stronger condition $\lim_{h\to 0^+} \|\mathsf{F}_h[I_h u]\|_{L^\infty(\bar\Omega_h)} = 0$. Then u_h converges locally uniformly to u.

Proof. By the given assumptions, there exists a linear operator G_h and a function I_h such that

$$\mathsf{F}_h[v_h](z) = \mathsf{G}_h[v_h](z) - \mathsf{I}_h(z) \quad z \in \bar\Omega_h.$$

The stability of the scheme shows that $\mathsf{G}_h[\cdot]$ is an isomorphism whose inverse is bounded independent of h. Since

$$\mathsf{G}_h[u_h - I_h u] = \mathsf{I}_h - \mathsf{G}_h[I_h u] = -\mathsf{F}_h[I_h u],$$

the consistency of the scheme implies that $\lim_{h\to 0^+} \|u_h - I_h u\|_{L^\infty(\bar\Omega_h)} = 0$, and thus, since I_h converges to the identity operator, u_h converges locally uniformly to u. □

While Theorem 3.4 is a useful result for a large class of problems, it is not applicable to fully nonlinear problems nor to weaker notions of solutions, in particular, viscosity solutions. The issue is that, if u is not a classical solution, then the approximation $I_h u$ may not be well-defined, and the consistency $\mathsf{F}_h[I_h u] \to 0$ used in the proof of Theorem 3.4 is no longer valid. Rather, to prove convergence to viscosity solutions, an additional structure condition is required. This requirement is summarized in the following definition.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Definition 3.5 (monotone operator). The discrete operator F_h is said to be monotone if, whenever $u_h - v_h$ has a global non-negative maximum at $z \in \bar{\Omega}_h$, we have

$$\mathsf{F}_h[u_h](z) \leq \mathsf{F}_h[v_h](z). \tag{3.4}$$

Remark 3.6 (monotonicity). The notion of monotonicity is essentially a discrete version of ellipticity. Indeed, following the proof of Theorem 2.16, if F is an elliptic operator in the sense of Definition 2.10, and if $u - v$, with $u, v \in C^2(\bar{\Omega})$, has a global non-negative maximum at $x \in \bar{\Omega}$, then $u(x) \geq v(x)$, $Du(x) = Dv(x)$ and $D^2u(x) \leq D^2v(x)$. Since F is non-increasing in its second argument, and non-decreasing in its fourth, we have

$$\mathsf{F}[u](x) = \mathsf{F}(x, u(x), Du(x), D^2u(x)) \leq \mathsf{F}(x, v(x), Dv(x), D^2v(x)) = \mathsf{F}[v](x).$$

Conversely, it can be shown that the operator F is elliptic if, whenever $u - v$ has a global non-negative maximum at $x \in \bar{\Omega}$, then $\mathsf{F}[u](x) \leq \mathsf{F}[v](x)$.

Consistency, stability and monotonicity are the three sufficient ingredients to guarantee convergence to viscosity solutions. The following result closely follows Barles and Souganidis (1991, Theorem 2.1).

Theorem 3.7 (Barles–Souganidis). Suppose that problem (3.3) is consistent, stable and monotone in the sense of Definitions 3.1, 3.3 and 3.5, respectively. Suppose further that F satisfies the comparison principle given in Definition 2.49. Then u_h converges locally uniformly to the unique continuous viscosity solution of (3.2).

Proof. Define $\bar{u} \in USC(\Omega)$ and $\underline{u} \in LSC(\Omega)$ by

$$\bar{u}(x) := \limsup_{\substack{y \to x \\ h \to 0^+}} u_h(y), \quad \underline{u}(x) := \liminf_{\substack{y \to x \\ h \to 0^+}} u_h(y), \quad x \in \bar{\Omega}. \tag{3.5}$$

Note that the stability of the scheme implies that both $\bar{u}$ and $\underline{u}$ are well-defined. The proof proceeds by showing that $\bar{u}$ and $\underline{u}$ are (viscosity) subsolutions and supersolutions to (3.2), respectively, and then appealing to the comparison principle.

To this end, suppose that $z_0 \in \Omega$ is a strict local maximum of $\underline{u} - \phi$ for some $\phi \in C^2(\Omega)$. Then standard arguments show that there exist sequences $\{h_n\}_{n=1}^{\infty}$ and $\{z_{h_n}\}_{n=1}^{\infty}$ such that

$$h_n \to 0, \quad z_{h_n} \to x_0, \quad \text{as } n \to \infty,$$

and $u_{h_n} - I_{h_n}\phi$ obtains a strict local maximum at z_{h_n}. Since $\mathsf{F}_{h_n}[u_{h_n}](z_{h_n}) = 0$, the monotonicity of the discrete operator implies that

$$\mathsf{F}_{h_n}[I_{h_n}\phi](z_{h_n}) \geq 0.$$

Passing to the limit, together with consistency of the scheme, yields

$$0 \leq \lim_{n \to \infty} \mathsf{F}_{h_n}[I_{h_n}\phi](z_{h_n}) = \mathsf{F}[\phi](z_0).$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Similar arguments show that if $\bar{u} - \phi$ obtains a strict minimum at $z_0 \in \Omega$, then $0 \geq \mathsf{F}[\phi](z_0)$. Thus, $\bar{u}$ and $\underline{u}$ are subsolutions and supersolutions to (3.2), respectively. Since $\underline{u} \leq \bar{u}$, the comparison principle of F implies that $\underline{u} = \bar{u} = u$, and u is the viscosity solution to (3.2). □

We once again mention that the problems and discretizations considered so far take into account the boundary conditions in a viscosity sense. While this set-up simplifies the proof of convergence, it may have practical limitations since the framework requires a consistent and monotone discretization of both the boundary conditions and the differential operator for all $x \in \partial\Omega$ and smooth functions ϕ. Also recall that, in general, viscosity boundary conditions are not equivalent to those imposed pointwise unless other conditions are assumed; see Proposition 2.55. Here we turn our attention to the elliptic boundary value problem (2.6), where the Dirichlet boundary condition is understood in the classical sense; see Definition 2.48.

To this end, we consider approximations of the form

$$F_h[u_h] = 0 \ \text{ in } \Omega_h^I, \quad u_h = g_h \ \text{ on } \Omega_h^B, \tag{3.6}$$

where $g_h = I_h g \in X_h$ is a discrete approximation to the Dirichlet data $g \in C(\bar{\Omega})$, and Ω_h^I and Ω_h^B are disjoint sets with $\Omega_h^I \to \Omega$ and $\Omega_h^B \to \partial\Omega$ as $h \to 0^+$. Note that with minor notational changes, the notions of consistency, stability, and monotonicity are applicable to the operator F_h. A natural question is whether the results of Theorem 3.7 carry over to the discrete problem (3.6). This issue is addressed in the next theorem.

Theorem 3.8 (convergence). Suppose that F_h is a consistent, stable and monotone operator. Suppose further that either

(i) $\underline{u}(x) \geq g(x)$ and $\bar{u}(x) \leq g(x)$ for all $x \in \partial\Omega$, where $\underline{u}, \bar{u}$ are given by (3.5), or

(ii) the sequence of solutions $\{u_h\}_{h>0}$ is equicontinuous.

Then u_h converges locally uniformly to the unique continuous viscosity solution of (3.1) with $B(x, u, Du) = g - u$.

Proof. The proof of case (i) follows directly from the arguments given in Theorem 3.7. The proof of case (ii) follows from the Arzelà–Ascoli theorem, and again appealing to the proof of Theorem 3.7. □

3.2. Monotonicity in finite difference schemes

Here we discuss basic monotonicity results of finite difference schemes. The main message given in this section is that for any uniformly elliptic operator, one can construct a consistent and monotone finite difference scheme. The drawback, however, is that monotonicity requires a wide stencil, which may severely impact its practical use. Much of the material in this section

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

is found in Kuo and Trudinger (1990, 1992), Kocan (1995), Motzkin and Wasow (1953) and Oberman (2006).

For simplicity we assume that the domain Ω is discretized on an equally spaced Cartesian grid and that each coordinate direction is discretized uniformly; in particular, by a possible change of coordinates, we assume that the grid is given by

$$\bar{\Omega}_h = \mathbb{Z}_h^d \cap \overline{\Omega}, \quad \text{with } \mathbb{Z}_h^d := \{he : e \in \mathbb{Z}^d\},$$

where $h > 0$ is the grid scale and $\mathbb{Z}^d$ is the set of d-tuples of integers. A finite subset $S \subset \mathbb{Z}^d \backslash \{0\}$ is called a *stencil*, and the space of *nodal functions*, denoted by X_h^{fd}, consists of real-valued functions with domain $\bar{\Omega}_h$. The canonical interpolant $I_h^{fd} : C^0(\bar{\Omega}) \to X_h^{fd}$ is the operator satisfying $I_h^{fd} v(z) = v(z)$ for all $z \in \bar{\Omega}_h$. We assume the existence of a positive integer m such that S is of the form

$$S = \{y : y \in \mathbb{Z}^d \backslash \{0\} : |y|_{\ell^\infty} \leq m\}. \tag{3.7}$$

The value m satisfying (3.7) is called the *stencil size* of S. The cardinality of S is $|S| := (2m+1)^d - 1$.

We consider finite difference schemes with stencil S acting on grid functions. These discrete operators are thus of the (implicit) form

$$F_h[v_h](z) = F_h(z, v_h(z), Tv_h(z)), \tag{3.8}$$

where $Tv_h(x) = \{v_h(x + hy) : y \in S\}$ is the set of translates of $v_h(x)$ with respect to the stencil. The method (3.8) is called a *one-step* scheme if $m = 1$, that is, the value $F_h[v_h](z)$ depends only on z, $v_h(z)$, and the values of v_h at neighbouring points of z. Otherwise, we call the scheme a *wide-stencil* scheme if $m \geq 2$.

To construct monotone schemes, we first reformulate this property so that it is easier to work with.

Definition 3.9 (non-negative operator). The operator F_h is of *non-negative type* (or simply non-negative) if

$$F_h(z, r, q + \tau) \geq F_h(z, r, q) \geq F_h(z, r + t, q + \tau), \tag{3.9}$$

for all $z \in \mathbb{R}^d$, $r, t \in \mathbb{R}$, and $q, \tau \in \mathbb{R}^{|S|}$ satisfying

$$0 \leq \tau_i \leq t \quad i = 1, 2, \ldots, |S|. \tag{3.10}$$

We see that if F_h is of non-negative type, then F_h is non-increasing in its second argument and non-decreasing in its third argument. If F_h is differentiable, then it is of non-negative type provided that

$$\frac{\partial F_h}{\partial q_i} \geq 0 \ (i = 1, 2, \ldots, |S|), \quad \frac{\partial F_h}{\partial r} + \sum_{i=1}^{|S|} \frac{\partial F_h}{\partial q_i} \leq 0.$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Remark 3.10 (reformulation). Alternatively, as in Oberman (2006), one can consider finite difference schemes of the form

$$F_h[u](z) = G_h(z, u(z), u(z) - Tu(z)).$$

Let $\mathbf{1} \in \mathbb{R}^{|S|}$ denote the vector of ones. Then using the correspondence $F_h(z, r, q) = G_h(z, r, r\mathbf{1} - q)$, one sees that F_h is of non-negative type if and only if G_h is non-increasing in its second and third arguments.

Let us now show that non-negativity is simply a reformulation of monotonicity.

Lemma 3.11 (equivalence). A finite difference scheme of the form (3.8) is monotone if and only if it is of non-negative type.

Proof. Suppose that F_h is of non-negative type. Let u_h and v_h be two grid functions such that $u_h - v_h$ has a global non-negative maximum at some grid point z. Set $r = v_h(z)$, $t = u_h(z) - v_h(z) \geq 0$, $q_i = v_h(z + hy_i)$, and $\tau_i = \max\{0, u_h(z + hy_i) - v_h(z + hy_i)\}$, so that $t \geq \tau_i$. Noting that $q_i + \tau_i = v_h(z + hy_i) + \max\{0, u_h(z + hy_i) - v_h(z + hy_i)\} \geq u_h(z + hy_i)$, and F_h is non-decreasing in its third argument, we find that $F_h[u_h](z) \leq F_h(z, r + t, q + \tau)$. Therefore, by the second inequality in (3.9) we have

$$F_h[u_h](z) \leq F_h(z, r + t, q + \tau) \leq F_h(z, r, q) = F_h[v_h](z).$$

Thus, F_h is monotone.

Now suppose that F_h is monotone. Let $z \in \mathbb{R}^d$ be fixed, and let $r, t \in \mathbb{R}$ and $q, \tau \in \mathbb{R}^{|S|}$ satisfy (3.9). Then define the grid functions u_h, v_h (locally) as

$$v_h(z) = r, \quad v_h(z + hy_i) = q_i, \quad u_h(z) = r + t, \quad u_h(z + y_i) = q_i + \tau_i.$$

Then

$$u_h(z) - v_h(z) = t \geq \tau_i = u_h(z + hy_i) - v_h(z + hy_i),$$

that is, $u_h - v_h$ has a non-negative maximum at z. The monotonicity of F_h yields $F_h[u_h](z) \leq F_h[v_h](z)$; thus

$$F_h(z, r + t, q + \tau) \leq F_h(z, r, q). \tag{3.11}$$

On the other hand, with v_h as before, we consider the grid function w_h with

$$w_h(z) = r, \quad w_h(z + hy_i) = q_i + \tau_i.$$

Then $v_h - w_h$ has a global maximum at z and thus

$$F_h(z, r, q) = F_h[v_h](z) \leq F_h[w_h](z) = F_h(z, r, q + \tau). \tag{3.12}$$

We conclude from (3.11)–(3.12) that F_h is of non-negative type. □

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Following the framework given in Kuo and Trudinger (1990, 1992) and Kocan (1995), we consider discrete operators constructed from the first- and second-order difference operators

$$\begin{aligned}
\delta^+_{y,h}u(z) &:= \frac{1}{h}(u(z+hy) - u(z)),\\
\delta^-_{y,h}u(z) &:= \frac{1}{h}(u(z) - u(z-hy)),\\
\delta_{y,h}u(z) &:= \frac{1}{2}(\delta^+_y + \delta^-_y)u(z) = \frac{1}{2h}(u(z+hy) - u(z-hy)),\\
\delta^2_{y,h}u(z) &:= \frac{1}{h^2}(u(z+hy) - 2u(x) + u(z-hy)),
\end{aligned}$$

with $y \in S$. Taylor's theorem shows that the differences $\delta^{\pm}_{y,h}u(z)$ are first-order approximations to

$$\frac{\partial u(z)}{\partial y} := Du(z)\cdot y,$$

whereas $\delta_{y,h}u(z)$ and $\delta^2_{y,h}u(z)$, respectively, are second-order approximations to

$$\frac{\partial u(z)}{\partial y} \quad \text{and} \quad \frac{\partial^2 u(z)}{\partial y^2} := y\cdot D^2u(z)y;$$

by this we mean that

$$\begin{aligned}
\left|\frac{\partial^{\pm}u(z)}{\partial y} - \delta^{\pm}_{y,h}u(z)\right| &= O(h|y|^2),\\
\left|\frac{\partial u(z)}{\partial y} - \delta_{y,h}u(z)\right| &= O(h^2|y|^3),\\
\left|\frac{\partial^2 u(z)}{\partial y^2} - \delta^2_{y,h}u(z)\right| &= O(h^2|y|^4),
\end{aligned}$$

for sufficiently smooth u.

Let

$$\delta_h u_h(z) = \{\delta_{y,h}u_h(z) : y \in S\} \quad \text{and} \quad \delta^2_h u_h(z) = \{\delta^2_{y,h}u_h(z) : y \in S\}.$$

Then a consistent and monotone finite difference scheme can be constructed in the form

$$F_h[u_h](z) = \mathcal{F}_h(z, u_h(z), \delta_h u_h(z), \delta^2_h u_h(z)) \tag{3.13}$$

(Kuo and Trudinger 1990, Kuo and Trudinger 1992, Motzkin and Wasow 1953), where $\mathcal{F}_h : \Omega_h \times \mathbb{R} \times \mathbb{R}^{|S|} \times \mathbb{R}^{|S|} \to \mathbb{R}$. Let (z, r, q, s) denote points in the domain of $\mathcal{F}_h$, and assume that $\mathcal{F}_h$ is symmetric with respect to $\pm q_{\pm i}$ and $s_{\pm i}$. Then from Definition 3.9 and Lemma 3.11 we see that F_h of the

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

form (3.13) is monotone provided that

$$\frac{h}{2}\left|\frac{\partial \mathcal{F}_h}{\partial q_i}\right| \leq \frac{\partial \mathcal{F}_h}{\partial s_i} \quad i = 1, 2, \ldots, |S|, \quad \text{and} \quad \frac{\partial \mathcal{F}_h}{\partial r} \leq 0. \tag{3.14}$$

Below, we will require slightly stronger conditions on the operator $\mathcal{F}_h$.

Definition 3.12 (positive operator). An operator of the form (3.13) is of *positive type* (or simply positive) if (3.14) is satisfied and there exists $\lambda_{0,h} > 0$ and an orthogonal set of vectors $\{y_i\}_{i=1}^d \subset S$ such that

$$\lambda_{0,h} + \frac{h}{2}\left|\frac{\partial \mathcal{F}_h}{\partial q_i}\right| \leq \frac{\partial \mathcal{F}_h}{\partial s_i}.$$

Remark 3.13 (discrete ellipticity). The discrete ellipticity constant $\lambda_{0,h}$ may depend on the stencil size m; see Theorem 3.36.

3.3. Finite difference stability estimates: Alexandrov estimates and Alexandrov–Bakelman–Pucci maximum principle

We now turn our attention to maximum principles of discrete operators, and correspondingly, stability estimates. As a starting point, we discuss monotone finite difference schemes for the linear non-divergence-form PDEs of Example 2.11

$$F[u] = \mathcal{L}u - f = A : D^2 u - f = 0. \tag{3.15}$$

While this setting may seem overly simplistic, as we shall see, the construction and theoretical results for the linear problem form all of the necessary tools to approximate viscosity solutions of nonlinear elliptic equations.

We assume that $f \in C(\bar{\Omega})$ and that the coefficient matrix A is bounded and uniformly symmetric positive definite. It is then reasonable to assume that F_h is linear and thus has the form

$$F_h[u_h](z) = \mathcal{L}_h u_h(z) - f(z) := \sum_{y \in S} a_y(z) \delta_{y,h}^2 u_h(z) - f(z) \tag{3.16}$$

for nodal functions (or coefficients) a_y. Applying Definition 3.9 to (3.16), we see that F_h is non-negative (and hence monotone) provided

$$a_y(z) \geq 0, \tag{3.17}$$

and of positive type if

$$a_{y_i}(z) \geq \lambda_{0,h} > 0 \quad i = 1, 2, \ldots, d.$$

for some orthogonal basis $\{y_i\}_{i=1}^d \subset S$. These inequalities suggest that the negation of the ensuing system is an M-matrix, and hence solutions to the discrete problem satisfy certain maximum principles, analogous to the continuous setting. This issue is discussed in the next section.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

3.3.1. Finite difference Alexandrov estimates

In this section we state and prove ABP maximum principles for grid functions. To get started, we first specify the fundamental notion of interior and boundary nodes used in this section.

Definition 3.14 (interior and boundary nodes). For a discrete operator F_h, we define the *set of interior nodes* Ω_h^I as the set of grid points $z \in \bar{\Omega}_h$ such that for any mesh function v_h, $F_h[v_h](z)$ depends only on the translates of v_h at points in $\bar{\Omega}_h$. The *set of boundary nodes* are given by $\Omega_h^B := \bar{\Omega}_h \backslash \Omega_h^I$.

Remark 3.15 (discrete domain). If F_h is a one-step method (*i.e.* $m = 1$), then $\Omega_h^B = \partial\Omega \cap \mathbb{Z}_h^d$ and $\Omega_h^I = \Omega \cap \mathbb{Z}_h^d$.

As a next step we introduce and discuss several basic properties of convexity and the subdifferential for discrete (nodal) functions.

Definition 3.16 (convex nodal function). We say that a nodal function $v_h \in X_h^{fd}$ is a *convex nodal function* if there is a supporting hyperplane of v_h at all interior nodes $z \in \Omega_h^I$.

Note that if v_h is the nodal interpolant of a convex function, then v_h is a convex nodal function.

Definition 3.17 (discrete convex envelope). Let $R > 0$ be sufficiently large so that $\bar{\Omega}$ (and hence $\bar{\Omega}_h$) is compactly contained in a ball B_R. For a nodal function (or continuous function) v_h with $v_h \geq 0$ on Ω_h^B, we extend v_h^- to $B_{R,h} \backslash \bar{\Omega}_h$ by zero, where $B_{R,h} = B_R \cap \mathbb{Z}_h^d$, and $v_h^- = -\max\{0, -v_h\}$ is the negative part of v_h. We define the *discrete convex envelope* of $-v_h^-$ as

$$\Gamma_h(v_h)(x) := \sup\{L(x) : L(z) \leq -v_h^-(z) \ \forall z \in B_{R,h},\ L \in \mathbb{P}_1\} \tag{3.18}$$

for all $x \in \bar{B}_R$.

Remark 3.18 (discrete convexity). There are subtle issues in the above definitions that require some elaboration.

- If $v_h \in X_h^{fd}$ is convex and $v_h \leq 0$, then we have
$$v_h(z) = \Gamma_h(v_h)(z) \quad \text{for all } z \in \Omega_h^I. \tag{3.19}$$
Thus, $\Gamma_h(v_h)$ is a natural convex extension of v_h.
- Since for every $x \in \partial B_R$, there exists an affine function L with $L(z) \leq -v_h^-(z)$ for all $z \in \bar{\Omega}_h$ and $L(x) = 0$, we conclude that $\Gamma_h(v_h) = 0$ on ∂B_R.
- Definition 3.17 implies that $\Gamma_h(v_h)$ is a convex, piecewise linear function with respect to a simplicial triangulation. The vertices of the triangulation are a subset of the gridpoints $B_{R,h}$, and its configuration depends on v_h; see Examples 3.44–3.45.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

- For $v_h \in X_h^{fd}$, let $\tilde{v}_h \in C(\bar{\Omega})$ denote the canonical multi-linear function. Then, since the inequality constraints in (3.18) are only posed on a discrete set of points, and since $\tilde{v}_h$ is not necessarily piecewise affine, we have $\Gamma(\tilde{v}_h) \neq \Gamma_h(v_h)$ in general (Carnicer and Dahmen 1994).

Next, we require the notion of a subdifferential acting on nodal functions. Recall from Definition 2.58 that the subdifferential requires function values at all points of the domain Ω, so this notion is not directly applicable to the discrete case. Instead, with a slight abuse of notation, we define its natural extension to nodal functions as follows:

$$\partial v_h(z) = \big\{\mathbf{p} \in \mathbb{R}^d : v_h(x) - v_h(z) \geq \mathbf{p} \cdot (x - z),\ \forall x \in \bar{\Omega}_h\big\}, \tag{3.20}$$

for all $z \in \bar{\Omega}_h$ and $v_h \in X_h^{fd}$.

Lemma 3.19 (discrete subdifferential). If $v_h \in X_h^{fd}$ is a convex nodal function, then $\partial v_h(z) = \partial \Gamma_h(v_h)(z)$ for all $z \in \Omega_h^I$.

Proof. Thanks to (3.19), if $\mathbf{p} \in \partial \Gamma_h(v_h)(z)$, that is,

$$\Gamma_h(v_h)(x) \geq \Gamma_h(v_h)(z) + \mathbf{p} \cdot (x - z), \quad \text{for all } x \in \Omega,$$

then $\mathbf{p} \in \partial v_h(z)$.

Conversely, let $\mathcal{T}_z$ be a local mesh induced by $\Gamma(u_h)(z)$, let $K \in \mathcal{T}_z$ be a d-dimensional simplex, and let $\{z_j\}_{j=1}^{d+1}$ be the vertices of K. If $\mathbf{p} \in \partial v_h(z)$, then we clearly have $v_h(x) \geq v_h(z) + \mathbf{p} \cdot (x - z)$ for all vertices $x \in \bar{\Omega}_h$. Again, thanks to (3.19), we have $\Gamma(v_h)(z_i) \geq \Gamma(v_h)(z) + \mathbf{p} \cdot (z_i - z)$ for all vertices. Since $\Gamma(v_h)$ is linear on element K, we have $\Gamma(v_h)(x) \geq \Gamma(v_h)(z) + \mathbf{p} \cdot (x - z)$ for any $x \in K$. This shows that $\mathbf{p} \in \partial \Gamma(u_h)(z)$ as well. □

Let us state two properties of subdifferentials. The proof of the first one follows directly from its definition.

Lemma 3.20 (monotonicity of subdifferential). Let $w_h, v_h \in X_h^{fd}$ be two convex nodal functions such that, for a fixed $z_* \in \bar{\Omega}_h$, $w_h(z_*) = v_h(z_*)$ and $w_h(z) \leq v_h(z)$ for all $z \neq z_*$. Then,

$$\partial w_h(z_*) \subset \partial v_h(z_*).$$

Lemma 3.21 (addition inequality). Let w_h and v_h be two convex nodal functions. Then

$$\partial w_h(z) + \partial v_h(z) \subset \partial (w_h + v_h)(z), \quad \text{for all } z \in \Omega_h^I,$$

where $\partial w_h(z) + \partial v_h(z)$ is the Minkowski sum:

$$\partial w_h(z) + \partial v_h(z) = \{\mathbf{p} + \mathbf{q} \in \mathbb{R}^d : \mathbf{p} \in \partial w_h(z), \mathbf{q} \in \partial v_h(z)\}.$$

Proof. We note that if $\mathbf{p} \in \partial w_h(z)$ and $\mathbf{q} \in \partial v_h(z)$, then

$$w_h(x) \geq w_h(z) + w \cdot (x - z) \quad \text{and} \quad v_h(x) \geq v_h(z) + v \cdot (x - z)$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

for all $x \in \bar{\Omega}_h$. Adding the two inequalities yields

$$w_h(x) + v_h(x) \geq w_h(z) + v_h(z) + (\mathbf{p} + \mathbf{q}) \cdot (x - z),$$

which implies that $(\mathbf{p} + \mathbf{q}) \in \partial(w_h + v_h)$. □

Given a convex nodal function u_h, computing its discrete subdifferential set is a non-trivial task. The following lemma shows that it involves computing the convex envelope of u_h.

Lemma 3.22 (characterization of subdifferential). Let u_h be a convex nodal function, and let $\mathcal{T}_h$ be the simplicial mesh induced by its convex envelope. The subdifferential of u_h at z is the convex hull of the piecewise gradient, that is,

$$\operatorname{conv}\{Du_h|_K, K \in \mathcal{T}_h,\ z \in \bar{K}\}.$$

Here, Du_h is the gradient of the piecewise linear polynomial induced by u_h and $\mathcal{T}_h$.

As final preparation for stating the finite difference version of the Alexandrov estimate, we define the nodal contact set.

Definition 3.23 (nodal contact set). Let v_h be either a nodal function or a continuous function with $v_h \geq 0$ on Ω_h^B. The (*lower*) *nodal contact set* of v_h is given by

$$\mathcal{C}_h^-(v_h) = \{z \in \Omega_h^I : \Gamma_h(v_h)(z) = v_h(z)\}.$$

Note that, for $x \in \mathcal{C}_h^-(v_h)$, we have

$$v_h(z) \geq \Gamma_h(v_h)(z) \geq v_h(x) + \mathbf{p} \cdot (z - x),$$

for all $z \in B_{R,h}$ and all $\mathbf{p} \in \partial\Gamma_h(v_h)(x)$.

We are now ready to state and prove the finite difference Alexandrov estimate. Recall that we assume Ω is compactly contained in a ball B_R of radius R, and that we set $B_{R,h} = B_R \cap \mathbb{Z}_h^d$.

Lemma 3.24 (finite difference Alexandrov estimate). Let $v_h \in X_h^{fd}$ with $v_h \geq 0$ on Ω_h^B. Then

$$\sup_{\bar{\Omega}_h} v_h^- \leq CR\Bigg(\sum_{z \in \mathcal{C}_h^-(v_h)} |\partial\Gamma_h(v_h)(z)|\Bigg)^{1/d}, \tag{3.21}$$

where the constant $C > 0$ depends only on d.

Proof. We follow the arguments given in Nochetto and Zhang (2017*a*, Proposition 5.1); see also Kuo and Trudinger (1990).

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Let $z_* \in B_{R,h}$ satisfy $\sup_{B_{R,h}} v_h^- = v_h^-(z_*)$, and let L be a horizontal plane touching v_h from below at z_*. By Definition 3.17 we have

$$\Gamma_h(v_h)(z) \geq L(z) = L(z_*) = v_h(z_*), \quad \text{for all } z \in B_{R,h}.$$

Thus, $\sup_{B_{R,h}} \Gamma_h(v_h)^- \leq v_h^-(z_*)$. Since $\Gamma_h(v_h) \leq v_h$ on $B_{R,h}$ implies

$$\sup_{B_{R,h}} v_h^- \leq \sup_{B_{R,h}} \Gamma_h(v_h)^-,$$

we conclude that

$$\sup_{\bar{\Omega}_h} v_h^- = \sup_{B_{R,h}} v_h^- = \sup_{B_{R,h}} \Gamma_h(v_h)^- = \sup_{B_R} \Gamma_h(v_h)^-.$$

Therefore, to conclude the proof it suffices to show that

$$\max_{\bar{\Omega}_h} \Gamma_h(v_h)^- \leq CR\left(\sum_{z \in \mathcal{C}_h^-(v_h)} |\partial \Gamma_h(v_h)(z)| \right)^{1/d}.$$

This is done in three steps.

Step 1. Let $K(x)$ be the cone with vertex z_* satisfying

$$K(z_*) = -\sup_{B_R} \Gamma_h(v_h)^- =: -M \quad \text{and} \quad K(x) = 0 \ \text{ on } \partial B_R,$$

and assume that $M > 0$, for otherwise (3.21) is trivial. We note that for any vector $\mathbf{p} \in B_{M/(2R)}(0)$, the affine function $L(x) = -M + \mathbf{p} \cdot x - z_*$ is a supporting plane of $K(x)$ at point z_*, namely $L(x) \leq K(x)$ for all $x \in B_R$ and $L(z_*) = K(z_*)$. This implies that $\partial K(z_*) \supset B_{M/(2R)}(0)$, and therefore

$$|\partial K(z_*)| \geq C \left(\frac{M}{R} \right)^d.$$

Step 2. We claim that

$$\partial K(z_*) \subset \bigcup_{z \in \mathcal{C}_h^-(v_h)} \partial \Gamma_h(v_h)(z). \tag{3.22}$$

This is equivalent to showing that for any supporting plane L of K at z_*, there is a parallel supporting plane $\tilde{L}$ for $\Gamma_h(v_h)$ at some contact node $y \in \mathcal{C}_h^-(v_h)$.

Consider the (nodal) function $v_h - L$, and observe that $v_h \geq 0$ on Ω_h^B and $v_h(z_*) = K(z_*) = L(z_*)$, whence

$$\begin{aligned} v_h(z) - L(z) &\geq K(z) - L(z) \geq 0 \ \text{ on } \Omega_h^B, \\ v_h(z_*) - L(z_*) &= K(z_*) - L(z_*) = 0. \end{aligned}$$

We infer that $v_h - L$ attains a non-positive minimum for some $x \in \Omega_h^I$. Hence, $\widetilde{L}(z) = L(z) + v_h(x) - L(x)$ satisfies $\tilde{L}(z) \leq v_h(z)$ for all $z \in B_{R,h}$ and

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

$\tilde{L}(x) = v_h(x)$. Applying Definition 3.17 we conclude that $\tilde{L} \le \Gamma_h(v_h) \le v_h$ and therefore $\Gamma_h(v_h)(x) = v_h(x)$; thus $x \in \mathcal{C}_h^-(v_h)$.

Step 3. Computing Lebesgue measures in (3.22) yields

$$C\left(\frac{M}{R}\right)^d \le |\partial K(z_*)| \le \sum_{z \in \mathcal{C}_h^-(v_h)} |\partial \Gamma_h(v_h)(z)|.$$

Finally, (3.44) follows from this last inequality and some simple algebraic manipulation. □

The following theorem states that positive finite difference operators satisfy a discrete Alexandrov–Bakelman–Pucci estimate (see Theorem 2.62 and Kuo and Trudinger 1990).

Theorem 3.25 (finite difference ABP estimate). Suppose that $\mathcal{L}_h$ is of positive type and of the form (3.16). Suppose that $u_h \in X_h^{fd}$ satisfies

$$\begin{cases} \mathcal{L}_h u_h \le f & \text{in } \Omega_h^I, \\ u_h = g_h & \text{on } \Omega_h^B \end{cases} \tag{3.23}$$

for some $g_h \in X_h^{fd}$. Then

$$\sup_{\bar{\Omega}_h} u_h^- \le \sup_{\Omega_h^B} g_h^- + C \frac{R}{\lambda_{0,h}} \left(\sum_{z \in \mathcal{C}_h^-(u_h)} h^d (f^+(z))^d \right)^{1/d},$$

where $\lambda_{0,h}$ is given in Definition 3.12, and the constant $C > 0$ depends only on d.

Proof. We follow the arguments given in Kuo and Trudinger (1990, Theorem 2.1).

Note that we can assume that $u_h \ge 0$ on Ω_h^B, by replacing u_h with $u_h + \max_{\Omega_h^B} g_h^-$. Moreover we can assume that $\max_{\bar{\Omega}_h} u_h^- > 0$, since otherwise the proof is trivial.

Let $z \in \mathcal{C}_h^-(u_h)$ and $y \in S$. As $\Gamma_h(u_h)$ is convex, we have $\delta_{y,h}^2 \Gamma_h(u_h)(z) \ge 0$. Thus, since the coefficients of F_h are positive, $\Gamma_h(u_h)(z) = u_h(z)$ and $\Gamma_h(u_h)(z \pm hy) \le u_h(z \pm hy)$, we have

$$0 \le a_y(z) \delta_{y,h}^2 \Gamma_h(u_h)(z) \le a_y(z) \delta_{y,h}^2 u_h(z).$$

Summing over $y \in S$ yields

$$0 \le a_y(z) \delta_{y,h}^2 \Gamma_h(u_h)(z) \le \sum_{y' \in S} a_{y'}(z) \delta_{y',h}^2 u_h(z) \le f(z) = f^+(z).$$

Take y to be the orthogonal set $\{y_i\}_{i=1}^d$ given in Definition 3.12. Expand

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

the left-hand side of the above inequality to get

$$\delta^+_{y_i,h}\Gamma_h(u_h)(z) - \delta^-_{y_i,h}\Gamma_h(u_h)(z) = h\delta^2_{y_i,h}\Gamma_h(u_h)(z) \leq \frac{h}{\lambda_{0,h}} f^+(z). \tag{3.24}$$

Now, let $\mathbf{p} \in \partial\Gamma_h(u_h)(z)$ so that $\Gamma_h(u_h)(z \pm hy_i) \geq \Gamma_h(u_h)(z) \pm h\mathbf{p}\cdot y_i$. By manipulating terms and applying inequality (3.24) we obtain

$$\delta^-_{y_i,h}\Gamma_h(u_h)(z) \leq \mathbf{p}\cdot y_i \leq \delta^+_{y_i,h}\Gamma_h(u_h)(z) \leq \delta^-_{y_i,h}\Gamma_h(u_h)(z) + \frac{h}{\lambda_{0,h}} f^+(z).$$

Since $\{y_i/|y_i|\}_{i=1}^d$ is an orthonormal basis of $\mathbb{R}^d$, these two inequalities show that the Lebesgue measure of $\partial\Gamma_h(u_h)(z)$ is bounded by

$$|\partial\Gamma_h(u_h)(z)| \leq \frac{h^d}{\lambda^d_{0,h}} |f^+(z)|^d,$$

and therefore

$$\sum_{z\in\mathcal{C}_h^-(u_h)} |\partial\Gamma_h(u_h)(z)| \leq \sum_{z\in\mathcal{C}_h^-(u_h)} \frac{(hf^+(z))^d}{\lambda^d_{0,h}}. \tag{3.25}$$

Combining (3.25) and Lemma 3.24 yields the desired result. □

Remark 3.26 (extensions). The finite difference ABP estimate given in Theorem 3.25 has been extended to operators with lower-order terms and to general meshes in Kuo and Trudinger (1996, 2000).

Theorem 3.25 implies that if u_h solves

$$\begin{cases} \mathcal{L}_h u_h = f & \text{in } \Omega_h^I, \\ u_h = g_h & \text{on } \Omega_h^B, \end{cases} \tag{3.26}$$

then

$$\max_{\bar\Omega_h} |u_h| \leq \max_{\Omega_h^B} |g_h| + \frac{CR}{\lambda_{0,h}} \left(\sum_{z\in\Omega_h^I} h^d |f(z)|^d \right)^{1/d}. \tag{3.27}$$

Since problem (3.26) is linear this estimate shows that there exists a unique solution to (3.26).

Similar to the continuous case (see Corollary 2.2), Theorem 3.25 implies a comparison principle.

Corollary 3.27 (discrete comparison). Suppose that $\mathcal{L}_h$ is of positive type and of the form (3.16). Let u_h and v_h be two nodal functions with $u_h \leq v_h$ on Ω_h^B and $\mathcal{L}_h u_h \geq \mathcal{L}_h v_h$ in Ω_h^I. Then $u_h \leq v_h$ in $\bar\Omega_h$.

Finally, since problems (3.16) and (3.26) are linear, the Lax–Richtmyer theorem immediately gives us error estimates.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Corollary 3.28 (rate of convergence). Let $I_h^{fd} : C(\bar{\Omega}) \to X_h^{fd}$ denote the canonical interpolant onto nodal functions. Let u be the solution to (3.15), and let $u_h \in X_h^{fd}$ be the unique solution to (3.26). Then

$$\|u_h - I_h^{fd}u\|_{L^\infty(\bar{\Omega}_h)} \leq \|g_h - I_h^{fd}u\|_{L^\infty(\Omega_h^B)} + \frac{CR}{\lambda_{0,h}}\left(\sum_{z\in\Omega_h^I} h^d |\mathcal{L}_h I_h^{fd}u(z))|^d\right)^{1/d}.$$

Based on Corollary 3.28 and the consistency of the approximation scheme, one can derive error estimates with explicit dependence on the discretization parameter h. For example, if we can show that

$$\|g_h - I_h^{fd}u\|_{L^\infty(\Omega_h^B)} = O(h^k) \quad \text{and} \quad \|\mathcal{L}_h I_h^{fd}u\|_{L^\infty(\Omega_h^I)} = O(h^k)$$

for some positive integer $k \in \mathbb{N}$, then Corollary 3.28 shows that the error satisfies $\|u - I_h^{fd}u\|_{L^\infty(\bar{\Omega}_h)} \leq Ch^k$. The value of k is determined by the consistency of the scheme, which typically follows from Taylor's theorem and the regularity of the exact solution. Unfortunately, the monotonicity of a scheme restricts the size of the order of convergence, as shown, for instance, by Oberman (2006) and Kossaczký, Ehrhardt and Günther (2016).

Theorem 3.29 (accuracy of monotone schemes). A monotone finite difference scheme of the form (3.8) is at most second-order accurate for second-order equations.

Finally, solutions to the discrete problem are Hölder-continuous (Kuo and Trudinger 1990, Corollary 4.6, Theorem 5.1).

Theorem 3.30 (Hölder continuity). Let $\mathcal{L}_h$ be of positive type, and let u_h satisfy $\mathcal{L}_h u_h = f$ in Ω_h^I. Assume that Ω satisfies a uniform exterior cone condition. Then there exist $\eta \in (0,1)$ and $C > 0$, independent of h, such that

$$|u_h(z) - u_h(y)| \leq C|z - y|^\eta,$$

for all $z, y \in \bar{\Omega}_h$.

As in the continuous setting, the development of Hölder estimates depends on discrete Harnack inequalities.

3.4. Construction of monotone finite difference schemes

Theorem 3.25 shows that linear, positive finite difference schemes are indeed uniquely solvable with solutions uniformly bounded with respect to the data. We will also see that many of these results carry over to the fully nonlinear case, and thus, applying the Barles–Souganidis framework, such schemes converge to the viscosity solution of the nonlinear PDE. However,

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

the theorem does not indicate how to construct such schemes. We now discuss this issue. First, we have the following classical results (Motzkin and Wasow 1953, Theorems 1 and 2).

Theorem 3.31 (impossibility). For a given (fixed) stencil width $m \in \mathbb{N}$, there exists an linear, elliptic operator $\mathcal{L}$ such that any linear and consistent finite difference scheme of the form (3.16) is not of positive type.

Theorem 3.32 (existence). Let $\mathcal{L}$ be a linear and uniformly elliptic operator satisfying (3.15). Then there exists, for sufficiently small h, a consistent finite difference scheme of the form (3.16) that is of positive type.

The main punchline of these theorems is that wide stencils are a necessary feature of consistent and positive-type finite difference discretizations, even for linear problems. The proof of Theorem 3.32, as presented in Motzkin and Wasow (1953, Theorem 2), is not constructive. On the other hand, the arguments given in Kuo and Trudinger (1990) and Kocan (1995) explicitly give an algorithm to construct consistent and positive schemes, and as a result, provide an estimate of the stencil width. We end this section by summarizing these results. As a first step, we state the following trivial observation.

Lemma 3.33 (positivity criterion). Suppose that $\mathcal{L}u = A : D^2u$, where the coefficient matrix is positive definite and has the form

$$A(x) = \sum_{\substack{y\in\mathbb{Z}^d \\ |y|_\infty \leq M}} a_y(x) y \otimes y, \tag{3.28}$$

for some $M \in \mathbb{N}$ and with coefficients $a_y(x) \geq 0$ for all $x \in \Omega$. Assume further that there exists an orthogonal set $\{y_i\}_{i=1}^d \subset \mathbb{Z}^d$ with $|y_i|_\infty \leq M$ such that $a_{y_i}(x) \geq c$ for some $c > 0$. Then the finite difference operator

$$\mathcal{L}_h u_h(z) = \sum_{\substack{y\in\mathbb{Z}^d \\ |y|_\infty \leq M}} a_y(z) \delta_{y,h}^2 u_h(z) \tag{3.29}$$

is of positive type and a consistent approximation to $\mathcal{L}$ with $\lambda_{0,h} = c$.

Of course, not all positive definite matrices are of the form (3.28). However, quite surprisingly, Lemma 3.33 provides the essential tools to construct consistent and positive finite difference schemes for linear (and nonlinear) elliptic equations. Let us consider an example.

Lemma 3.34 (diagonally dominant matrix). Suppose that $\mathcal{L}u = A : D^2u$ with

$$A(x) = \sum_{i,j=1}^{d} a_{i,j}(x) y_i \otimes y_j, \tag{3.30}$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

where $\{y_i\}_{i=1}^d \subset \mathbb{Z}^d$ is an orthogonal basis of $\mathbb{R}^d$, that is,

$$a_{i,j} = y_i \cdot Ay_j/(|y_i|\,|y_j|).$$

Suppose, in addition, that for some $c > 0$ we have

$$\sum_{\substack{i,j=1\\ j\neq i}}^{d} |a_{i,j}(x)| \leq a_{i,i}(x) - c \quad i = 1, 2, \ldots, \quad x \in \Omega, \tag{3.31}$$

Then there exists a positive finite difference method $\mathcal{L}_h$ that is consistent with $\mathcal{L}$ and with $\lambda_{0,h} = c$.

Proof. By manipulating terms, we may write

$$\begin{aligned} A = &\sum_{i=1}^{d}\left(a_{ii} - \sum_{\substack{j=1\\ j\neq i}}^{d} |a_{i,j}|\right) y_i \otimes y_i \\ &+ \frac{1}{4}\sum_{\substack{i,j=1\\ i\neq j}}^{d} (|a_{i,j}| + a_{i,j})(y_i + y_j) \otimes (y_i + y_j) \\ &+ \frac{1}{4}\sum_{\substack{i,j=1\\ i\neq j}}^{d} (|a_{i,j}| - a_{i,j})(y_i - y_j) \otimes (y_i - y_j). \end{aligned}$$

Thus, A satisfies the conditions in Lemma 3.33 with $M \leq 2\max_i |y_i|_\infty$. The result now follows from Lemma 3.33. □

Remark 3.35 (stencil size). By taking $\{y_i\}_{i=1}^d \subset \mathbb{Z}^d$ as the standard basis of $\mathbb{R}^d$ in Lemma 3.34, we deduce that if A is strictly diagonally dominant, then there exists a consistent and positive one-step finite difference method.

Now, following Kocan (1995, Theorem 5), we can estimate the stencil size for a general positive definite matrix.

Theorem 3.36 (existence). Consider the elliptic operator $\mathcal{L}u = A : D^2u$, where A is uniformly positive definite in Ω with

$$\lambda I \leq A \leq \Lambda I$$

for positive constants $\lambda \leq \Lambda$. Then there exists a consistent and positive operator $\mathcal{L}_h$. The stencil size m satisfies $m \leq 2M_d(8d^3\mathcal{E})$, where $\mathcal{E} := \Lambda/\lambda$ and

$$M_d(s) = C\begin{cases} s & d = 2, \\ s^{5/2} & d = 3, \\ s^{2d-4} & d \geq 4. \end{cases}$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Here the constant $C > 0$ depends only on the dimension d. Moreover, one can take the discrete ellipticity constant to be $\lambda_{0,h} = \lambda/(2dm^2)$.

Proof. Let $\{\lambda_i\}_{i=1}^d \subset [\lambda, \Lambda]$ denote the eigenvalues of A and let $\{\varphi_i\}_{i=1}^d$ be an orthonormal set of eigenvectors of A, labelled such that

$$A = \sum_{i=1}^{d} \lambda_i \varphi_i \otimes \varphi_i.$$

By Kocan (1995, Theorem 2), for $s > 0$ to be determined, there exists $y_i \in \mathbb{Z}^d$ such that

$$\left| \varphi_i - \frac{y_i}{|y_i|} \right|_\infty \leq \frac{1}{s}, \quad \frac{M_d(s)}{2} \leq |y_i|_\infty \leq M_d(s). \tag{3.32}$$

Since $\{y_i \otimes y_j\}_{i,j=1}^d$ spans $\mathbb{S}^d$, we may write

$$\begin{aligned} A &= \sum_{i=1}^{d} \frac{\lambda_i}{|y_i|^2} y_i \otimes y_i + \sum_{i=1}^{d} \lambda_i \left(\varphi_i \otimes \varphi_i - \frac{y_i \otimes y_i}{|y_i|^2} \right) \\ &= \sum_{i=1}^{d} \frac{\lambda_i}{|y_i|^2} y_i \otimes y_i + \sum_{i,j=1}^{d} \frac{B_{i,j}}{|y_i|\,|y_j|} y_i \otimes y_j \end{aligned}$$

for some $B \in \mathbb{R}^{d \times d}$. Thus, A is of the form (3.30) with $a_{i,i} = (\lambda_i + B_{i,i})/|y_i|^2$ and $a_{i,j} = B_{i,j}/(|y_i|\,|y_j|)$.

Applying (3.32) and the inequalities $\lambda \leq \lambda_i \leq \Lambda$, we obtain

$$|B_{i,j}| \leq \frac{2\Lambda d^{3/2}}{s},$$

and hence

$$|a_{i,i}| \geq \frac{1}{|y_i|^2} \left(\lambda - \frac{2\Lambda d^{3/2}}{s} \right) \geq \frac{\lambda}{|y_i|^2}, \quad |a_{i,j}| \leq \frac{2d^{3/2}\Lambda}{s|y_i|\,|y_j|}.$$

Thus (3.31) will be satisfied if

$$\lambda \geq \max_i \left(\frac{2d^{3/2}\Lambda}{s} \sum_{j=1}^{d} \frac{|y_i|}{|y_j|} + c|y_i|^2 \right)$$

for some constant $c > 0$. Taking $c = \lambda/(2dm^2)$, and noting that $|y_i|/|y_j| \leq 2d^{1/2}$, we conclude that if

$$s \geq 8d^3\mathcal{E},$$

then (3.31) is satisfied. The desired result now follows from Lemma 3.34. □

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Finally we end this section with a result which shows that the uniform ellipticity of $\mathcal{L}$ condition in Theorem 3.36 cannot be relaxed; see Kocan (1995, Theorem 1).

Example 3.37 (uniform ellipticity is necessary). Suppose that $d = 2$ and $Lu = A : D^2u$, where

$$A = \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix},$$

$\alpha, \gamma > 0$ and $\alpha\gamma = \beta^2$ (so that $\det(A) = 0$). Then if $\sqrt{\alpha/\gamma}$ is irrational, there does not exist a consistent and non-negative scheme for the problem $\mathcal{L}u = f$.

3.5. Monotonicity in finite element methods

In this section we discuss the construction of monotone finite element methods for second-order elliptic problems. As in the previous section, we focus on the linear case, where the elliptic problem is given by (3.15), and extend these results to nonlinear problems in subsequent sections. We further simplify the presentation and analysis by assuming that the coefficient matrix is the identity matrix, $A = I$, and assume Dirichlet boundary conditions. Thus, we focus on the Poisson problem

$$\Delta u = f \ \text{ in } \Omega, \quad u = g \ \text{ on } \partial\Omega. \tag{3.33}$$

Quite surprisingly, the results given here extend to fully nonlinear problems.

Let $\mathcal{T}_h$ be a simplicial, conforming and quasi-uniform triangulation of Ω (Ciarlet 2002). For simplicity, and to communicate the essential ideas, we shall ignore the approximation of Ω by the triangulation-induced polytope, and simply assume throughout the paper that $\bar{\Omega} = \cup_{T\in\mathcal{T}_h}\bar{T}$. Let X_h^l be the linear Lagrange finite element space, that is,

$$X_h^l = \{v_h \in C(\bar{\Omega}) : v_h|_T \in \mathbb{P}_1 \ \forall T \in \mathcal{T}_h\}. \tag{3.34}$$

We extend the definitions given in the previous sections to unstructured meshes by letting Ω_h^I and Ω_h^B denote the sets of vertices (or nodes) of the triangulation $\mathcal{T}_h$ that belong to Ω and $\partial\Omega$ respectively. Then any function $w_h \in X_h^l$ is uniquely determined by the values $w_h(z)$ for all $z \in \bar{\Omega}_h := \Omega_h^I \cup \Omega_h^B$.

To describe the finite element method and to facilitate further developments, we assume that the vertices are labelled such that $\bar{\Omega}_h = \{z_i\}_{i=1}^{N+M}$ for positive integers N, M, with $\Omega_h^I = \{z_i\}_{i=1}^{N}$ and $\Omega_h^B = \{z_i\}_{i=N+1}^{M}$. The fact that continuous piecewise linear polynomials are uniquely determined by their values at the vertices induces a basis of *hat functions* $\{\tilde{\phi}_i\} \subset X_h^l$, with the unique property $\tilde{\phi}_i(z_j) = \delta_{i,j}$. We define the *normalized hat functions*

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

as $\phi_i = c_i^{-1}\tilde{\phi}_i$ with $c_i = \left(\int_\Omega \tilde{\phi}_i\right) > 0$, and note that

$$v_h = \sum_{i=1}^{N+M} c_i v_h(z_i)\phi_i, \quad \text{for all } v_h \in X_h^l,$$

and

$$v_h = \sum_{i=1}^{N} c_i v_h(z_i)\phi_i, \quad \text{for all } v_h \in X_{0,h}^l := X_h^l \cap H_0^1(\Omega). \tag{3.35}$$

We set $\omega_{z_i} = \text{supp}(\phi_i)$, which is the union of elements in $\mathcal{T}_h$ that have z_i as a vertex.

A finite element method for the Poisson problem simply restricts the variational formulation (2.9) (with $A = I$) onto the piecewise polynomial space X_h^l. Thus, we consider the following problem: find $u_h \in X_h^l$ with $u_h(z_i) = g(z_i)$ $(N+1 \le i \le N+M)$ and

$$-\int_\Omega Du_h \cdot Dv_h = \int_\Omega f v_h, \quad \text{for all } v_h \in X_{0,h}^l. \tag{3.36}$$

As in the continuous setting, the existence and uniqueness of u_h readily follows from the Lax–Milgram lemma. An application of Céa's lemma and interpolation results also show that the error satisfies $\|\nabla(u - u_h)\|_{L^2(\Omega)} = O(h)$ provided $u \in H^2(\Omega)$; we refer the reader to Ciarlet (2002), Brenner and Scott (2008) and Ern and Guermond (2004), for example, for proofs of these basic results.

To pose this problem in the operator framework of the previous sections, we first note that (3.36) is equivalent to the conditions

$$-\int_\Omega Du_h \cdot D\phi_i = \int_\Omega f\phi_i \quad i = 1, 2, \ldots, N. \tag{3.37}$$

Define $\mathcal{L}_h$ such that for all interior vertices $z_i \in \Omega_h^I$,

$$\mathcal{L}_h u_h(z_i) = \Delta_h u_h(z_i), \tag{3.38}$$

where the *finite element Laplacian* is defined by

$$\Delta_h u_h(z_i) := -\int_\Omega Du_h \cdot D\phi_i. \tag{3.39}$$

Set

$$f_h(z_i) = \int_\Omega f\phi_i.$$

We further define the piecewise linear function g_h on $\partial\Omega$ with the property

$$g_h(z_i) = g(z_i) \quad i = N+1, \ldots, N+M, \tag{3.40}$$

that is, $g_h = I_h^{fe} g$ is the nodal interpolant of g. With this notation, we see

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

that the finite element method (3.36) is equivalent to problem (3.6) with $F_h = \mathcal{L}_h - f_h$.

Before discussing the monotonicity of the scheme (3.36), let us first point out that, unlike the finite difference scheme, one cannot take $I_h = I_h^{fe}$, the nodal interpolant, in Definition 3.1 to deduce the (operator) consistency of the finite element approximation. The next lemma exemplifies this point. For further details the reader is referred to Jensen and Smears (2013*b*) and Nochetto and Zhang (2017*a*).

Lemma 3.38 (finite element inconsistency). Let Δ_h be the finite element Laplacian, defined in (3.39), and denote by $I_h^{fe} : C(\bar{\Omega}) \to X_h^l$, the nodal interpolant onto the linear Lagrange finite element space. Then, in general, we have

$$\Delta_h(I_h^{fe}u)(z) \not\to \Delta u(z), \quad \text{for all } z \in \bar{\Omega}_h \ \text{ as } h \to 0$$

for all $u \in C^2(\Omega)$.

Proof. Let $d = 2$, and consider the triangulation $\mathcal{T}_h$ with four triangles and vertices $z_1 = (0,0)$, $z_2 = (h,0)$, $z_3 = (0,h)$, $z_4 = (-h,0)$ and $z_5 = (0,-h)$. Let u be a C^2-function that vanishes at the origin. Then a calculation shows that

$$\Delta_h I_h^{fe}u(z_1) = \frac{3}{2}h^{-2}\sum_{i=2}^{5} u(z_i).$$

Taking, for example, $u(x_1,x_2) = x_1^2$ then yields

$$\Delta_h I_h^{fe}u(z_1) = 3 \neq 2 = \Delta u(z_1), \quad \text{for all } h > 0.$$

In other words, Δ_h and I_h^{fe} are not a consistent approximation scheme. □

The inconsistency in Lemma 3.38 is caused by the wrong choice of interpolation operator. The so-called elliptic projection gives a correct one.

Definition 3.39 (elliptic projection). The elliptic projection

$$I_h^{ep} : H^1(\Omega) \cap C(\bar{\Omega}) \to X_h^l$$

is defined by

$$\Delta_h I_h^{ep}u(z_i) = -\int_\Omega Du \cdot D\phi_i \quad i = 1,2,\ldots,N, \tag{3.41}$$

and $I_h^{ep}u = u$ on Ω_h^B.

The elliptic projection is (almost) quasi-optimal in the L^∞-norm (Schatz and Wahlbin 1982).

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Proposition 3.40 (properties of I_h^{ep}). Let $I_h^{ep}u \in X_{0,h}^l$ be the elliptic projection of $u \in C(\bar{\Omega}) \cap H_0^1(\Omega)$ defined by (3.41). Then

$$\|u - I_h^{ep}u\|_{L^\infty(\Omega)} \leq C\,|\log h| \inf_{v_h \in X_{0,h}^l} \|u - v_h\|_{L^\infty(\Omega)}.$$

If $u \in W^{2,\infty}(\Omega)$, then

$$\|u - I_h^{ep}u\|_{L^\infty(\Omega)} \leq C\,|\log h| h^2 \|u\|_{W^{2,\infty}(\Omega)}.$$

More importantly, the finite element method is consistent when one uses the elliptic projection I_h^{fe}.

Lemma 3.41 (finite element consistency). Let $\{z_h\}_{h>0}$ with $z_h \in \bar{\Omega}_h^I$ and $z_h \to z_0 \in \Omega$. Then, for all $u \in C^2(\bar{\Omega})$,

$$\mathcal{L}_h I_h^{ep}u(z_h) = \Delta_h(I_h^{ep}u)(z_h) \to \Delta u(z_0), \quad \text{for all } z \in \bar{\Omega}_h,$$

as $h \to 0+$. Moreover, $I_h^{ep}u \to u$ on $\partial\Omega$.

Proof. The convergence $I_h^{ep}u \to u$ on $\partial\Omega$ follows from the definition of I_h^{ep} and standard interpolation theory.

Owing to the regularity of u, we have $\Delta u(z_h) \to \Delta u(z_0)$ as $h \to 0^+$. Now, let $\{\phi_h\}_{h>0} \subset X_{0,h}^l$ denote the normalized hat functions. Then, since $\|\phi_h\|_{L^1(\Omega)} = 1$, and $\phi_h \geq 0$, we have

$$\int_\Omega (\Delta u)\phi_h = \int_{\omega_{z_h}} (\Delta u)\phi_h \to \Delta u(z_0).$$

Therefore, by integration by parts,

$$\Delta_h I_h^{ep}u(z_h) = \int_\Omega (\Delta u)\phi_h \to \Delta u(z_0)$$

as $h \to 0+$, and consequently Δ_h is consistent. □

Lemma 3.42 (finite element monotonicity). Suppose that the bases satisfy

$$\int_\Omega D\phi_i \cdot D\phi_j \leq 0. \tag{3.42}$$

for $i, j = 1, 2, \ldots, N+M$ and $i \neq j$. Then $\mathcal{L}_h$, given by (3.38), is monotone.

Proof. Suppose that $v_h, w_h \in X_h^l$ and $w_h - v_h$ has a non-negative maximum at an interior vertex $z_i \in \Omega_h^B$. Without loss of generality we may assume

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

that $w_h \le v_h$ and $w_h(z_i) = v_h(z_i)$. Then we find that

$$\begin{aligned}\mathcal{L}_h w_h(z_i) - \mathcal{L}_h v_h(z_i) &= -\int_\Omega D(w_h - v_h)\cdot D\phi_i \\ &= -\sum_{j=1}^{N+M} c_j(w_h(z_j) - v_h(z_j))\int_\Omega D\phi_j \cdot D\phi_i \le 0.\end{aligned}$$

Thus, $\mathcal{L}_h w_h(z_i) \le \mathcal{L}_h v_h(z_i)$, and therefore $\mathcal{L}_h$ is monotone. □

Lemma 3.42 indicates that the finite element method is monotone provided that a certain mesh condition is satisfied. Indeed, for an edge $E \subset \partial T$, we let θ_E^T denote the angle between the faces not containing E, and let κ_E^T be the $(d-2)$-dimensional simplex opposite to E. Then we have

$$\int_\Omega D\tilde{\phi}_i \cdot D\tilde{\phi}_j = -\frac{1}{d(d-1)}\sum_{T\supset E}|\kappa_E^T|\cot\theta_E^T$$

(Xu and Zikatanov 1999, Strang and Fix 1973), where E is the edge with vertices z_i and z_j. The condition (3.42) is satisfied if the mesh is weakly acute. For example, in two dimensions, this condition means that the sum of the angles opposite to any edge is less than or equal to π.

Corollary 3.43 (maximum principle). Let $u_h \in X_h^l$ solve (3.37). Suppose that (3.42) is satisfied and $f \ge 0$ and $g \le 0$. Then $u_h \le 0$.

Proof. Define the stiffness matrix $S \in \mathbb{R}^{(N+M)\times(N+M)}$ by

$$S_{i,j} = -\int_\Omega D\phi_i \cdot D\phi_j, \tag{3.43}$$

and note that condition (3.42) is equivalent to $S_{i,j} \ge 0$ for $i \ne j$. Moreover, since the hat functions form a partition of unity,

$$\sum_{j=1}^{N+M} c_j S_{i,j} = 0.$$

Now, suppose that u_h attains a strict positive maximum at an interior node z_i. We then find

$$\begin{aligned}0 = \mathcal{L}_h u_h(z_i) &= -\int_\Omega Du_h\cdot D\phi_i - \int_\Omega f\phi_i \\ &\le -\int_\Omega Du_h\cdot D\phi_i \\ &= \sum_{j=1}^{N+M} c_j u_h(z_j)S_{i,j} = \sum_{j=1}^{N+M} c_j(u_h(z_j)-u_h(z_i))S_{i,j} < 0,\end{aligned}$$

a contradiction. □

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

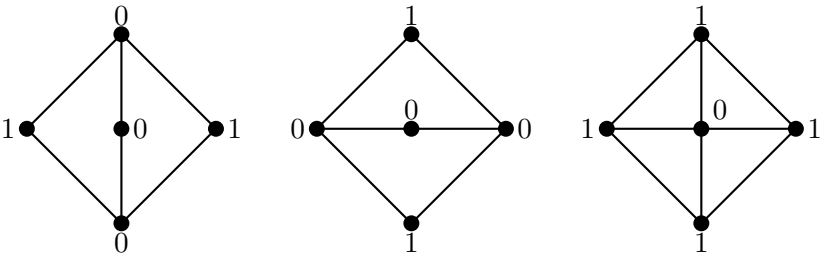


Figure 3.1. Meshes corresponding to convex envelopes $\Gamma(v_1) = |x_1|$ (a) and $\Gamma(v_2) = |x_2|$ (b), and $\Gamma(v_3) = |x_1| + |x_2|$ (c).

3.6. Finite element stability estimates: Alexandrov estimates and Alexandrov–Bakelman–Pucci maximum principle

As in the finite difference schemes discussed in the previous section, we develop some discrete Alexandrov estimates for finite element functions and analogous ABP maximum principles. Before stating and proving these results, it is useful to discuss some properties of the convex envelope of piecewise linear polynomials.

For $v_h \in X_h^l$ with $v_h \geq 0$ on $\partial\Omega$, let $\Gamma(v_h)$ and $\Gamma_h(v_h)$ denote the convex envelope and discrete convex envelope of v_h given in Definitions 2.59 and 3.17, respectively. Then, since v_h is piecewise affine, we find that $\Gamma_h(v_h) = \Gamma(v_h)$, and furthermore, $\Gamma_h(v_h)$ is also piecewise affine. However, perhaps unexpectedly, $\Gamma(v_h)$ is not necessarily piecewise linear subordinate to $\mathcal{T}_h$! The following examples illustrate this feature.

Example 3.44 (convex envelope). Consider a triangulation with vertices $z_1 = (1,0)$, $z_2 = (0,1)$, $z_3 = (-1,0)$, $z_4 = (0,-1)$ and $z_5 = (0,0)$. Consider the piecewise linear functions satisfying

$$v_1(z_1) = v_1(z_3) = 1, \quad v_2(z_2) = v_2(z_4) = 1,$$
$$v_3(z_1) = v_3(z_2) = v_3(z_3) = v_3(v_4) = 1,$$

and $v_j(z_i) = 0$ otherwise. The convex envelopes are $\Gamma(v_1) = |x_1|$, $\Gamma(v_2) = |x_2|$, and $\Gamma(v_3) = |x_1| + |x_2|$. The convex envelopes are subordinate to the meshes depicted in Figure 3.1.

As shown in the example above, since $\Gamma(v_h)$ is a piecewise linear function, it induces a mesh $\widetilde{\mathcal{T}}_h$ which depends on v_h. The following example shows that if v_h is the nodal interpolant of a function v, and if the Hessian $D^2 v$ is degenerate (or nearly degenerate), the induced mesh may be anisotropic.

Example 3.45 (anisotropy). Let $\Omega = \mathbb{R}^2$ and $\bar{\Omega}_h = \{(k,m)\}$. Let $v(x) = (x \cdot e)^2$, where $e = (1, A)$ for some integer $0 < A$ and $v_h(z) = v(z)$ for

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

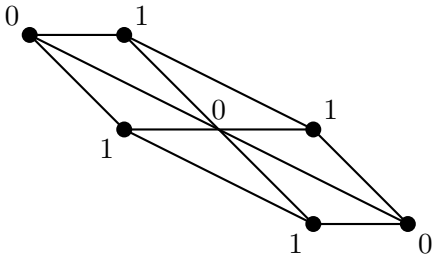


Figure 3.2. Mesh induced by the nodal interpolant of $v(x) = (x \cdot e)^2$ where $e = (1, 2)$. Its convex envelope equals $|x \cdot e|$ in the star of $(0, 0)$.

all $z \in \bar{\Omega}_h$. Then the convex envelope induces an anisotropic mesh depicted in Figure 3.2. The convex envelope in the star of the origin is $|x \cdot e|$.

Let us now state the Alexandrov estimate for finite element functions. Recall that Ω is compactly contained in a ball B_R, and the nodal contact set $\mathcal{C}_h^-(v_h)$ is given in Definition 3.23.

Lemma 3.46 (finite element Alexandrov estimate). For every $v_h \in X_h^l$ such that $v_h \geq 0$ on $\partial\Omega$, we have

$$\sup_{\bar{\Omega}} v_h^- \leq CR\Big(\sum_{z \in \mathcal{C}_h^-(v_h)} |\partial\Gamma(v_h)(z)| \Big)^{1/d}, \tag{3.44}$$

where the constant C depends only on the dimension d and the domain Ω.

Proof. The result directly follows from the proof of Lemma 3.24. Indeed it suffices to realize that, for every $v_h \in X_h^l$, $\sup_{\bar{\Omega}} v_h^- = \sup_{\bar{\Omega}_h} v_h^-$ and $\Gamma_h(v_h) = \Gamma(v_h)$; see Nochetto and Zhang (2017*a*, Proposition 5.1) for details. □

Let us point out that, with Lemma 3.46 in hand, it may be possible to extend the arguments given in Theorem 3.25 to develop ABP estimates for piecewise linear polynomials. Instead, following Nochetto and Zhang (2017*a*, Section 5), we outline a proof which is more geometric and is based on the characterization of the subdifferential of piecewise linear functions.

As a first step we define the local convex envelope and local subdifferential at a point $z \in \mathcal{C}_h^-(v_h)$ (see Definition 3.23).

Definition 3.47 (local convex envelope). For $v_h \in X_h^l$ and contact node $z \in \mathcal{C}_h^-(v_h)$, let ω_z denote the union of elements in $\mathcal{T}_h$ that have z as a vertex. We then define the *local convex envelope* by

$$\Gamma_z(v_h)(x) = \sup\{L(x) : L \leq v_h \ \text{ in } \omega_z,\ L \in \mathbb{P}_1,\ L(z) = v_h(z)\}$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

for all $x \in \omega_z$. Its *local subdifferential* is

$$\partial\Gamma_z(v_h)(z) = \{\mathbf{p} \in \mathbb{R}^d : \Gamma_z(v_h)(x) \geq \Gamma_z(v_h)(z) + \mathbf{p} \cdot (x - z),\ \forall x \in \omega_z\}. \tag{3.45}$$

Comparing (3.45) with Definition 2.59, we easily deduce that

$$\partial\Gamma(v_h)(z) \subset \partial\Gamma_z(v_h)(z), \quad \text{for all } z \in \mathcal{C}_h^-(v_h), \tag{3.46}$$

and therefore, by Lemma 3.46,

$$\sup_{\bar{\Omega}} v_h^- \leq CR\Big(\sum_{z \in \mathcal{C}_h^-(v_h)} |\partial\Gamma_z(v_h)(z)|\Big)^{1/d}. \tag{3.47}$$

Less obvious is the following result.

Proposition 3.48 (subordination). Suppose that $d = 2$ and, for $v_h \in X_h^l$ and contact point $z \in \mathcal{C}_h^-(v_h)$, let $\Gamma_z(v_h)$ be given by Definition 3.47. Then $\Gamma_z(v_h)$ is subordinate to $\mathcal{T}_h$.

We refer the reader to Nochetto and Zhang (2017*a*, Lemma 5.1) for a proof of this result. Here, instead, we will show that if $d \geq 3$ the assertion is no longer true. Set

$$z_0 = (0, 0, -1), \quad z_1 = (-1, 0, 0), \quad z_2 = (0, 1, 0), \quad z_3 = (1, 0, 0),$$

and let T_1, T_2 be the convex hulls of z_0, z_1, z_2, z_3 and $z_0, z_1, -z_2, z_3$. Consider the piecewise linear function v_h with values $v_h(z_0) = -1$, $v_h(z_1) = v_h(z_3) = 0$ and $v_h(\pm z_2) = -1$. Then $\Gamma_{z_0}(v_h)(x) = |x_1| - 1$ is not affine on T_i for each $i = 1, 2$.

Next, to derive an ABP maximum principle, we state the relation between the subdifferential of a convex, piecewise linear polynomial with its finite element Laplacian. As a first step, we first integrate by parts in (3.38) to get the identity

$$\Delta_h v_h(z_i) = -\sum_{F \in \mathcal{F}_{z_i}} \int_F [[Dv_h]]\phi_i, \quad \text{for all } v_h \in X_h^l.$$

Here, $\mathcal{F}_{z_i}$ is the set of (interior) $(d-1)$-dimensional simplices that have z_i as a vertex, and, for a vector-valued function $\mathbf{w}$, the *jump of* $\mathbf{w}$ across the face F is given by

$$[[\mathbf{w}]]\big|_F := \begin{cases} \mathbf{n}_F^+ \cdot \mathbf{w}^+\big|_F + \mathbf{n}_F^- \cdot \mathbf{w}^-\big|_F & \text{if } F = \partial K^+ \cap \partial K^-, \\ \mathbf{n}_F^+ \cdot \mathbf{w}^+\big|_F & \text{if } F = \partial K^+ \cap \partial\Omega, \end{cases} \tag{3.48}$$

with $\mathbf{w}^\pm = \mathbf{w}|_{K^\pm}$, and $\mathbf{n}_F^\pm$ denoting the outward unit normal vectors of $K^\pm$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

on F. We also define the jump of a scalar function v across F as

$$[[v]]\big|_F := \begin{cases} \mathbf{n}_F^+ v^+\big|_F + \mathbf{n}_F^- v^-\big|_F & \text{if } F = \partial K^+ \cap \partial K^-, \\ \mathbf{n}_F^+ v^+\big|_F & \text{if } F = \partial K^+ \cap \partial\Omega. \end{cases} \tag{3.49}$$

Now, since $[[Dv_h]]$ is constant on F, and

$$\int_F \phi_i = \frac{d+1}{d}\frac{|F|}{|\omega_{z_i}|},$$

we can obtain an expression on the finite element Laplacian with explicit dependence on the jumps:

$$\Delta_h v_h(z_i) = \frac{-(d+1)}{d} \sum_{F\in\mathcal{F}_{z_i}} \frac{|F|}{|\omega_{z_i}|}[[Dv_h]]\big|_F. \tag{3.50}$$

A relationship between the jumps of the gradients (and hence the discrete Laplacian) and the subdifferential of a convex, piecewise affine function is now given.

Proposition 3.49 (subdifferential versus jumps). Let γ be a piecewise affine convex function on a patch ω_z for some node $z\in\Omega_h^I$, and let $\mathcal{F}_z$ denote the set of $(d-1)$-dimensional simplices that touch z. Then, for any $F\in\mathcal{F}_z$, the jump $[[D\gamma]]|_F$ is non-positive and

$$|\partial\gamma(z)| \le C\bigg(\sum_{F\in\mathcal{F}_z} -[[D\gamma]]\big|_F\bigg)^d. \tag{3.51}$$

We will not give a complete proof of Proposition 3.49, but rather give a rough idea of how such a result is obtained in two dimensions. Further details can be found in Nochetto and Zhang (2017*a*, Section 5.2).

Without loss of generality, assume that $z=0$ and $\gamma(0)=0$. We further denote by $\{z_j\}_{j=1}^m$ the set of nodes in ω_z. Now, since γ is a piecewise affine function, a vector $\mathbf{p}\in\partial\gamma(0)$ is characterized by the inequalities

$$\mathbf{p}\cdot z_j \le \gamma(z_j) \quad 1\le j\le m.$$

Therefore, we conclude that the subdifferential of $\gamma(0)$ is a convex polygon determined by the intersection of the half-spaces

$$S_j := \{\mathbf{p}\in\mathbb{R}^2 : \mathbf{p}\cdot z_j \le \gamma(z_j)\}, \tag{3.52}$$

and that a vector $\mathbf{p}$ is in the interior of $\partial\gamma(0)$ if and only if

$$\mathbf{p}\cdot z_j < \gamma(z_j), \quad 1\le j\le m,$$

and it is on the boundary of $\partial\gamma(0)$ if

$$\mathbf{p}\cdot z_j = \gamma(z_j)$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

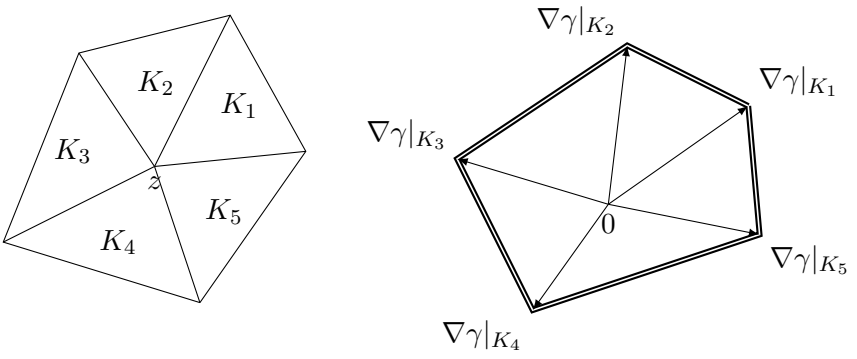


Figure 3.3. A pictorial description of the dual set of a patch.

for some j. This characterization of the boundary motivates the introduction of a $\partial\gamma(0)$-induced *dual mesh*, which we now explain.

Let T be an n-dimensional simplex in ω_z with $0 \leq n \leq 2$ such that $0 \in T$. We then define the $(2-n)$-dimensional dual set T^* as follows (see Figure 3.3).

- If $n = 0$, so that $T = \{0\}$, then we define T^* as the subdifferential $\partial\gamma(0)$.
- If $n = 2$, so that $T = K$ is an element of ω_z, then T^* is the vector $\partial\gamma\big|_K$.
- If $n = 1$, so that $T = F = \partial K^+ \cap \partial K^-$ is an (interior) edge in ω_z, then T^* is the line segment jointing the two vectors $\partial\gamma|_{K^\pm} = D\gamma_{K^\pm}$.

Note that T^* is a convex polytope contained in the $(2-n)$-dimensional plane

$$P_T = \{\mathbf{p} \in \mathbb{R}^2 : \mathbf{p} \cdot z = \gamma(z) \ \forall z \in T\},$$

and therefore, for arbitrary $\mathbf{p}_1, \mathbf{p}_2 \in P_T$, $(\mathbf{p}_1 - \mathbf{p}_2) \cdot z = 0$ for all $z \in T$, that is, P_T is orthogonal to T.

Now, the proceeding discussion implies that the boundary of $\partial\gamma(0)$ is given by

$$\bigcup \big\{F^* : \text{edges } F \text{ in } \omega_z \text{ with } 0 \in F\big\},$$

in other words, the boundary of $\partial\gamma$ is made of line segments that join $\partial\gamma(0)$ on neighbouring triangles; for further details see Nochetto and Zhang (2017*a*, Proposition 5.6). If $F = \partial K^+ \cap \partial K^-$, then it follows that the length of F^* is given by

$$|F^*| = \big|D\gamma|_{K^+} - D\gamma|_{K^-}\big| = -[[D\gamma]]|_F,$$

where we have used the fact that $D\gamma|_{K^+} - D\gamma_{K^-}$ is perpendicular to F

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

and the non-positivity of $[[D\gamma]]$. Putting everything together, we conclude that the boundary of $\partial\gamma(0)$ is bounded by $\sum_{F\in\mathcal{F}_z} -[[D\gamma]]\big|_F$. Thus, by the isoperimetric inequality,

$$|\partial\gamma(0)| \le C\bigg(\sum_{F\in\mathcal{F}_z} -[[D\gamma]]\big|_F\bigg)^2.$$

This last statement is (3.51) for $d = 2$.

Theorem 3.50 (finite element ABP estimate). Suppose that the simplicial mesh $\mathcal{T}_h$ satisfies (3.42) and that $u_h \in X_h^l$ satisfies

$$\begin{cases} \mathcal{L}_h u_h \le f_h & \text{in } \Omega_h^I, \\ u_h = g_h & \text{on } \Omega_h^B, \end{cases}$$

where $\mathcal{L}_h$ and g_h are given by (3.38) and (3.40), respectively, and X_h^l is the linear Lagrange space defined by (3.34). Then

$$\sup_{\bar\Omega} u_h^- \le \sup_{\Omega_h^B} g_h^- + CR\bigg(\sum_{z\in\mathcal{C}_h^-(u_h)} |\omega_z|(f_h^+(z))^d\bigg)^{1/d}. \tag{3.53}$$

Proof. We give the proof under the assumption that $\Gamma_z(u_h)$ is subordinate to $\mathcal{T}_h$ (which is the case in two dimensions). For the proof of the general case, we refer interested readers to Nochetto and Zhang (2017*a*).

As in the proof of Theorem 3.25, we may assume $u_h \ge 0$ on Ω_h^B and $\sup_{\bar\Omega} u_h^- > 0$.

Let $z \in \mathcal{C}_h^-(u_h)$, and note that $\Gamma_z(u_h)(x) \le u_h(x)$ for all $x \in \omega_z$ with equality at z. Then by (3.42) and Lemma 3.42, and since $\Gamma_z(u_h)$ is subordinate to $\mathcal{T}_h$, we find that

$$\Delta_h\Gamma_z(u_h)(z) \le \Delta_h u_h(z).$$

Moreover, by (3.50), we have

$$-\sum_{F\in\mathcal{F}_z} \frac{|F|}{|\omega_z|}[[D\Gamma_z(u_h)]]\big|_F = \frac{d}{d+1}\Delta_h\Gamma_z(u_h)(z) \le \frac{d}{d+1}\Delta_h u_h(z).$$

Applying Proposition 3.49, and using $|F| \approx |\omega_z|^{1-1/d}$, we conclude that

$$|\partial\Gamma_z(u_h)(z)| \le C\bigg(-\sum_{F\in\mathcal{F}_z} \frac{|F|}{|\omega_z|}[[D\Gamma_z]]\big|_F\bigg)^d |\omega_z| \le C\frac{|\omega_z|}{d+1}\Delta_h u_h(z).$$

Combining this last inequality with (3.47) yields the result. □

Remark 3.51 (contact set). The proof of Theorem 3.50 shows that if $\mathcal{L}_h u_h \le 0$ only on $\mathcal{C}_h^-(u_h)$, then (3.53) is still satisfied. This will be important in subsequent developments.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

4. Finite element methods for elliptic problems in non-divergence form

In this section we summarize recent advances in finite element methods for elliptic problems in non-divergence form with non-smooth coefficients. For simplicity, and to illustrate the main ideas, we consider problems of the form (2.5) with no lower-order terms and with homogeneous Dirichlet boundary conditions. In this setting the problem reads

$$\mathcal{L}u = A : D^2 u = f \ \text{ in } \Omega, \quad u = 0 \ \text{ on } \partial\Omega, \tag{4.1}$$

where A is a symmetric positive definite in $\bar{\Omega} \subset \mathbb{R}^d$ with either $A \in C(\bar{\Omega}, \mathbb{S}^d)$ or $A \in L^\infty(\Omega, \mathbb{S}^d)$. Further assumptions of the domain Ω, its boundary $\partial\Omega$, the coefficient matrix A, and the source function f will be made as they become necessary.

Recall from Section 2.1 that if the coefficient matrix is sufficiently smooth, then we can write the PDE in divergence form with A as the diffusion coefficient and $D\cdot A$ (taken column-wise) as the convective coefficient. In this setting, (weak) solutions are defined by an integration-by-parts argument (see Section 2.3), and as such, finite element methods are easily constructed. However, when A is not differentiable, the clear-cut methodology of Galerkin methods is no longer valid.

This section summarizes three classes of finite element methods for problem (4.1), each motivated by the different solution concepts presented in Sections 2.4–2.5. The first class considers problem (4.1) on convex domains with $A \in L^\infty(\Omega, \mathbb{S}^d)$ satisfying the Cordes condition. The second class of methods is motivated by the notion of strong solutions (see Section 2.4) under the assumption that $A \in C(\bar{\Omega}, \mathbb{S}^d)$. Finally, the third method is motivated by the notion of viscosity solutions (see Section 2.5), where comparison principles and monotonicity of the scheme are the central themes.

4.1. Discretization of non-divergence-form PDEs satisfying the Cordes condition

Here we discuss recent numerical methods for second-order elliptic PDEs in non-divergence form satisfying the Cordes condition (2.15). Recall from Section 2.4.1 that, if the domain is convex, this condition on the coefficient matrix ensures that the bilinear mapping

$$a(\cdot,\cdot) : (H^2(\Omega) \cap H_0^1(\Omega))^2 \ni (v, w) \to a(v, w) = \int_\Omega \gamma \mathcal{L}v \Delta w \tag{4.2}$$

is coercive, thus allowing one to define strong solutions via variational principles. Here, the function γ is given by (2.13). In this setting the existence of a strong solution $u \in H^2(\Omega) \cap H_0^1(\Omega)$ to (4.1) is deduced from the variational

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

formulation

$$a(u,v) = \int_\Omega \gamma f \Delta v \, dx, \quad \text{for all } v \in H^2(\Omega) \cap H_0^1(\Omega) \tag{4.3}$$

and appealing to the Lax–Milgram lemma. Since the mapping $\Delta : H^2(\Omega) \cap H_0^1(\Omega) \to L^2(\Omega)$ is surjective on convex domains, one concludes that a function $u \in H^2(\Omega) \cap H_0^1(\Omega)$ satisfying (4.3) satisfies (4.1) almost everywhere, that is, $u \in H^2(\Omega) \cap H_0^1(\Omega)$ is a strong solution to (4.1). We refer the reader to Theorem 2.39 for details.

One immediately sees that the solution concept lends itself to a finite element approximation which would pose (4.3) over a finite-dimensional subspace of $H^2(\Omega) \cap H_0^1(\Omega)$ consisting of piecewise polynomials. To describe this procedure, we let $\mathcal{T}_h$ denote a simplicial, conforming and shape-regular triangulation of Ω parametrized by $h > 0$, and let $X_h^{cd} \subset H^2(\Omega) \cap H_0^1(\Omega)$ be a finite-dimensional subspace consisting of piecewise polynomials with respect to $\mathcal{T}_h$. A conforming finite element approximation to (4.3) seeks a function $u_h \in X_h^{cd}$ satisfying the discrete variational formulation

$$a(u_h, v_h) = \int_\Omega \gamma f \Delta v_h, \quad \text{for all } v_h \in X_h^{cd}. \tag{4.4}$$

Problem (4.4) represents a square linear system of equations. The coercivity of the bilinear form $a(\cdot,\cdot)$ over $H^2(\Omega) \cap H_0^1(\Omega)$ implies that this system is invertible, and thus there exists a unique solution $u_h \in X_h^{cd}$ to problem (4.4). Moreover, the continuity of $a(\cdot,\cdot)$ and Céa's lemma show that such approximations are quasi-optimal in the sense that

$$\|u - u_h\|_{H^2(\Omega)} \leq \frac{C}{\alpha} \inf_{v_h \in X_h^{cd}} \|u - v_h\|_{H^2(\Omega)},$$

where $C > 0$ and $\alpha = 1 - \sqrt{1-\epsilon}$ are the continuity and coercivity constants of $a(\cdot,\cdot)$, respectively.

While method (4.4) is a stable and convergent numerical scheme for computing solutions to (4.3), it has some potential practical drawbacks. Piecewise polynomial subspaces of $H^2(\Omega)$ are difficult to construct and implement, and are not a practical option for solving second-order PDEs. These properties are further exacerbated in three dimensions, where, for example, polynomials of degree of at least nine are required to construct H^2-conforming finite element spaces on general simplicial partitions; see Lai and Schumaker (2007, Remark 1) and Ženíšek (1973). Nevertheless, this path is explored by Gallistl (2017), who also presents a mixed formulation.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

4.1.1. C^0 finite element approximations

We now discuss finite element methods for problem (4.1) that use continuous basis functions, commonly used for second-order problems in divergence form. To this end, with $\mathcal{T}_h$ given in the previous section, we define the Lagrange finite element space

$$X_h^{cg} = \{v_h \in H_0^1(\Omega) : v_h\big|_T \in \mathbb{P}_k(T)\ \forall T \in \mathcal{T}_h\} \tag{4.5}$$

with $k \in \mathbb{N}$. Note that functions in X_h^{cg} are locally smooth, yet not globally $H^2(\Omega)$, and therefore second-order derivatives are only defined piecewise with respect to $\mathcal{T}_h$. To simplify the presentation, we shall write

$$\|v\|_{L^2(\mathcal{T}_h)} := \left(\sum_{K\in\mathcal{T}_h} \|v\|_{L^2(K)}^2\right)^{1/2}$$

for piecewise L^2-functions.

Since $X_h^{cg} \not\subset H^2(\Omega) \cap H_0^1(\Omega)$, the bilinear form $a(\cdot,\cdot)$ defined by (4.2) is not well-defined on $X_h^{cg} \times X_h^{cg}$. Moreover, a piecewise version of the bilinear form is not generally coercive on X_h^{cg}, since, for example,

$$\sum_{K\in\mathcal{T}_h} \int_K \gamma \mathcal{L} v_h \Delta v_h = 0$$

for all piecewise linear $v_h \in X_h^{cg}$. This stems from the fact that a piecewise version of the Miranda–Talenti estimate $\|D^2 v\|_{L^2(\Omega)} \leq \|\Delta v\|_{L^2(\Omega)}$ is not satisfied on X_h^{cg}. As such, the coercivity proof of $a(\cdot,\cdot)$ found at the continuous level does not directly carry over to the discrete setting.

To overcome this we develop a discrete Miranda–Talenti estimate suitable for piecewise polynomials. To do so, we introduce some notation. Let $\mathcal{F}_h^I$ and $\mathcal{F}_h^B$ denote the set of interior and boundary edges/faces, respectively, and set $\mathcal{F}_h := \mathcal{F}_h^I \cup \mathcal{F}_h^B$ and $h_F := \operatorname{diam}(F)$ for $F \in \mathcal{F}_h$. We recall that the jump of a vector-valued function $\mathbf{w}$ is given by (3.48).

A discrete Miranda–Talenti estimate is based on the following result, whose proof can be found in Georgoulis, Houston and Virtanen (2011).

Lemma 4.1 (enrichment operator). Suppose that $d = 2$. Then there exists a finite-dimensional space $X_h^{cd} \subset H^2(\Omega) \cap H_0^1(\Omega)$ and an (enrichment) operator $E_h : X_h^{cg} \to X_h^{cd}$ satisfying

$$\|D^2(v_h - E_h v_h)\|_{L^2(\mathcal{T}_h)} \leq C\left(\sum_{F\in\mathcal{F}_h^I} h_F^{-1} \big\|[\![Dv_h]\!]\big\|_{L^2(F)}^2\right)^{1/2} \tag{4.6}$$

for all $v_h \in X_h^{cg}$. Here, the constant $C > 0$ depends on the polynomial degree k and the shape-regularity of $\mathcal{T}_h$, but is independent of h.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Theorem 4.2 (discrete Miranda–Talenti estimate). Suppose that $\Omega \subset \mathbb{R}^2$ is a convex polygon. Then, for all $v_h \in X_h^{cg}$,

$$\|D^2 v_h\|_{L^2(\mathcal{T}_h)} \le \|\Delta v_h\|_{L^2(\mathcal{T}_h)} + C\Big(\sum_{F\in\mathcal{F}_h^I} h_F^{-1}\big\|[[Dv_h]]\big\|_{L^2(F)}^2\Big)^{1/2}. \tag{4.7}$$

Proof. Fix $v_h \in X_h^{cg}$ and let $E_h : X_h^{cg} \to X_h^{cd}$ be the enrichment operator of Lemma 4.1. Since $E_h v_h \in H^2(\Omega)\cap H_0^1(\Omega)$ and Ω is convex, the Miranda–Talenti estimate

$$\|D^2 E_h v_h\|_{L^2(\Omega)} \le \|\Delta E_h v_h\|_{L^2(\Omega)} = \|\Delta E_h v_h\|_{L^2(\mathcal{T}_h)}$$

is satisfied. Applying the triangle inequality, the inequality $\|\Delta E_h v_h\|_{L^2(\Omega)} \le \sqrt{2}\|D^2 E_h v_h\|_{L^2(\Omega)}$, and estimate (4.6) yields

$$\begin{aligned}\|D^2 v_h\|_{L^2(\mathcal{T}_h)} &\le \|D^2 E_h v_h\|_{L^2(\Omega)} + \|D^2(v_h - E_h v_h)\|_{L^2(\mathcal{T}_h)}\\ &\le \|\Delta E_h v_h\|_{L^2(\Omega)} + \|D^2(v_h - E_h v_h)\|_{L^2(\mathcal{T}_h)}\\ &\le \|\Delta v_h\|_{L^2(\mathcal{T}_h)} + C\Big(\sum_{F\in\mathcal{F}_h^I} h_F^{-1}\big\|[[Dv_h]]\big\|_{L^2(F)}^2\Big)^{1/2},\end{aligned}$$

where in the last step we used that

$$\begin{aligned}\|\Delta E_h v_h\|_{L^2(\Omega)} &\le \|\Delta v_h\|_{L^2(\mathcal{T}_h)} + \|\Delta(E_h v_h - v_h)\|_{L^2(\mathcal{T}_h)}\\ &\le \|\Delta v_h\|_{L^2(\mathcal{T}_h)} + \sqrt{2}\|D^2(E_h v_h - v_h)\|_{L^2(\mathcal{T}_h)}.\end{aligned}$$

This concludes the proof. □

Motivated by Theorem 4.2, we define the bilinear form

$$a_h(v,w) := \sum_{K\in\mathcal{T}_h}\int_K \gamma\mathcal{L}v\Delta w + \sum_{F\in\mathcal{F}_h^I}\mu h_F^{-1}\int_F [[Dv]]\,[[Dw]], \tag{4.8}$$

where μ is a positive penalty parameter. We also define the discrete H^2-type norm

$$\|v\|_{H_h^2(\Omega)}^2 := \|\Delta v\|_{L^2(\mathcal{T}_h)}^2 + \sum_{F\in\mathcal{F}_h^I} h_F^{-1}\big\|[[Dv]]\big\|_{L^2(F)}^2. \tag{4.9}$$

We consider the following finite element method: find $u_h \in X_h^{cg}$ satisfying

$$a_h(u_h, v_h) = \sum_{K\in\mathcal{T}_h}\int_K \gamma f\Delta v_h, \quad \text{for all } v_h \in X_h^{cg}. \tag{4.10}$$

Note that the additional penalization term does not affect the consistency of the scheme; that is, we obtain $a_h(u, v_h) = \sum_{K\in\mathcal{T}_h}\int_K \gamma f\Delta v_h$ for all $v_h \in X_h^{cg}$. The role of this term is to weakly enforce H^2-regularity and to ensure that the bilinear form $a_h(\cdot,\cdot)$ is coercive provided μ is sufficiently large.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Lemma 4.3 (coercivity). Suppose that $\Omega \subset \mathbb{R}^2$ is convex and that $A \in L^\infty(\Omega, \mathbb{S}^2)$ satisfies the Cordes condition (2.15) with parameter ϵ. There exists $\mu_* > 0$ depending on the shape-regularity of the mesh, polynomial degree k, and the parameter ϵ such that, for $\mu \geq \mu_*$,

$$\frac{\alpha}{2}\|v_h\|^2_{H^2_h(\Omega)} \leq a_h(v_h, v_h), \quad \text{for all } v_h \in X_h^{cg},$$

where $\alpha = 1 - \sqrt{1-\epsilon}$ is the coercivity constant of $a(\cdot,\cdot)$.

Proof. Adding a subtracting Δv_h and applying the Cordes condition yields

$$\begin{aligned} a_h(v_h, v_h) &= \sum_{K\in\mathcal{T}_h} \int_K \gamma \mathcal{L} v_h \Delta v_h + \mu \sum_{F\in\mathcal{F}_h^I} h_F^{-1} \big\| [[Dv_h]] \big\|^2_{L^2(F)} \\ &= \|\Delta v_h\|^2_{L^2(\mathcal{T}_h)} + \sum_{K\in\mathcal{T}_h} \int_K (\gamma\mathcal{L}v_h - \Delta v_h)\Delta v_h \\ &\quad + \mu \sum_{F\in\mathcal{F}_h^I} h_F^{-1} \big\| [[Dv_h]] \big\|^2_{L^2(F)} \\ &\geq \|\Delta v_h\|^2_{L^2(\mathcal{T}_h)} - \sqrt{1-\epsilon}\|\Delta v_h\|_{L^2(\mathcal{T}_h)}\|D^2 v_h\|_{L^2(\mathcal{T}_h)} \\ &\quad + \mu \sum_{F\in\mathcal{F}_h^I} h_F^{-1} \big\| [[Dv_h]] \big\|^2_{L^2(F)}. \end{aligned}$$

Applying the discrete Miranda–Talenti estimate and the Cauchy–Schwarz inequality, we obtain, for any $\tau > 0$,

$$\begin{aligned} a_h(v_h, v_h) &\geq \left(\alpha - \frac{\tau}{2}\sqrt{1-\epsilon}\right)\|\Delta v_h\|^2_{L^2(\mathcal{T}_h)} \\ &\quad + \left(\mu - \frac{1}{2\tau}\sqrt{1-\epsilon}\right) \sum_{F\in\mathcal{F}_h^I} h_F^{-1} \big\| [[Dv_h]] \big\|^2_{L^2(F)}. \end{aligned}$$

Taking $\tau = \alpha/\sqrt{1-\epsilon} = 1/\sqrt{1-\epsilon} - 1$ and $\mu_* = \alpha/2 + \sqrt{1-\epsilon}(1 - 1/\alpha)$ yields

$$\frac{\alpha}{2}\|v_h\|^2_{H^2_h(\Omega)} \leq a_h(v_h, v_h). \qquad \square$$

The coercivity stated in Lemma 4.3 shows that there exists a unique solution to the finite element method (4.10). Combined with the consistency of the scheme, we immediately obtain quasi-optimal error estimates in the discrete H^2-norm.

Theorem 4.4 (existence and error estimates). Suppose that $\Omega \subset \mathbb{R}^2$ is convex and that $A \in L^\infty(\Omega, \mathbb{S}^2)$ satisfies the Cordes condition (2.15) with parameter ϵ. Suppose that $\mu \geq \mu_*$, and let $u_h \in X_h^{cg}$ be the unique solution to (4.10) with X_h^{cg} given by (4.5). Then, if the solution to (4.1) satisfies

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

$u \in H^s(\Omega)$ with $2 \leq s \leq k+1$,

$$\alpha \|u - u_h\|_{H_h^2(\Omega)} \leq C \inf_{v_h \in X_h^{cg}} \|u - v_h\|_{H_h^2(\Omega)} \leq C h^{s-2} \|u\|_{H^s(\Omega)},$$

where $\alpha = 1 - \sqrt{1-\epsilon}$.

Remark 4.5 (linear case). Note that in the piecewise linear case ($k = 1$), Theorem 4.4 does not give a convergence result. In fact, it is easy to see that in this case the solution to (4.10) is the trivial one $u_h \equiv 0$.

Remark 4.6 (three dimensions). The results in Lemma 4.3 and Theorem 4.4 are restricted to the two-dimensional case due to Lemma 4.1. If there exists an enrichment operator satisfying (4.6) with $d = 3$, then these results carry over to the three-dimensional case.

4.1.2. Discontinuous Galerkin approximations

In this section we summarize the discretization developed and analysed in in Smears and Süli (2013), where a consistent discontinuous Galerkin (DG) method is constructed. Instead of developing a discrete Miranda–Talenti via penalization, the key idea of this approach is to add auxiliary terms in the bilinear form to bypass the Miranda–Talenti estimate found at the continuous level.

For $k \in \mathbb{N}$, define the piecewise polynomial space without continuity:

$$X_h^{dg} = \{v \in L^2(\Omega) : v|_T \in \mathbb{P}_k(T) \ \forall T \in \mathcal{T}_h\}. \tag{4.11}$$

We note that the method in Smears and Süli (2013) considers discretizations in an *hp*-framework where the polynomial degree is element-dependent. For simplicity, and to ease the presentation, we consider here only the h-version of the method, where the polynomial degree is globally fixed and we do not trace the dependence of the constants on the polynomial degree k.

To motivate the method, we again emphasize that a Miranda–Talenti estimate fails to hold for piecewise polynomials, and as such, the coercivity proof found at the continuous levels fails in the discrete setting. Indeed, mimicking the calculations in Lemma 2.38 element-wise over X_h^{dg} leads to

$$\sum_{K \in \mathcal{T}_h} \int_K \gamma \mathcal{L} v_h \Delta v_h = \sum_{K \in \mathcal{T}_h} \int_K \Delta v_h \Delta v_h + \sum_{K \in \mathcal{T}_h} \int_K (\gamma \mathcal{L} v_h - \Delta v_h) \Delta v_h. \tag{4.12}$$

Applying the Cordes condition and the Cauchy–Schwarz inequality yields the inequality

$$\sum_{K \in \mathcal{T}_h} \int_K \gamma \mathcal{L} v_h \Delta_h v \geq \|\Delta v_h\|^2_{L^2(\mathcal{T}_h)} - \sqrt{1-\epsilon} \|D^2 v_h\|_{L^2(\mathcal{T}_h)} \|\Delta v_h\|_{L^2(\mathcal{T}_h)}.$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Since the piecewise Hessian matrix of v_h cannot be controlled by its piecewise Laplacian (*e.g.* if v_h is piecewise harmonic), one concludes that the bilinear mapping $(v_h, w_h) \to \sum_{K\in\mathcal{T}_h} \int_K \gamma \mathcal{L} v_h \Delta w_h$ is not coercive over $X_h^{dg} \times X_h^{dg}$ in general.

The essential idea presented in Smears and Süli (2013) is to replace the bilinear form $(v_h, w_h) \to \sum_{K\in\mathcal{T}_h} \int_K \Delta v_h \Delta w_h$ implicit in the right-hand side of (4.12) with a consistent bilinear form that is coercive with a discrete H^2-type norm.

For a face $F \in \mathcal{F}_h$, let $\{t_i\}_{i=1}^{d-1}$ be an orthonormal coordinate system, and define the tangential gradient, tangential divergence, and tangential Laplacian, respectively, as

$$D_T v = \sum_{i=1}^{d-1} t_i \frac{\partial v}{\partial t_i}, \quad D_T \cdot \mathbf{w} = \sum_{i=1}^{d-1} \frac{\partial w_i}{\partial t_i}, \quad \Delta_T v = D_T \cdot D_T v.$$

We also define the average of a scalar or vector-valued function as

$$\{\!\{v\}\!\}|_F := \begin{cases} \frac{1}{2}(v_+ + v_-) & \text{if } F = \partial K^+ \cap \partial K^- \in \mathcal{F}_h^I, \\ v_+ & \text{if } F = \partial K^+ \cap \partial\Omega \in \mathcal{F}_h^B. \end{cases} \tag{4.13}$$

Then define the bilinear form

$$\begin{aligned} B_h(u_h, v_h) = &\sum_{K\in\mathcal{T}_h} \int_K \left(D^2 u_h : D^2 v_h + \Delta u_h \Delta v_h\right) \\ &+ \sum_{F\in\mathcal{F}_h^I} \int_F \left(\{\!\{\Delta_T u_h\}\!\}[\![Dv_h]\!] + \{\!\{\Delta_T v_h\}\!\}[\![Du_h]\!]\right) \\ &- \sum_{F\in\mathcal{F}_h} \int_F \left(D_T\{\!\{Du_h \cdot n\}\!\} \cdot [\![D_T v_h]\!]_T + D_T\{\!\{Dv_h \cdot n\}\!\} \cdot [\![D_T u_h]\!]_T\right) \\ &+ \sum_{F\in\mathcal{F}_h^I} \mu h_F^{-1} \int_F [\![Du_h]\!][\![Dv_h]\!] + \sum_{F\in\mathcal{F}_h} \mu h_F^{-3} \int_F [\![u_h]\!] \cdot [\![v_h]\!], \end{aligned} \tag{4.14}$$

where $[\![\mathbf{v}]\!]_T|_F := \mathbf{v}_+ - \mathbf{v}_-$ on $\mathcal{F}_h^I$ and $[\![\mathbf{v}]\!]_T|_F := \mathbf{v}_+$ on $\mathcal{F}_h^B$, and $\mu > 0$ is a penalization parameter.

For $\theta \in [0, 1]$, let us define the discrete DG-norm

$$\begin{aligned} \|v\|_{DG(\theta)}^2 = &(1-\theta)\|\Delta v\|_{L^2(\mathcal{T}_h)}^2 + \theta \|D^2 v\|_{L^2(\mathcal{T}_h)}^2 \\ &+ c_* \left(\sum_{F\in\mathcal{F}_h^I} h_F^{-1} \left\| [\![Dv]\!] \right\|_{L^2(F)}^2 + \sum_{F\in\mathcal{F}_h} h_F^{-3} \left\| [\![v]\!] \right\|_{L^2(F)}^2 \right). \end{aligned} \tag{4.15}$$

The seemingly abstruse bilinear form $B_h(\cdot,\cdot)$ is carefully defined to satisfy the following properties (Smears and Süli 2013, Lemmas 5 and 7).

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Lemma 4.7 (properties of B_h). The bilinear form $B_h : X_h^{dg} \times X_h^{dg} \to \mathbb{R}$, defined in (4.14), satisfies the following properties.

- *Consistency.* If $u \in H^s(\Omega) \cap H_0^1(\Omega)$ for some $s > 5/2$, then
$$B_h(u, v_h) = 2 \sum_{K \in \mathcal{T}_h} \int_K \Delta u \Delta v_h$$
for all $v_h \in X_h^{dg}$.
- *Coercivity.* For any $\kappa > 1$, there exists a $\mu_* = C\kappa/(\kappa - 1)$ with $C > 0$ depending only on the shape-regularity of $\mathcal{T}_h$ and k such that for $\mu \geq \mu_*$,
$$2\|v_h\|_{DG(1/2)}^2 \leq \kappa B_h(v_h, v_h), \quad \text{for all } v_h \in X_h^{dg}, \tag{4.16}$$
for some constant $c_* > 0$ independent of the discretization parameter h and polynomial degree k.

The properties of $B_h(\cdot, \cdot)$ shown above allow us to define the following DG method: find $u_h \in X_h^{dg}$ such that, for all $v_h \in X_h^{dg}$,
$$\begin{aligned} a_h^{DG}(u_h, v_h) &:= \sum_{K \in \mathcal{T}_h} \int_K \gamma(\mathcal{L}u_h - \Delta u_h)\Delta v_h + \frac{1}{2} B_h(u_h, v_h) \\ &= \sum_{K \in \mathcal{T}_h} \int_K \gamma f \Delta v_h. \end{aligned} \tag{4.17}$$

Due to the consistency of $B_h(\cdot, \cdot)$ we see that the scheme is consistent provided the exact solution is sufficiently smooth. If u is the solution to (4.1) and satisfies $u \in H^s(\Omega)$ for some $s > 5/2$, then $a_h^{DG}(u, v_h) = \int_\Omega \gamma f \Delta_h v_h$ for all $v_h \in X_h^{dg}$. In addition, the coercivity of $B_h(\cdot, \cdot)$ implies the coercivity of $a_h^{DG}(\cdot, \cdot)$; see Smears and Süli (2013, Theorem 8).

Lemma 4.8 (coercivity). Suppose that $\Omega \subset \mathbb{R}^d$ is convex and that A satisfies the Cordes condition (2.15). Then there exists $\mu_* = O(\epsilon^{-1}) > 0$ such that
$$C\|v_h\|_{DG(1)}^2 \leq a_h^{DG}(v_h, v_h), \quad \text{for all } v_h \in X_h^{dg}.$$
Consequently, there exists a unique solution $u_h \in X_h^{dg}$ to (4.17).

Combined with consistency of $a_h^{DG}(\cdot, \cdot)$, Lemma 4.8 implies quasi-optimal error estimates in the discrete H^2-type norm (Smears and Süli 2013, Theorem 9).

Theorem 4.9 (existence and error estimates). Suppose the hypotheses of Lemma 4.8 hold. Further, suppose that the solution to (4.1) satisfies $u \in H^s(\Omega)$ for some $5/2 < s \leq k+1$. Then there exists an h-independent

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

constant $C > 0$ such that

$$\|u - u_h\|_{DG(1)} \leq Ch^{s-2}\|u\|_{H^s(\Omega)}.$$

Remark 4.10 (regularity). The regularity assumption $u \in H^s(\Omega)$ with $s > 5/2$ ensures that $a_h^{DG}(u, v_h)$ is well-defined.

Remark 4.11 (extensions). Wang and Wang (2017) recently proposed and analysed a primal–dual discontinuous Galerkin method for second-order elliptic equations in non-divergence form.

4.2. Discrete finite element Calderón–Zygmund estimates

In this section we describe finite element discretizations to non-divergence-form elliptic operators based on the notion and theory of strong solutions. Recall from Definition 2.28 that a function $u \in W^{2,p}(\Omega)$ is a strong solution to (4.1) if the equation and boundary conditions hold almost everywhere in $\bar{\Omega}$. Such solutions exist provided the data is sufficiently regular and $A \in C(\bar{\Omega}, \mathbb{S}^d)$; see Theorem 2.29. This result is obtained by using the Calderón–Zygmund decomposition technique when $A = I$, and then extended to general $\mathcal{L}$ using the continuity of the coefficient matrix. In this section, we develop a discrete version of this theory to develop *a priori* estimates and convergence results of finite element solutions. Let us first present the derivation of the method.

Assume for the moment that the coefficient matrix A in (4.1) is sufficiently smooth. Then, as explained in Example 2.8, we can write problem (4.1) in divergence form:

$$D \cdot (ADu) - (D \cdot A) \cdot Du = f \quad \text{in } \Omega.$$

A standard finite element method for this problem (without stabilization) reads as follows: find $u_h \in X_h^{cg}$ such that

$$-\int_\Omega (ADu_h) \cdot Dv_h - \int_\Omega ((D \cdot A) \cdot Du_h)v_h = \int_\Omega f v_h, \quad \text{for all } v_h \in X_h^{cg}, \tag{4.18}$$

where $X_h^{cg} \subset H_0^1(\Omega)$ is the Lagrange finite element space of degree $k \geq 1$ defined by (3.34). It is well known that, for h sufficiently small, there exists a unique solution to (4.18).

If A is not sufficiently smooth and/or if the locations of the singularities are complex/unknown, then the classical finite element method (4.18) is not viable due to the differential operators acting on A. However, we easily circumvent this issue by using the integration-by-parts identity

$$\int_\Omega \boldsymbol{\tau} \cdot Dv_h = -\sum_{K \in \mathcal{T}_h} \int_K (D \cdot \boldsymbol{\tau})v_h + \sum_{F \in \mathcal{F}_h^I} \int_F [[\boldsymbol{\tau}]]v_h,$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

which holds for all piecewise smooth $\boldsymbol{\tau}$ and $v_h \in X_h^{cg}$. Taking $\boldsymbol{\tau} = ADu_h$ and applying the product rule yields

$$-\int_\Omega (ADu_h)\cdot Dv_h = \sum_{K\in\mathcal{T}_h}\int_K (A : D^2u_h)v_h + \int_\Omega ((D\cdot A)\cdot Du_h)v_h + \sum_{F\in\mathcal{F}_h^I}\int_F [[ADu_h]]v_h.$$

Substituting this identity into the (ill-posed) formulation leads to the finite element method

$$b_h(u_h,v_h) := \sum_{K\in\mathcal{T}_h}\int_K (A : D^2u_h)v_h - \sum_{F\in\mathcal{F}_h^I}\int_F [[ADu_h]]v_h = \int_\Omega fv_h \quad (4.19)$$

for all $v_h \in X_h^{cg}$. In contrast to (4.18), the formulation (4.19) is well-defined for non-differentiable A. Furthermore, by reversing the arguments, we see that (4.19) is equivalent to (4.18) if A is sufficiently smooth. In particular, when $A(x) \equiv \bar{A}$ is a constant SPD matrix, the method (4.19) reduces to the (well-posed) problem

$$b_{h,0}(u_h,v_h) := -\int_\Omega (\bar{A}Du_h)\cdot Dv_h = \int_\Omega fv_h, \quad \text{for all } v_h \in X_h^{cg}. \quad (4.20)$$

We point out that method (4.19) is consistent and meaningful in the piecewise linear case ($k = 1$).

While the derivation of the finite element method (4.19) is relatively simple, a stability and convergence analysis of the method is less obvious. The key difficulty is that integration by parts is not at our disposal, and it is unclear whether a clever choice of test function will render a coercivity or inf-sup condition. Rather, the stability analysis of (4.19) mimics the techniques found in the PDE theory, where Calderón–Zygmund estimates are the essential tools.

To describe the stability and convergence theory, we first define a discrete $W^{2,p}$-type norm:

$$\|v\|^p_{W_h^{2,p}(\Omega)} := \sum_{K\in\mathcal{T}_h}\|D^2v\|^p_{L^p(K)} + \sum_{F\in\mathcal{F}_h^I} h_F^{1-p}\big\|[[Dv]]\big\|^p_{L^p(F)}, \quad \text{for } 1<p<\infty.$$

A discrete Calderón–Zygmund-type estimate with respect to this norm in the case of constant coefficients is now given.

Lemma 4.12 (discrete Calderón–Zygmund estimate). Let L be the elliptic, divergence-form operator (2.4), where the coefficient matrix A is constant and SPD. Suppose that the *a priori* estimate $C\|w\|_{W^{2,p}(\Omega)} \le \|Lw\|_{L^p(\Omega)}$ is satisfied for all $w \in W^{2,p}(\Omega)\cap W_0^{1,p}(\Omega)$. Let $b_{h,0}(\cdot,\cdot)$ be defined

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

by (4.20). Then, for h sufficiently small,

$$C\|v_h\|_{W_h^{2,p}(\Omega)} \leq \sup_{w_h \in X_h^{cg}\setminus\{0\}} \frac{b_{h,0}(v_h, w_h)}{\|w_h\|_{L^{p'}(\Omega)}}, \quad \text{for all } v_h \in X_h^{cg},$$

where $1/p + 1/p' = 1$.

Proof. We prove the simpler case $p = 2$, and refer the reader to Feng, Hennings and Neilan (2017, Lemma 2.6) and Neilan (2013, Lemma 4.1) for general $p \in (1, \infty)$.

First, let $\mathcal{B}_h(v_h) \in X_h^{cg}$ be the unique solution to the problem

$$\int_\Omega \mathcal{B}_h(v_h) w_h \, dx = b_{h,0}(v_h, w_h), \quad \text{for all } w_h \in X_h^{cg},$$

and let $\varphi \in H_0^1(\Omega)$ be the unique (weak) solution to $L\varphi = -\mathcal{B}_h(v_h)$ in Ω. Then, for $w_h \in X_h^{cg}$, we find

$$b_{h,0}(v_h, w_h) = \int_\Omega \mathcal{B}_h(v_h) w_h = -\int_\Omega (AD\varphi) \cdot Dw_h = b_{h,0}(\varphi, w_h).$$

Thus, v_h is the elliptic projection of φ with respect to $b_{h,0}(\cdot, \cdot)$. Therefore, Céa's lemma and the hypothesis $\varphi \in H^2(\Omega)$ with

$$\|\varphi\|_{H^2(\Omega)} \leq C\|\mathcal{B}_h(v_h)\|_{L^2(\Omega)}$$

yield

$$\|\varphi - v_h\|_{H^1(\Omega)} \leq Ch\|\varphi\|_{H^2(\Omega)} \leq Ch\|\mathcal{B}_h(v_h)\|_{L^2(\Omega)}. \tag{4.21}$$

Next, for any $\varphi_h \in X_h^{cg}$, the triangle inequality and a scaling argument show that

$$\begin{aligned}\|v_h\|_{H_h^2(\Omega)} &\leq \|v_h - \varphi_h\|_{H_h^2(\Omega)} + \|\varphi_h - \varphi\|_{H_h^2(\Omega)} + \|\varphi\|_{H_h^2(\Omega)} \\ &\leq Ch^{-1}(\|v_h - \varphi\|_{H^1(\Omega)} + \|\varphi - \varphi_h\|_{H^1(\Omega)}) + \|\varphi_h - \varphi\|_{H_h^2(\Omega)} + \|\varphi\|_{H^2(\Omega)}.\end{aligned}$$

By taking φ_h to be the nodal interpolant of φ and applying (4.21) and the definition of $\mathcal{B}_h(v_h)$, we obtain

$$\begin{aligned}\|v_h\|_{H_h^2(\Omega)} &\leq C\big[h^{-1}\|v_h - \varphi\|_{H^1(\Omega)} + \|\varphi\|_{H^2(\Omega)}\big] \leq C\|\mathcal{B}_h(v_h)\|_{L^2(\Omega)} \\ &= C \sup_{w_h \in X_h^{cg}\setminus\{0\}} \frac{\int_\Omega \mathcal{B}_h(v_h) w_h}{\|w_h\|_{L^2(\Omega)}} = C \sup_{w_h \in X_h^{cg}\setminus\{0\}} \frac{b_{h,0}(v_h, w_h)}{\|w_h\|_{L^2(\Omega)}}. \qquad \square\end{aligned}$$

Remark 4.13 ($p = 2$). Note that when Ω is convex, the assumptions of Lemma 4.12 hold for $p = 2$.

A corollary of this result is a local stability estimate of the discrete adjoint problem.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Corollary 4.14 (local stability). For a domain $D \subset \Omega$, let ρ_D denote the radius of the largest ball inscribed in D, and let $X_h^{cg}(D)$ denote the set of functions in X_h^{cg} that vanish outside D. Suppose that the hypothesis of Lemma 4.12 are satisfied. Then, if h and ρ_D are sufficiently small,

$$C\|w_h\|_{L^{p'}(D)} \leq \sup_{v_h \in X_h^{cg}(D_h)\setminus\{0\}} \frac{b_h(v_h, w_h)}{\|v_h\|_{W_h^{2,p}(D_h)}}, \quad \text{for all } w_h \in X_h^{cg}(D).$$

with $D_h = \{x \in \Omega : \operatorname{dist}(x, D) \leq h\}$.

Proof. Again, we prove the case $p = p' = 2$ and refer the reader to Feng *et al.* (2017, Appendix B) for the general result.

Let $w_h \in X_h^{cg}(D)$, and let $\varphi_h \in X_h^{cg}$ satisfy

$$b_{h,0}(\varphi_h, v_h) = \int_\Omega w_h v_h, \quad \text{for all } v_h \in X_h^{cg},$$

where $b_{h,0}(\cdot,\cdot)$ is defined by (4.20) with

$$\bar{A} = \frac{1}{|D|}\int_D A.$$

Taking $v_h = w_h$ in the method yields

$$\|w_h\|^2_{L^2(D)} = b_{h,0}(\varphi_h, w_h) = b_h(\varphi_h, w_h) + (b_{h,0}(\varphi_h, w_h) - b_h(\varphi_h, w_h)).$$

Using the (uniform) continuity of A, for any $\tau > 0$, we have

$$|b_{h,0}(\varphi_h, w_h) - b_h(\varphi_h, w_h)| \leq \tau\|\varphi_h\|_{H_h^2(\Omega)}\|w_h\|_{L^2(D)}$$

provided ρ_D is sufficiently small.

Applying the estimate

$$\|\varphi_h\|_{H_h^2(\Omega)} \leq C\|w_h\|_{L^2(\Omega)} = C\|w_h\|_{L^2(D)}$$

established in Theorem 4.12, we obtain

$$\begin{aligned}(1 - C\tau)\|w_h\|^2_{L^2(D)} \leq b_h(\varphi_h, w_h) &= \left(\frac{b_h(\varphi_h, w_h)}{\|\varphi_h\|_{H_h^2(D_h)}}\right)\|\varphi\|_{H_h^2(\Omega)} \\ &\leq C\left(\sup_{v_h \in X_h^{cg}(D_h)\setminus\{0\}} \frac{b_h(v_h, w_h)}{\|v_h\|_{H_h^2(D_h)}}\right)\|w_h\|_{L^2(\Omega)}.\end{aligned}$$

Taking τ sufficiently small and manipulating terms in the last inequality yields the result. □

The local stability result for the discrete adjoint problem given in Corollary 4.14 leads to a stability estimate for method (4.19).

Theorem 4.15 (global stability). Suppose that $A \in C(\bar{\Omega}, \mathbb{S}^d)$, and that elliptic and divergence-form operators with constant coefficients inherit

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

$W^{2,p}$-regularity ($1 < p < \infty$). Then there exists $h_* > 0$ depending on the modulus of continuity of A and p such that, for $h \le h_*$,

$$C\|v_h\|_{W_h^{2,p}(\Omega)} \le \sup_{w_h \in X_h^{cg}\setminus\{0\}} \frac{b_h(v_h, w_h)}{\|w_h\|_{L^{p'}(\Omega)}}, \quad \text{for all } v_h \in X_h^{cg}. \tag{4.22}$$

Proof. We outline the main steps of the proof and refer to Feng *et al.* (2017) for details.

Combining Corollary 4.14 with cut-off functions techniques and a covering argument leads to the Gärding-type inequality

$$C\|w_h\|_{L^{p'}(\Omega)} \le \sup_{v_h \in X_h^{cg}\setminus\{0\}} \frac{b_h(v_h, w_h)}{\|v_h\|_{W_h^{2,p}(\Omega)}} + \|w_h\|_{W^{-1,p'}(\Omega)}, \quad \text{for all } w_h \in X_h^{cg}.$$

A standard duality argument then shows that, for h sufficiently small,

$$C\|w_h\|_{L^{p'}(\Omega)} \le \sup_{v_h \in X_h^{cg}\setminus\{0\}} \frac{b_h(v_h, w_h)}{\|v_h\|_{W_h^{2,p}(\Omega)}}, \quad \text{for all } w_h \in X_h^{cg}.$$

This estimate shows that, for fixed $v_h \in X_h^{cg}$, there exists a unique $w_h \in X_h^{cg}$ satisfying

$$\begin{aligned} b_h(z_h, w_h) = &\sum_{K\in\mathcal{T}_h} \int_K |D^2 v_h|^{p-2} D^2 v_h : D^2 z_h \\ &+ \sum_{F\in\mathcal{F}_h} h_F^{-1} \int_F |[[Dv_h]]|^{p-2} [[Dv_h]][[Dz_h]], \quad \text{for all } z_h \in X_h^{cg}. \end{aligned} \tag{4.23}$$

Applying the global stability estimate for the adjoint problem and Hölder's inequality, we obtain

$$C\|w_h\|_{L^{p'}(\Omega)} \le \sup_{z_h \in X_h^{cg}\setminus\{0\}} \frac{b_h(z_h, w_h)}{\|z_h\|_{W_h^{2,p}(\Omega)}} \le C\|v_h\|_{W_h^{2,p}(\Omega)}^{p-1}.$$

On the other hand, setting $z_h = v_h$ in (4.23) yields

$$\begin{aligned} \|v_h\|_{W_h^{2,p}(\Omega)}^p = b_h(v_h, w_h) &\le \left(\sup_{z_h \in X_h^{cg}\setminus\{0\}} \frac{b_h(v_h, z_h)}{\|z_h\|_{L^{p'}(\Omega)}} \right) \|w_h\|_{L^{p'}(\Omega)} \\ &\le C \left(\sup_{z_h \in X_h^{cg}\setminus\{0\}} \frac{b_h(v_h, z_h)}{\|z_h\|_{L^{p'}(\Omega)}} \right) \|v_h\|_{W_h^{2,p}(\Omega)}^{p-1}. \end{aligned}$$

Dividing by $\|v_h\|_{W_h^{2,p}(\Omega)}$, we obtain (4.22). □

Theorem 4.16 (existence and error estimates). Suppose that the hypotheses of Theorem 4.15 are satisfied. Then there exists a unique solution $u_h \in X_h^{cg}$ to (4.19). If the solution to (4.1) satisfies $u \in W^{s,p}(\Omega)$ for

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

$2 \le s \le k+1$, then

$$\|u - u_h\|_{W_h^{2,p}(\Omega)} \le Ch^{s-2}\|u\|_{W^{s,p}(\Omega)}.$$

Proof. The existence and uniqueness of a solution to (4.19) follows from the stability estimate (4.22). The error estimate follows from the stability and continuity of the bilinear form $b_h(\cdot,\cdot)$, the consistency of the scheme, and approximation properties of X_h^{cg} with respect to the discrete $W^{2,p}$-norm. □

Remark 4.17 (extensions). The ideas and analysis presented in this section has been extended to discontinuous Galerkin methods (Feng, Neilan and Schnake 2016) and mixed finite element methods (Lakkis and Pryer 2011, Neilan 2017).

4.3. Finite element method based on integro-differential approximation

In this section we consider a two-scale finite element discretization for problem (4.1) developed in Nochetto and Zhang (2017*a*), which is based on a regularized, integro-differential approximation proposed in Caffarelli and Silvestre (2010*b*). As in the previous sections we assume that the PDE operator $\mathcal{L}$ is uniformly elliptic, that is, there exist strictly positive constants λ, Λ satisfying $\lambda I \le A(x) \le \Lambda(x) I$, for all $x \in \bar{\Omega}$. We further make the simplifying assumption that $A \in C(\bar{\Omega}, \mathbb{S}^d)$, and make remarks when this regularity can be relaxed.

To explain and motivate the method, we first perform some algebraic manipulations, and rewrite the PDE as

$$A : D^2 u = \frac{\lambda}{2}\Delta u + A_\lambda^2 : D^2 u, \quad A_\lambda := \left(A - \frac{\lambda}{2} I\right)^{1/2}. \tag{4.24}$$

Let φ be a radially symmetric function with compact support in the unit ball satisfying $\int_{\mathbb{R}^d} |z|^2 \varphi(z) = d$. We then find that

$$\int_{\mathbb{R}^d} z_i z_j \varphi(z) = 0 \quad \text{for } i \ne j, \quad \text{and} \quad \int_{\mathbb{R}^d} z_i^2 \varphi(z) = 1.$$

Consequently, we have

$$\int_{\mathbb{R}^d} z \otimes z \varphi(z) = I,$$

and therefore

$$(A_\lambda(x))^2 : D^2 u(x) = A_\lambda(x)\left(\int_{\mathbb{R}^d} z \otimes z \varphi(z)\right) A_\lambda(x) : D^2 u(x).$$

For a regularization parameter $\epsilon > 0$, we make the change of variables

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

$y = \epsilon A_\lambda(x) z$ in the integral to obtain

$$(A_\lambda(x))^2 : D^2 u(x) = \int_{\mathbb{R}^d} \frac{(y \otimes y) : D^2 u(x)}{\epsilon^{d+2} \det(A_\lambda(x))} \varphi\left(\frac{A_\lambda^{-1}(x) y}{\epsilon}\right).$$

Set

$$Q = \left(\Lambda - \frac{\lambda}{2}\right)^{1/2}, \quad \Omega_\epsilon = \{x \in \Omega : \operatorname{dist}(x, \partial\Omega) > Q\epsilon\}, \quad \omega_\epsilon = \Omega \backslash \Omega_\epsilon, \tag{4.25}$$

and note that $\varphi(A_\lambda^{-1}(x) y/\epsilon)$ has support in the ball $B_{Q\epsilon}(0)$. For $x \in \Omega$, let $\theta = \theta(x) \in (0, 1]$ be the largest number such that $x \pm \theta y \in \Omega$ for all $y \in B_{Q\epsilon}(0)$. Recall that the second difference operator is given by

$$\delta^2_{y,\theta} u(x) := \frac{u(x + \theta y) - 2u(x) + u(x - \theta y)}{\theta^2},$$

and note that $\delta^2_{\theta y, \theta} u(x) = (y \otimes y) : D^2 u(x)$ if u is a quadratic polynomial, and that $\theta = 1$ for $x \in \Omega_\epsilon$.

Combining these calculations and identities, we are led to the approximation

$$(A_\lambda(x))^2 : D^2 u(x) \approx \int_{\mathbb{R}^d} \frac{|y|^2 \delta^2_{\theta y, \theta} u(x)}{\epsilon^{d+2} \det(A_\lambda(x))} \varphi\left(\frac{A_\lambda^{-1}(x) y}{\epsilon}\right) =: I_\epsilon u(x). \tag{4.26}$$

The approximation is quantified in the next lemma (Nochetto and Zhang 2017*a*, Lemma 2.1).

Lemma 4.18 (rate of convergence of integral transform). Let I_ϵ be the integral operator defined by (4.26), and let $U_{Q\epsilon}(x) := \bar{B}_{Q\epsilon}(x) \cap \bar{\Omega}$.

- If $u \in C^2(\bar{\Omega})$, then

$$I_\epsilon u(x) \to \left(A(x) - \frac{\lambda}{2} I\right) : D^2 u(x) \quad \text{as } \epsilon \to 0^+,$$

  for all $x \in \Omega$.

- If $u \in C^{2,\alpha}(U_{Q\epsilon}(x))$ for some $\alpha \in (0, 1]$, then

$$\left| I_\epsilon u(x) - \left(A(x) - \frac{\lambda}{2} I\right) : D^2 u(x) \right| \le C \|u\|_{C^{2,\alpha}(U_{Q\epsilon})} \theta^\alpha \epsilon^\alpha,$$

  for all $x \in \Omega$.

We approximate equation (4.1) by the integro-differential equation

$$\mathcal{L}^\epsilon u^\epsilon := \frac{\lambda}{2} \Delta u^\epsilon + I_\epsilon u^\epsilon = f \ \text{ in } \Omega. \tag{4.27}$$

We refer the reader to Caffarelli and Silvestre (2010*b*) for details about the existence, uniqueness and regularity estimates of solution u^ϵ.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

We now describe a convergent finite element scheme for the non-divergence-form problem (4.1) based on the regularized problem (4.27). To this end, we let $X^l_{0,h} = X^l_h \cap H^1_0(\Omega)$ be the linear, Lagrange finite element space with vanishing trace. Let $\phi_i \in X^l_{0,h}$ denote the normalized hat function with respect to the interior node $z_i \in \Omega^I_h$, and let Δ_h be the finite element Laplacian defined by (3.39). We consider the following finite element method: find $u_h \in X^l_{0,h}$ such that

$$\mathcal{L}^\epsilon_h u^\epsilon_h(z_i) := \frac{\lambda}{2}\Delta_h u^\epsilon_h(z_i) + I_\epsilon u^\epsilon_h(z_i) = f_i := \int_\Omega f\phi_i, \quad \text{for all } z_i \in \Omega^I_h. \tag{4.28}$$

Note that the formulation (4.28) is *not* obtained by testing (4.27) with ϕ_i (which would introduce the term $\int_\Omega I_\epsilon u^\epsilon_h \phi_i$). Instead, mass lumping is used to preserve the monotonicity of the scheme.

Lemma 4.19 (monotonicity). Suppose that $v_h, w_h \in X^l_{0,h}$ satisfy $v_h \le w_h$ with equality at $z \in \Omega^I_h$. Then $I_\epsilon v_h(z) \le I_\epsilon w_h(z)$. Consequently, if $\mathcal{T}_h$ satisfies (3.42), then $\mathcal{L}^\epsilon_h$ is monotone.

Proof. From the hypotheses and the definition of $\delta^2_{\theta y,\theta}$, we have $\delta^2_{\theta y,\theta} v(z) \le \delta^2_{\theta y,\theta} w(z)$, and therefore $I_\epsilon v_h(z_i) \le I_\epsilon w_h(z_i)$. The monotonicity of $\mathcal{L}^\epsilon_h$ then follows from Lemma 3.42. □

The monotonicity, along with the Alexandrov–Bakelman–Pucci estimate for the finite element Laplacian, yields the following maximum principle.

Theorem 4.20 (discrete ABP estimate for $\mathcal{L}^\epsilon_h$). Suppose that $\mathcal{T}_h$ satisfies (3.42). Then, for $v_h \in X^l_{0,h}$ satisfying

$$\mathcal{L}^\epsilon_h v_h(z_i) \le f_i, \quad \text{for all } z_i \in \Omega^I_h,$$

we obtain

$$\sup_\Omega v^-_h \le \frac{C}{\lambda}\Bigg(\sum_{z_i \in \mathcal{C}^-_h(v_h)} |f^+_i|^d |\omega_{z_i}|\Bigg)^{1/d},$$

where $\mathcal{C}^-_h(v_h)$ is the nodal contact set given in Definition 3.23.

Proof. Let $\Gamma_h(v_h) = \Gamma(v_h)$ be the convex envelope of v_h. Then, for a contact point $z_i \in \mathcal{C}^-_h(v_h)$,

$$0 \le I_\epsilon \Gamma(v_h)(z_i) \le I_\epsilon v_h(z_i),$$

where the first inequality follows from the convexity of $\Gamma(v_h)$ and the second follows from the monotonicity of I_ϵ in Lemma 4.19. Consequently,

$$\frac{\lambda}{2}\Delta_h v_h(z_i) \le \mathcal{L}^\epsilon_h v_h(z_i) \le f^+_i$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

since $f_i \geq 0$ for $z_i \in \mathcal{C}_h^-(v_h)$. The result now follows from this inequality and Theorem 3.50 (see Remark 3.51). □

Since the method (4.28) is linear, a corollary of the ABP estimate is the existence and uniqueness of a solution u_h.

Corollary 4.21 (existence and uniqueness). Suppose that $\mathcal{T}_h$ satisfies (3.42). Then there exists a unique $u_h \in X_{0,h}^l$ satisfying (4.28).

We now turn our attention to error estimates of the finite element method (4.28) and derive a rate of convergence in the L^∞-norm. To do so we assume that the solution to (4.1) satisfies $u \in C^{2,\alpha}(\Omega)$. Recall (see Theorem 2.17) that this regularity is guaranteed provided that A is Hölder-continuous and $\partial\Omega$ is sufficiently smooth.

Now, since the method is linear and the problem is stable, such estimates reduce to the consistency of the method. However, as shown in Lemma 3.38, the finite element method is not consistent, in the sense of Definition 3.1, when supplemented with the canonical interpolant I_h^{fe}. Instead, we make use of the elliptic projection I_h^{ep} defined in (3.41) and its properties (see Lemma 3.41 and Proposition 3.40).

In conclusion, in order to derive error estimates, it suffices to derive upper bounds for the difference $u_h^\epsilon - I_h^{ep}u$. To this end, we apply the definition of the method (4.28) to obtain the *error equation*

$$\mathcal{L}_h^\epsilon[I_h^{ep}u - u_h^\epsilon](z_i) = \int_{\omega_{z_i}} (T_1^{(i)} + T_2^{(i)} + T_3^{(i)})\phi_i \tag{4.29}$$

with

$$\begin{aligned} T_1^{(i)} &= I_\epsilon[I_h^{ep}u](z_i) - I_\epsilon u(z_i), \\ T_2^{(i)} &= I_\epsilon u(z_i) - \left(A(z_i) - \frac{\lambda}{2}I\right) : D^2u(z_i), \\ T_3^{(i)} &= \left(A(z_i) - \frac{\lambda}{2}I\right) : (D^2u(z_i) - D^2u(x)). \end{aligned}$$

Note that, with the finite element ABP estimate given in Theorem 4.20 and the approximation results of the elliptic projection stated in Proposition 3.40, upper-bound estimates of $T_j^{(i)}$ yield error estimates of $u - u_h$.

With the assumed regularity $u \in C^{2,\alpha}(\Omega)$, we immediately find that $T_3^{(i)}$ can be bounded by

$$|T_3^{(i)}| \leq Ch^\alpha \|u\|_{C^{2,\alpha}(\Omega)}. \tag{4.30}$$

For $T_2^{(i)}$, we apply Lemma 4.18 to obtain

$$|T_2^{(i)}| \leq C\theta^\alpha \epsilon^\alpha \|u\|_{C^{2,\alpha}(\Omega)} \leq C\epsilon^\alpha \|u\|_{C^{2,\alpha}(\Omega)}. \tag{4.31}$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Finally, we apply the approximation results of Proposition 3.40 and the definition of the second-order difference operator $\delta^2_{\theta y,\theta}$ to obtain

$$|\delta^2_{\theta y,\theta}(I_h^{ep}u(z_i) - u(z_i))| \le C\frac{h^2}{\theta^2}\,|\log h|\|u\|_{W^{2,\infty}(\Omega)}.$$

This leads to the estimate

$$|T_1^{(i)}| \le C\|u\|_{W^{2,\infty}(\Omega)}\left(\frac{h^2}{\epsilon^2}\,|\log h| + \frac{h^2}{\theta^2\epsilon^2}\,|\log h|\chi_{\omega_\epsilon}(z_i)\right), \tag{4.32}$$

where χ_{ω_ϵ} is the indicator function of ω_ϵ, which is defined in (4.25). Combining (4.30)–(4.32), we obtain

$$\begin{aligned}&|T_1^{(i)} + T_2^{(i)} + T_3^{(i)}| \\ &\quad\le C\left(h^\alpha + \epsilon^\alpha + \frac{h^2}{\epsilon^2}\,|\log h| + \frac{h^2}{\theta^2\epsilon^2}\,|\log h|\,|\chi_{\omega_\epsilon}(z_i)\right),\end{aligned} \tag{4.33}$$

where we have absorbed the factor $\|u\|_{C^{2,\alpha}(\Omega)}$ into the constant C.

Note that, owing to the last term on the right-hand side, estimate (4.33) reduces to order 1 in the boundary layer ω_ϵ. In order to derive meaningful estimates in this region, we introduce the barrier layer function

$$b(x) := \xi(\operatorname{dist}(x,\partial\Omega)),\ \text{with } \xi(s) := \begin{cases} Q^{-2}(s - Q\epsilon)^2 - \epsilon^2 & \text{if } s \le Q\epsilon, \\ -\epsilon^2 & \text{if } s > Q_\epsilon, \end{cases} \tag{4.34}$$

The discrete boundary layer function is defined as $b_h := I_h^{fe}b$. The next result summarizes key properties of b_h; see Nochetto and Zhang (2017*a*, Lemma 6.1)

Lemma 4.22 (properties of barrier layer function). Let $b_h = I_h^{fe}b$, where b is given by (4.34). Then

$$\mathcal{L}_h^\epsilon b_h(z_i) \ge C\chi_{\omega_\epsilon(z_i)}, \quad |b_h(z_i)| \le C\epsilon^2.$$

Note that since every $z_i \in \omega_\epsilon$ is at most $O(h)$ from $\partial\Omega$, we have $\epsilon\theta \ge Ch$ on ω_ϵ. Consequently, the last term in (4.33) can be bounded by the discrete barrier function as follows:

$$\frac{h^2}{\theta^2\epsilon^2}\,|\log h|\chi_{\omega_\epsilon(z_i)}\|u\|_{C^{2,\alpha}(\Omega)} \le C\,|\log h|\chi_{\omega_\epsilon}(z_i) \le C\,|\log h|\mathcal{L}_h^\epsilon b_h(z_i).$$

Thus, combining this estimate with (4.33) and (4.29) leads to

$$\mathcal{L}_h^\epsilon\left[I_h^{ep}u - u_h^\epsilon - C\,|\log h|b_h\right](z_i) \le C\left(h^\alpha + \epsilon^\alpha + \frac{h^2}{\epsilon^2}\,|\log h|\right). \tag{4.35}$$

From this expression, we easily obtain estimates of $u - u_h$.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Theorem 4.23 (rate of convergence). Suppose that $\mathcal{T}_h$ satisfies (3.42), and that the solution to (4.1) satisfies $u \in C^{2,\alpha}(\Omega)$. Let $u_h \in X^l_{0,h}$ be the solution to (4.28) with $\epsilon = C(h^2 |\log h|)^{1/(2+\alpha)}$. Then

$$\|u - u_h\|_{L^\infty(\Omega)} \leq C(h^2 |\log h|)^{\alpha/(2+\alpha)}. \tag{4.36}$$

Proof. Applying the finite element ABP for $\mathcal{L}^\epsilon_h$ (see Theorem 4.20) to (4.35) yields

$$\sup_\Omega (I^{ep}_h u - u^\epsilon_h - C |\log h| b_h)^- \leq C\left(h^\alpha + \epsilon^\alpha + \frac{h^2}{\epsilon^2} |\log h|\right).$$

Therefore, since $|b_h| \leq \epsilon^2$, we get

$$\sup (I^{ep}_h u - u^\epsilon_h)^- \leq C\left(h^\alpha + \epsilon^\alpha + \left(\epsilon^2 + \frac{h^2}{\epsilon^2}\right) |\log h|\right).$$

Similar estimates are obtained for $\sup_\Omega (I^{ep}_h u - u^\epsilon_h)^+$, thus leading to

$$\|I^{ep}_h u - u^\epsilon_h\|_{L^\infty(\Omega)} \leq C\left(h^\alpha + \epsilon^\alpha + \left(\epsilon^2 + \frac{h^2}{\epsilon^2}\right) |\log h|\right).$$

If $\epsilon = C(h^2 |\log h|)^{1/(2+\alpha)}$, then

$$(h^2/\epsilon^2) |\log h| \leq C\epsilon^\alpha \leq C(h^2 |\log h|)^{\alpha/(2+\alpha)},$$

and therefore

$$\|I^{ep}_h u - u^\epsilon_h\|_{L^\infty(\Omega)} \leq C(h^2 |\log h|)^{\alpha/(2+\alpha)}.$$

Finally, applying Proposition 3.40 and the triangle inequality yields (4.36). The proof is complete. □

Theorem 4.23 shows that if $\alpha = 1$, then the error satisfies $\|u - u_h\|_{L^\infty(\Omega)} = O(h^{2/3-\tau})$ for arbitrary $\tau > 0$. If more regularity is assumed then an almost linear rate is obtained (Nochetto and Zhang 2017*a*, Corollary 6.8)

Theorem 4.24 (improved rate). Let h and ϵ satisfy $\epsilon = Ch^{2/(3+\alpha)}$. If the solution to (4.1) has the regularity $u \in C^{3,\alpha}(\Omega)$, and if $\mathcal{T}_h$ satisfies (3.42), then

$$\|u - u^\epsilon_h\|_{L^\infty(\Omega)} \leq Ch^{2(1+\alpha)/(3+\alpha)} |\log h|.$$

In the opposite direction, if we assume less regularity of the solution and data, then convergence is still obtained, although without rates (Nochetto and Zhang 2017*a*, Corollary 6.5).

Theorem 4.25 (convergence). Assume that $u \in C^2(\bar{\Omega})$, that the coefficient matrix satisfies $A \in VMO(\Omega, \mathbb{S}^d)$, and that the two scales ϵ and h satisfy $\epsilon = Ch |\log h|$. Let $u^\epsilon_h \in X^l_{0,h}$ satisfy (4.28) with $A(z_i)$ replaced by

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

its average

$$\bar{A}(z_i) := \frac{1}{|\omega_{z_i}|}\int_{\omega_{z_i}} A(y).$$

If the mesh condition (3.42) is satisfied, then $\lim_{h\to 0^+} \|u - u_h^\epsilon\|_{L^\infty(\Omega)} = 0$.

5. Discretizations of convex second-order elliptic equations

Up to this point we have discussed general issues regarding stability and convergence of numerical methods for general fully nonlinear equations, as well as the construction of suitable schemes for linear problems in non-divergence form. For the rest of this overview we will merge the ideas presented in the previous sections. Since convergence of these schemes has already been discussed in Theorems 3.7 and 3.8, we will pay special attention to obtaining rates of convergence for them. In this section we will focus on *convex* equations. Moreover, as explained Section 1.3, there is no loss of generality in assuming that we are dealing with the Hamilton–Jacobi–Bellman equation, which we recall reads

$$\begin{cases} F(x,u,Du,D^2u) := \sup_{\alpha\in\mathcal{A}}[\tilde{\mathcal{L}}^\alpha u(x) - f^\alpha(x)] = 0 & \text{in } \Omega, \\ u = g & \text{on } \partial\Omega, \end{cases} \tag{5.1}$$

where

$$\tilde{\mathcal{L}}^\alpha u(x) := A^\alpha(x) : D^2u(x) + \mathbf{b}(x)\cdot Du(x) + c(x)u(x).$$

We assume that the linear elliptic operators are such that, for every $\alpha \in \mathcal{A}$,

$$A^\alpha \geq \lambda I, \quad \text{for all } \alpha \in \mathcal{A},$$

so that the operator F is uniformly elliptic.

The numerical schemes for problem (5.1) can be roughly classified as finite difference, finite element and semi-Lagrangian methods. In this section we will focus on the first two. Semi-Lagrangian schemes will be illustrated in a particular case, the Monge–Ampère equation, in Section 6.2.

5.1. Finite difference methods

The analysis of the convergence properties of finite difference schemes for (5.1) dates back to Krylov (1997, 2000), who considered the problem for $\Omega = \mathbb{R}^d$ and constant 'coefficients', that is, when the operators $\tilde{\mathcal{L}}^\alpha$ are x-independent. These results were later extended by Barles and Jakobsen (2002) to Lipschitz coefficients. Before embarking on the technical details of their results, let us give some intuition. Recall that, in general, a finite difference scheme is written in the form (3.13). Ideally, to approximate

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

u, the solution of (5.1), one would first construct a smooth function u_ε, parametrized by ε which, for a constant C independent of ε, satisfies

$$\|u - u_\varepsilon\|_{L^\infty(\Omega)} \leq C\varepsilon^{\kappa_1}.$$

A strengthened notion of consistency (see Definition 3.1) would then imply that $F_h[I_h^{fd}u_\varepsilon] \approx h^{\kappa_2}$, where the hidden constants in this expression may depend on ε. By stability, Definition 3.3, and monotonicity, Definition 3.5, we obtain

$$\|I_h^{fd}u_\varepsilon - u_h\|_{L^\infty(\bar{\Omega}_h)} \leq Ch^{\kappa_2}.$$

An application of the triangle inequality and relating the smoothing parameter ε to the discretization h would yield a rate of convergence.

Unfortunately, the construction of such a smooth approximation u_ε is not immediate in practice. The groundbreaking idea of Krylov was to 'shake the coefficients'. He introduced u_ε as the solution of

$$\inf_{|e|\leq\varepsilon} F(x+e, u_\varepsilon, Du_\varepsilon, D^2u_\varepsilon) = 0, \tag{5.2}$$

and, from u_ε, he was able to construct a smooth subsolution to (5.1) so that by comparison we can obtain an upper bound for $u_h - u$. To obtain a lower bound, the original work of Krylov invoked arguments that some authors have characterized as probabilistic. However, Krylov (2015, page 3) disagrees with this statement. Another line of reasoning was given by Barles and Jakobsen (2002) who, instead, proposed that the problem and scheme should play a symmetric role. In other words, they introduce u_h^ε which solves

$$\inf_{|e|\leq\varepsilon} F_h[u_h^\varepsilon](z+e) = 0. \tag{5.3}$$

Under suitable assumptions they show that this family of operators possesses unique smooth solutions, where the smoothness is independent of ε. This then allows us to compare u and u_h^ε and obtain a rate of convergence. The assumptions that they rely on, however, must be checked for every particular instance.

Let us now proceed with the details. Recall that we are operating in the whole space $\mathbb{R}^d$ and that the mesh is given by $\bar{\Omega}_h = \mathbb{Z}_h^d$. We assume that all the coefficients of $\tilde{\mathcal{L}}^\alpha$ and f^α are Lipschitz-continuous uniformly in α and that $\sup_{\alpha\in\mathcal{A}} c^\alpha(x) \leq c < 0$ for every $x \in \mathbb{R}^d$. This, as indicated in Theorem 2.52(b), implies that the operator has a comparison principle. The structural requirements on the discretization scheme are summarized below.

Assumption 5.1 (finite differences). The finite difference scheme

$$F_h = F_h(z, r, q)$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

satisfies the following.

(i) The scheme is consistent in a sense stronger than Definition 3.1, namely, for some $\kappa > 0$,

$$\|F_h[I_h^{fd}\phi] - F[\phi]\|_{L^\infty(\bar{\Omega}_h)} \leq Ch^\kappa,$$

for all sufficiently smooth ϕ.

(ii) The scheme is of non-negative type: see Definition 3.9. In addition, there exists $\bar{c} > 0$ such that $F_h(z, r + t, q + \mathbf{1}t) \leq F_h(z, r, q) - \bar{c}t$ for all $t \geq 0$.

(iii) The scheme is convex in the r and q variables.

(iv) The scheme is uniformly solvable and smooth under perturbations. In other words, for $h > 0$ small enough and $\varepsilon \in [0, 1]$, problem (5.3) has a unique solution u_h^ε and, moreover, for some $\delta > 0$ we have

$$\|u_h - u_h^\varepsilon\|_{L^\infty(\bar{\Omega}_h)} \leq C\varepsilon^\delta.$$

Note that conditions (i)–(iii) of Assumption 5.1 are relatively easy to enforce. However, condition (iv) must be verified in each case. As a first step, we show that schemes satisfying the non-negativity condition satisfy a comparison principle.

Lemma 5.2 (discrete comparison principle). Let F_h satisfy Assumption 5.1(ii). Let $v_h, w_h \in X_h^{fd}$ be two bounded nodal functions that are sub- and supersolutions to the discrete problem $F_h[u_h] = 0$, respectively. Then $v_h \leq w_h$.

Proof. We assume $r = \sup_{\bar{\Omega}_h}(v_h - w_h) > 0$ and derive a contradiction. For simplicity we assume there exists $z \in \bar{\Omega}_h$ such that $v_h(z) - w_h(z) = r$, for otherwise we can apply a standard limiting argument.

As in Section 3.2 we write $F_h[v_h](z) = F_h(z, v_h(z), Tv_h(z))$, where the set of translates of $v_h(z)$ is $Tv_h(z) = \{v_h(z + hy) : y \in S\}$, and S is the stencil. Note that $v_h(z) = w_h(z) + r$ and $v_h(z + hy) \leq w_h(z + hy) + r$ for $y \in S$. Since F_h is increasing in its third argument (see Definition 3.9), we have

$$\begin{aligned} 0 &\leq F_h[v_h](z) - F_h[w_h](z) \\ &= F_h(z, w_h(z) + r, Tv_h(z)) - F_h(z, w_h(z), Tw_h(z)) \\ &\leq F_h(z, w_h(z) + r, Tw_h(z) + \mathbf{1}r) - F_h(z, w_h(z), Tw_h(z)). \end{aligned}$$

Applying Assumption 5.1(ii) then yields

$$0 \leq (F_h(z, w_h(z), Tw_h(z)) - \bar{c}r) - F_h(z, w_h(z), Tw_h(z)) = -\bar{c}r,$$

which is a contradiction. □

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Having shown the comparison principle, we now state the convergence rate of the scheme.

Theorem 5.3 (rate of convergence). Assume that the finite difference scheme satisfies Assumption 5.1. Then there are constants $C_1, C_2 > 0$ such that, for every $z \in \bar{\Omega}_h$,

$$I_h^{fd} u(z) - u_h(z) \leq C_1 h^{\kappa_1}, \quad u_h(z) - I_h^{fd} u(z) \leq C_2 h^{\kappa_2},$$

where the rates $\kappa_1, \kappa_2 > 0$ are not necessarily the same.

Proof. Let us sketch the proof of each of these bounds.

Proof of $I_h^{fd} u(z) - u_h(z) \leq C_1 h^{\kappa_1}$. Let u_ε be the solution of (5.2). By definition, after the change of variables $y = x + e$, we realize that $u_\varepsilon(\cdot - e)$ is, for $|e| \leq \varepsilon$, a subsolution to the equation, that is, it satisfies, in the viscosity sense,

$$F\big(y, u_\varepsilon(\cdot - e), Du_\varepsilon(\cdot - e), D^2 u_\varepsilon(\cdot - e)\big) \geq 0.$$

We regularize this function by the mollification $u^\varepsilon = u_\varepsilon \star \rho_\varepsilon$, where ρ_ε are the standard mollifiers. From the convexity of the operator F and Jensen's inequality, it follows that u^ε is also a subsolution. Since u^ε is now a smooth function, we can invoke consistency to obtain

$$F_h[I_h^{fd} u^\varepsilon](z) \geq F(z, u^\varepsilon(z), Du^\varepsilon(z), D^2 u^\varepsilon(z)) - Ch^\kappa \geq -Ch^\kappa,$$

for some constant C that depends on the smoothness of u^ε which, in turn, scales like negative powers of ε, say $C \leq \varepsilon^{-\delta_1}$. In conclusion, we have obtained that

$$F_h[I_h^{fd} u^\varepsilon](z) \geq -Ch^\kappa \varepsilon^{-\delta_1}.$$

Assumption 5.1(ii) shows that the function $I_h^{fd}(u^\varepsilon - Ch^\kappa \varepsilon^{-\delta_1})$ is, for a suitably chosen C, a subsolution of the scheme, that is, $F_h[I_h^{fd}(u^\varepsilon - Ch^\kappa \varepsilon^{-\delta_1})] \geq 0$. Therefore, by the comparison principle given in Lemma 5.2,

$$I_h^{fd} u^\varepsilon(z) - u_h(z) \leq Ch^\kappa \varepsilon^{-\delta_1}.$$

In conclusion, using the continuity properties of equation (5.1) and properties of mollifiers, we obtain

$$\begin{aligned} I_h^{fd} u(z) - u_h(z) &= I_h^{fd}(u - u_\varepsilon)(z) + I_h^{fd}(u_\varepsilon - u^\varepsilon)(z) + I_h^{fd}(u^\varepsilon - u_h)(z) \\ &\leq \frac{C_1}{2}(\varepsilon^{\delta_2} + h^\kappa \varepsilon^{-\delta_1}), \end{aligned}$$

where $\delta_2 > 0$ depends on the smoothness of u (see Theorem 2.67). Optimizing with respect to ε, we get the result.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Proof of $u_h(z) - I_h^{fd}u(z) \le C_2 h^{\kappa_2}$. We follow similar reasoning but this time, we interchange the roles that the equation and the scheme have played in the previous step. Indeed, by Assumption 5.1(iv) we know that there is a $u_h^\varepsilon \in X_h^{fd}$ that solves (5.3) and, again with the change of variables $\tilde{z} = z + e$, this function satisfies

$$F_h[u_h^\varepsilon](z) \ge 0,$$

so that it is a subsolution of the scheme. Now, convexity of F_h implies that $I_h^{fd}(u_h^\varepsilon \star \rho_\varepsilon)$ is also a subsolution of the scheme. Moreover, we have that $u_h^\varepsilon \star \rho_\varepsilon$ is a smooth function, and therefore consistency implies that for some $\delta_3 > 0$ we have

$$F[u_h^\varepsilon \star \rho_\varepsilon](z) \ge -Ch^\kappa \varepsilon^{-\delta_3}.$$

Monotonicity of the equation shows that, for a suitably chosen C, the function $u_h^\varepsilon \star \rho_\varepsilon - Ch^\kappa \varepsilon^{-\delta_3}$ is a subsolution which, by comparison, readily implies that

$$u_h^\varepsilon \star \rho_\varepsilon(z) - u(z) \le Ch^\kappa \varepsilon^{-\delta_3}.$$

Properties of convolutions together with Assumption 5.1(iv) then imply

$$\begin{aligned} u_h(z) - I_h^{fd}u(z) &= (u_h - u_h^\varepsilon)(z) + (u_h^\varepsilon - I_h^{fd}(u_h^\varepsilon \star \rho_\varepsilon))(z) \\ &\quad + I_h^{fd}(u_h^\varepsilon \star \rho_\varepsilon - u)(z) \le \frac{C_2}{3}(\varepsilon^\delta + \varepsilon^{\delta_2} + h^\kappa \varepsilon^{-\delta_3}). \end{aligned}$$

Optimization in ε once again yields the result. □

Let us now give an example of finite difference schemes for which the assumptions of Theorem 5.3 can be verified.

Example 5.4 (monotone finite differences). Let A^α be independent of x for every α and, possibly after a renormalization, verify

$$\sum_{i=1}^d \left[a_{i,i}^\alpha - \sum_{j \ne i} |a_{i,j}^\alpha| \right] \le 1.$$

More importantly, we assume that these matrices are diagonally dominant, that is, (3.31) holds. For simplicity, assume also that $\mathbf{b}^\alpha \equiv 0$ for all $\alpha \in \mathcal{A}$. As shown in Lemma 3.34, there exists a monotone finite difference $\mathcal{L}_h^\alpha + c^\alpha$ that is consistent with $\tilde{\mathcal{L}}^\alpha$. We then define

$$F_h[u_h](z) = \sup_{\alpha \in \mathcal{A}} \left[\mathcal{L}_h^\alpha u_h(z) + c^\alpha(z) u_h(z) - f^\alpha(z) \right] = 0.$$

In this case, Barles and Jakobsen (2002, Section 4) show that all the assumptions are verified and, moreover, that $\kappa_1 = \kappa_2 = 1/3$.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Remark 5.5 (examples and improvements). Let us comment on the various improvements and refinements of Theorem 5.3 as well as some extensions of Example 5.4.

- In Example 5.4 it is assumed that the leading coefficients A^α do not depend on the spatial variable x. Barles and Jakobsen (2005) studied the relation between (5.1) and a certain system of quasi-variational inequalities of compliance obstacle type (see (5.4) below). They used this system instead of Assumption 5.1(iv) to obtain an upper bound for $u_h - I_h^{fd} u$ for the case when A^α is Lipschitz-continuous in the space variable. Their results show that $\|I_h^{fd} u - u_h\|_{L^\infty(\bar{\Omega}_h)} \leq Ch^{1/5}$.
- Let $\{\mathbf{e}_i\}_{i=1}^d$ be an orthonormal basis of $\mathbb{R}^d$ and assume that the matrices have the form
$$A^\alpha(x) = \sum_{i=1}^d a_i^\alpha(x) \mathbf{e}_i \otimes \mathbf{e}_i$$
for some a_i^α that is Lipschitz in x uniformly in α. Recall that, with this assumption, Lemma 3.33 guarantees the existence of a monotone finite difference scheme. In this setting Krylov (2005) shows that $\kappa_1 = \kappa_2 = 1/2$. Moreover, in the same setting but assuming that the coefficients are $C^{1,1}$, Krylov (2014) shows that $\kappa_1 = \kappa_2 = 2/3$.
- All the aforementioned results consider the case $\Omega = \mathbb{R}^d$, so that boundary conditions are not an issue: see Theorem 3.8. Dong and Krylov (2007) consider the boundary value problem (5.1) under the assumption that the boundary conditions are attained classically; see Definitions 2.48 and 2.53. Under the assumption that a *barrier* function can be constructed, that is, a smooth b such that $b > 0$ in Ω, $b = 0$ on $\partial\Omega$ and
$$F[b](x) \leq -1,$$
they were able to extend the results presented here and show that the rate of convergence is $\|I_h^{fd} u - u_h\|_{L^\infty(\bar{\Omega}_h)} \leq Ch^{1/2}$.

Other results, which invoke the probabilistic interpretation of (5.1), can be found in the literature; see, for instance, Fleming and Soner (2006) and Menaldi (1989). The reader is also referred to the introduction of Krylov (2015) for a detailed account of the development of error estimates for (5.1).

5.2. Finite element methods

We now focus on the construction and analysis of finite element schemes for (5.1). We will divide the exposition in two cases. First we will discuss the discretization for a variant of the problem when the operators $\tilde{\mathcal{L}}^\alpha$ are replaced by operators in divergence form $\tilde{L}^\alpha$ with smooth coefficients. Then

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

we will discuss the case of (5.1) where the coefficients for $\tilde{\mathcal{L}}^\alpha$ satisfy the Cordes condition of Definition 2.32, which is based on the discretization of non-divergence-form operators of Section 4.1. We must remark that it is possible to construct discrete schemes using the integro-differential approximation of Section 4.3. However, to avoid repetition, its discussion will be illustrated in Section 7.3 for a non-convex operator of Isaacs type (see Example 2.13). Setting $\#\mathcal{B} = 1$ there, we reduce the scheme and its analysis to the case we are concerned with here.

5.2.1. Discretization for divergence-form operators

Let us consider (5.1) but where the operators $\tilde{\mathcal{L}}^\alpha$ are replaced by divergence-form elliptic operators $\tilde{L}^\alpha$ with $C^2(\bar{\Omega})$ coefficients; see Definition 2.3. In addition, we assume that $\partial\Omega$ is sufficiently smooth, $g = 0$ and that, for every $\alpha \in \mathcal{A}$, we have $0 \le f^\alpha \in L^\infty(\Omega)$. Finally, we assume that $\mathcal{A} = \{1, \dots, M\}$ for some $M \in \mathbb{N}$.

Remark 5.6 (smooth coefficients). Note that, if the coefficients of $\tilde{\mathcal{L}}^\alpha$ in (5.1) are sufficiently smooth, one can rewrite this operator in divergence form. Therefore, this reformulation is sufficiently general.

Recall that, by integration by parts, to every operator $\tilde{L}^\alpha$ we can associate the bilinear form $a^\alpha : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$, defined by

$$a^\alpha(v, w) = \int_\Omega (A^\alpha Dv \cdot Dw + v^\alpha \mathbf{b}^\alpha \cdot Dw + \mathbf{c}^\alpha \cdot Dvw + d^\alpha vw).$$

To simplify the discussion, we will assume that these forms are coercive uniformly in α, that is, there is a constant λ_0 such that

$$\inf_{\alpha\in\mathcal{A}} a^\alpha(v, v) \ge \lambda_0 \|Dv\|^2_{L^2(\Omega)}, \quad \text{for all } v \in H^1_0(\Omega).$$

Given $k > 0$ and $w \in H^1_0(\Omega)$, define

$$\mathcal{K}(k, w) = \{v \in H^1_0(\Omega) : v \le k + w\},$$

which is a closed and convex subset of $H^1_0(\Omega)$. For $k > 0$ we introduce a system of quasi-variational inequalities of compliance obstacle type as follows: find $u^\alpha_k \in \mathcal{K}(k, u^{\alpha+1})$ such that

$$a^\alpha(u^\alpha_k, u^\alpha_k - v) \le (f^\alpha, u^\alpha_k - v), \quad \text{for all } v \in \mathcal{K}(k, u^{\alpha+1}), \tag{5.4}$$

with $u^{M+1}_k = u^1_k$. Standard results on quasi-variational inequalities, like those presented in Bensoussan and Lions (1984, Chapter 4), imply the existence and uniqueness of $\mathbf{u}_k := \{u^\alpha_k\}_{\alpha\in\mathcal{A}} \subset H^1_0(\Omega)$ that solves (5.4). In addition, adaptions of the results of Section 2.3 to the case of (quasi-) variational inequalities (*i.e.* by penalization and a limiting argument) allow us to conclude that $\mathbf{u}_k \subset W^{1,\infty}(\Omega) \cap W^{2,p}_{loc}(\Omega)$ $(p < \infty)$. The purpose of this

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

system lies in the fact that, as $k \to 0$, the solutions to (5.4) converge to the solution to (5.1). For a proof of the following result see Evans and Friedman (1979, Theorem 7.2) and Bensoussan and Lions (1984, Section 4.6).

Theorem 5.7 (convergence as $k \to 0$). In the setting described above there exists a unique strong solution u to (5.1). Moreover, defining $\mathbf{u} = \{u\}_{\alpha \in \mathcal{A}}$, we have

$$\lim_{k \to 0} \|\mathbf{u} - \mathbf{u}_k\|_{\ell^\infty(\mathcal{A}, L^\infty(\Omega))} = 0.$$

It is remarkable that this result was shown before the development of the viscosity solutions of Section 2, and it was originally used to show the well-posedness of the Hamilton–Jacobi–Bellman equations. In addition, it can be used to propose finite element discretizations of (5.1) by instead discretizing (5.4). Then, if one is able to extend the standard $L^\infty(\Omega)$-norm estimates for variational inequalities (see Nochetto, Otárola and Salgado 2015, Theorem 2.9) and give a rate for the limit in Theorem 5.7, we obtain a convergent (with rates) finite element method. This programme has been carried out by Cortey-Dumont (1987), Boulbrachene and Haiour (2001) and Boulbrachene and Cortey-Dumont (2009), with some success.

With the notation of Section 3.5 we begin by defining, for $k > 0$ and $w_h \in X^l_{0,h}$, the set

$$\mathcal{K}_h(k, w_h) = \{v_h \in X^l_{0,h} : v_h \leq k + w_h\},$$

which is a closed and convex subset of $X^l_{0,h}$. Moreover, we remark that it is sufficient to impose the inequality at the nodes $z \in \Omega^I_h$. We approximate the solution to (5.4) by the following set of discrete quasi-variational inequalities: find $u^{\alpha,k}_h \in \mathcal{K}_h(k, u^{\alpha+1,k}_h)$ such that

$$a^\alpha(u^{\alpha,k}_h, u^{\alpha,k}_h - v_h) \leq (f^\alpha, u^{\alpha,k}_h - v_h), \quad \text{for all } v_h \in \mathcal{K}_h(k, u^{\alpha+1,k}_h), \tag{5.5}$$

with $u^{M+1,k}_h = u^{1,k}_h$. Once again, it can be shown that this problem always has a unique solution $\mathbf{u}_{h,k} = \{u^{\alpha,k}_h\}_{\alpha \in \mathcal{A}} \subset X^l_{0,h}$.

To establish the $L^\infty(\Omega)$-norm convergence of the solutions to (5.5) to the solution to (5.4), we must assume that the ensuing stiffness matrices are M-matrices. Examining the proof of Lemma 3.42, we realize that for this to hold, it is sufficient to require that

$$a^\alpha(\phi_i, \phi_j) \leq 0, \quad \text{for all } \alpha \in \mathcal{A},\ i \neq j, \tag{5.6}$$

which we assume below.

Remark 5.8 (lack of generality). The discussion of Section 3.5 shows that condition (5.6) is satisfied if the mesh is weakly acute *in the metric induced by* A^α *for all* $\alpha \in \mathcal{A}$. This is a severe restriction in practice, as it

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

is not clear how to impose such a condition for one matrix, let alone for a family of them.

Although a rate of convergence for the limit in Theorem 5.7 does not seem possible, there is a rate for the approximation of (5.4) by (5.5).

Theorem 5.9 (rate of convergence). Assume that, for all $h > 0$, the family of triangulations $\mathcal{T}_h$ satisfies condition (5.6). Then we have

$$\|\mathbf{u}_k - \mathbf{u}_{h,k}\|_{\ell^\infty(\mathcal{A},L^\infty(\Omega))} \leq Ch^2 \,|\log h|^3,$$

where the constant $C > 0$ depends on $M = \#\mathcal{A}$ and k.

Proof. We will follow the ideas of Boulbrachene and Haiour (2001) which, in turn, borrow from the iterative schemes used to prove existence of elliptic quasi-variational inequalities considered in Bensoussan and Lions (1984, Chapter 4).

Step 1. We introduce the following iterative scheme. Define $\hat{\mathbf{u}}^0 = \{\hat{u}^{\alpha,0}\}_{\alpha\in\mathcal{A}}$ as the solutions to the unconstrained problems, that is,

$$a^\alpha(\hat{u}^{\alpha,0}, v) = (f^\alpha, v), \quad \text{for all } v \in H_0^1(\Omega).$$

Assuming that, for $n \geq 0$, $\hat{\mathbf{u}}^n$ has been defined, we look for $\hat{u}^{\alpha,n+1} \in \mathcal{K}(k, \hat{u}^{\alpha+1,n})$ such that

$$a^\alpha(\hat{u}^{\alpha,n+1}, \hat{u}^{\alpha,n+1} - v) \leq (f^\alpha, \hat{u}^{\alpha,n+1} - v), \quad \text{for all } v \in \mathcal{K}(k, \hat{u}^{\alpha+1,n}).$$

Using the positivity and order-preserving properties of the associated map, it is possible then to show that for $\lambda < \min\{1, k/\|\hat{\mathbf{u}}^0\|_{\ell^\infty(\mathcal{A},L^\infty(\Omega))}\}$ we have

$$\|\mathbf{u}_k - \hat{\mathbf{u}}^n\|_{\ell^\infty(\mathcal{A},L^\infty(\Omega))} \leq (1-\lambda)^n \|\hat{\mathbf{u}}^0\|_{\ell^\infty(\mathcal{A},L^\infty(\Omega))}. \tag{5.7}$$

Step 2. Define $\tilde{\mathbf{u}}_h^0 = \{\tilde{u}_h^{\alpha,0}\}_{\alpha\in\mathcal{A}}$ as the finite element approximation of the unconstrained problems, that is,

$$a^\alpha(\tilde{u}_h^{\alpha,0}, v_h) = (f^\alpha, v_h), \quad \text{for all } v_h \in X_{0,h}^l.$$

Using (5.6), we can invoke standard finite element error estimates for linear problems to obtain

$$\|\hat{\mathbf{u}}^0 - \tilde{\mathbf{u}}_h^0\|_{\ell^\infty(\mathcal{A},L^\infty(\Omega))} \leq Ch^2 \,|\log h|^2. \tag{5.8}$$

For each $n \geq 0$ we define $\tilde{u}_h^{\alpha,n+1} \in \mathcal{K}_h(k, \hat{u}^{\alpha+1,n})$ as the solution of

$$a^\alpha(\tilde{u}_h^{\alpha,n+1}, \tilde{u}_h^{\alpha,n+1} - v_h) \leq (f^\alpha, \tilde{u}_h^{\alpha,n+1} - v_h), \quad \text{for all } v_h \in \mathcal{K}_h(k, I_h^{fe}\hat{u}^{\alpha+1,n}).$$

Note that this is simply the finite element approximation of $\hat{u}^{\alpha+1,n}$ as the solution to an obstacle problem. Once again using (5.6), we can invoke pointwise estimates for obstacle problems (see Baiocchi 1977, and Theorem 2.9 of Nochetto, Otárola and Salgado 2015) to conclude that

$$\|\hat{\mathbf{u}}^n - \tilde{\mathbf{u}}_h^n\|_{\ell^\infty(\mathcal{A},L^\infty(\Omega))} \leq Ch^2 \,|\log h|^2, \tag{5.9}$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

where the constant is independent of n.

Step 3. Introduce a discrete iterative scheme analogous to that given in Step 1. In other words, set $\hat{\mathbf{u}}_h^0 = \tilde{\mathbf{u}}_h^0$ and, for $n \geq 0$, find $\hat{u}_h^{\alpha,n+1} \in \mathcal{K}_h(k, \hat{u}_h^{\alpha+1,n})$ as the solution of

$$a^\alpha(\hat{u}_h^{\alpha,n+1}, \hat{u}_h^{\alpha,n+1} - v_h) \leq (f^\alpha, \hat{u}_h^{\alpha,n+1} - v_h), \quad \text{for all } v \in \mathcal{K}_h(k, \hat{u}_h^{\alpha+1,n}).$$

Similar techniques to those that led to (5.7) yield

$$\|\mathbf{u}_{h,k} - \hat{\mathbf{u}}_h^n\|_{\ell^\infty(\mathcal{A},L^\infty(\Omega))} \leq (1-\lambda)^n \|\hat{\mathbf{u}}_h^0\|_{\ell^\infty(\mathcal{A},L^\infty(\Omega))}. \tag{5.10}$$

Step 4. By induction, it can be shown that

$$\|\hat{\mathbf{u}}^n - \hat{\mathbf{u}}_h^n\|_{\ell^\infty(\mathcal{A},L^\infty(\Omega))} \leq \sum_{k=0}^{n} \|\hat{\mathbf{u}}^k - \tilde{\mathbf{u}}_h^k\|_{\ell^\infty(\mathcal{A},L^\infty(\Omega))}.$$

With this at hand, the triangle inequality yields

$$\begin{aligned}\|\mathbf{u}_k - \mathbf{u}_{h,k}\|_{\ell^\infty(\mathcal{A},L^\infty(\Omega))} \leq \|\mathbf{u}_k - \hat{\mathbf{u}}^n\|_{\ell^\infty(\mathcal{A},L^\infty(\Omega))} + \|\hat{\mathbf{u}}_h^n - \mathbf{u}_{h,k}\|_{\ell^\infty(\mathcal{A},L^\infty(\Omega))} \\ + \sum_{k=0}^{n} \|\hat{\mathbf{u}}^k - \tilde{\mathbf{u}}_h^k\|_{\ell^\infty(\mathcal{A},L^\infty(\Omega))},\end{aligned}$$

so that by using (5.7)–(5.10) we obtain

$$\|\mathbf{u}_k - \mathbf{u}_{h,k}\|_{\ell^\infty(\mathcal{A},L^\infty(\Omega))} \leq C\big[(1-\lambda)^n + nh^2\,|\log h|^2\big].$$

Now choose n so that $(1-\lambda)^n \approx h^2$ to obtain the result. □

Remark 5.10 ($k = 0$). An algorithm similar to that used in the proof of Theorem 5.9 is studied in Lions and Mercier (1980, Algorithme I) for $k = 0$. It is shown there that the ensuing iterates converge monotonically to $\mathbf{u}$, but no rate is given.

5.2.2. Discretization for HJB satisfying the Cordes condition

Let us now discuss the case of (5.1) with non-divergence-form operators without lower-order terms, that is, $\mathcal{L}^\alpha$. For simplicity, we set $g = 0$. More importantly, we will assume that the coefficient matrices A^α satisfy the Cordes condition of Definition 2.32. Furthermore, we will assume that the domain Ω is convex and, finally, that $\mathcal{A}$ is a compact metric space.

The stated assumptions imply, invoking Theorem 2.39, that each one of the operators $\mathcal{L}^\alpha$ is an isomorphism between $H^2(\Omega) \cap H_0^1(\Omega)$ and $L^2(\Omega)$. Consequently, we can apply the techniques of Section 4.1 to each one. Now, following the arguments of Smears and Süli (2014), let us show that these assumptions also imply that problem (5.1) is also well-posed and that its solution is strong.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

To do so we must assume that the Cordes condition holds uniformly in $\mathcal{A}$ which, from (2.16), yields the existence of $\epsilon > 0$ for which

$$\sup_{\alpha\in\mathcal{A}} \|\gamma^\alpha \mathcal{L}^\alpha v - \Delta v\|_{L^2(\Omega)} \leq \sqrt{1-\epsilon}\|D^2 v\|_{L^2(\Omega)}, \quad \text{for all } v \in H^2(\Omega). \tag{5.11}$$

This inequality motivates the definition of the (elliptic) nonlinear operator

$$F_\gamma[v] := \sup_{\alpha\in\mathcal{A}} \left[\gamma^\alpha(\mathcal{L}^\alpha v - f^\alpha)\right]. \tag{5.12}$$

The equivalence between problem $F_\gamma[u] = 0$ and (5.1) essentially follows from the continuity of the data and the positivity of γ^α. This result is summarized in the next lemma.

Lemma 5.11 (equivalence). Suppose that f^α and A^α are uniformly continuous for each $\alpha \in \mathcal{A}$, $\mathcal{A}$ is compact, and Ω is convex. The function $u \in H^2(\Omega) \cap H_0^1(\Omega)$ satisfies $F_\gamma[u] = 0$ a.e. Ω (*i.e.* u is a strong solution to this problem) if and only if it is a strong solution to (5.1).

Identity (5.11) and the algebraic identity $|\sup_{\alpha\in\mathcal{A}} x^\alpha - \sup_{\alpha\in\mathcal{A}} y^\alpha| \leq \sup_{\alpha\in\mathcal{A}} |x^\alpha - y^\alpha|$ for bounded sequences $\{x^\alpha\}, \{y^\alpha\} \subset \mathbb{R}$ then yields the following result.

Lemma 5.12 (continuity). For all $u, v \in H^2(\Omega)$ we have

$$|F_\gamma[v] - F_\gamma[w] - \Delta(v-w)| \leq \sqrt{1-\epsilon}|D^2(v-w)|.$$

Define

$$G(v,w) := \int_\Omega F_\gamma[v]\Delta w, \quad \text{for all } v, w \in H^2(\Omega) \cap H_0^1(\Omega).$$

The estimate of Lemma 5.12 leads to the following (strong) monotonicity property of G.

Theorem 5.13 (properties of G). Under the given assumptions, there is a positive constant C for which

$$G(v, v-w) - G(w, v-w) \geq C(1-\sqrt{1-\epsilon})\|v-w\|^2_{H^2(\Omega)} \tag{5.13}$$

for all $v, w \in H^2(\Omega) \cap H_0^1(\Omega)$. In addition, if $A^\alpha \in C(\bar{\Omega}, \mathbb{S}^d)$ and $f^\alpha \in C(\bar{\Omega})$, then

$$|G(v,u) - G(w,u)| \leq C\|v-w\|_{H^2(\Omega)}\|u\|_{H^2(\Omega)}. \tag{5.14}$$

Proof. Applying Lemma 5.12 yields

$$\begin{aligned} G(v, v-w) - G(w, v-w) &= \int_\Omega (F_\gamma[v] - F_\gamma[w])\Delta(v-w) \\ &= \|\Delta(v-w)\|^2_{L^2(\Omega)} + \int_\Omega (F_\gamma[v] - F_\gamma[w] - \Delta(v-w))\Delta(v-w) \\ &\geq \|\Delta(v-w)\|^2_{L^2(\Omega)} - \sqrt{1-\epsilon}\|D^2(v-w)\|_{L^2(\Omega)}\|\Delta(v-w)\|_{L^2(\Omega)}. \end{aligned}$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Since Ω is convex, we can apply the Miranda–Talenti estimate (2.17) to get

$$G(v, v-w) - G(w, v-w) \geq (1 - \sqrt{1-\epsilon})\|\Delta(v-w)\|^2_{L^2(\Omega)}.$$

Inequality (5.13) now follows from the equivalence of $\|\Delta\cdot\|_{L^2(\Omega)}$ and $\|\cdot\|_{H^2(\Omega)}$ on convex domains.

Finally, the Lipschitz property (5.14) follows from the continuity of the data and the Cauchy–Schwarz inequality. □

Theorem 5.13 and Lemma 5.11, along with the Browder–Minty theorem and the fact that $\Delta : H^2(\Omega) \cap H^1_0(\Omega) \to L^2(\Omega)$ is surjective on convex domains, yield the following existence and uniqueness result for the HJB problem (5.1); see Smears and Süli (2014, Theorem 3) for details.

Theorem 5.14 (existence and uniqueness). Suppose that $\mathcal{A}$ is a compact metric space, $f^\alpha \in C(\bar{\Omega})$, and $A^\alpha \in C(\bar{\Omega}, \mathbb{S}^d)$ satisfies the Cordes condition for each $\alpha \in \mathcal{A}$. Then there exists a unique strong solution $u \in H^2(\Omega) \cap H^1_0(\Omega)$ to $F_\gamma[u] = 0$. Further, u is also the unique strong solution to (5.1).

Let us now discuss how the finite element methods for linear problems given in Section 4.1 are extended to the nonlinear problem (5.1). While any of the methods given in that section can be adopted for the nonlinear problem, here we focus on the DG approximation (Smears and Süli 2013, 2014, 2016).

Recall that X_h^{dg}, defined by (4.11), is the piecewise polynomial space of degree k with respect to a conforming and shape-regular, simplicial partition of Ω. As in Section 4.1, we will only consider the h-version of the method, where the polynomial degree is globally fixed, and we will not follow the dependence on the polynomial degree k of all the ensuing constants. We refer the reader to Smears and Süli (2014), who consider hp-DG methods.

The DG method for the linear problem (4.17) extends to the nonlinear problem (5.1) by essentially taking the supremum over $\mathcal{A}$. This yields the following scheme: find $u_h \in X_h^{dg}$ such that

$$G_h^{dg}(u_h, v_h) := \sum_{K \in \mathcal{T}_h} \int_K \big[F_\gamma[u_h] - \Delta u_h\big]\Delta v_h + \frac{1}{2} B_h(u_h, v_h) = 0, \qquad (5.15)$$

where $B_h(\cdot,\cdot)$ is the bilinear form defined by (4.14) with penalty parameter $\mu > 0$. Recall from Section 4.1 that if $u \in H^s(\Omega) \cap H^1_0(\Omega)$ with $s > 5/2$, then

$$B_h(u, v_h) = 2 \sum_{K \in \mathcal{T}_h} \int_K \Delta u \Delta v_h, \quad \text{for all } v_h \in X_h^{dg}.$$

Therefore (5.15) is a consistent method, that is, $G_h^{dg}(u, v_h) = 0$ for all $v_h \in X_h^{dg}$ provided u is sufficiently smooth. In addition, as in the linear case, the

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

coercivity properties of B_h ensure that the structural properties found at the continuous level carry over. That is, we have the following result (see Smears and Süli 2014, Theorem 7).

Theorem 5.15 (properties of G_h^{dg}). Suppose that the hypotheses in Theorem 5.13 are satisfied. Then there exists $\mu_* = O(\epsilon^{-1})$ such that if $\mu \geq \mu_*$, then

$$C\|v_h - w_h\|_{DG(1)}^2 \leq G_h^{dg}(v_h, v_h - w_h) - G_h^{dg}(w_h, v_h - w_h),$$

for all $v_h, w_h \in X_h^{dg}$, where the discrete $H^2(\Omega)$-norm $\|\cdot\|_{DG(1)}$ is defined by (4.15). Moreover, there exists a constant $C > 0$ such that, for all $v_h, w_h, u_h \in X_h^{cg}$,

$$|G_h^{dg}(v_h, u_h) - G_h^{dg}(w_h, u_h)| \leq C\|v_h - w_h\|_{DG(1)}\|u_h\|_{DG(1)}.$$

Similar to the continuous setting, Theorem 5.15 yields existence and uniqueness for the DG method.

Corollary 5.16 (existence and uniqueness). If the hypotheses of Theorem 5.15 hold, then there exists a unique $u_h \in X_h^{dg}$ satisfying (5.15).

Finally, stability/monotonicity and the consistency of the method lead to the following error estimates. We refer to Smears and Süli (2014, Theorem 8) for a proof.

Theorem 5.17 (rates of convergence). Suppose that the hypotheses of Theorem 5.15 hold. Let $u_h \in X_h^{dg}$ be the solution to (5.15), and suppose that the solution to (5.1) has regularity $u \in H^s(\Omega)$ with $5/2 < s \leq k+1$. Then the error satisfies

$$\|u - u_h\|_{DG(1)} \leq Ch^{s-2}\|u\|_{H^s(\Omega)}.$$

5.2.3. Finite element methods for parabolic isotropic HJB problems

In this section we depart from the elliptic framework and discuss the finite element method for *parabolic* (time-dependent) Hamilton–Jacobi–Bellman problems developed by Jensen and Smears (2013*b*). Assuming homogeneous Dirichlet boundary conditions, we consider numerical approximations of the problem:

$$\begin{cases} \dfrac{\partial u}{\partial t} - \sup\limits_{\alpha \in \mathcal{A}}[\tilde{\mathcal{L}}^\alpha u - f^\alpha] = 0 & \text{in } \Omega \times (0,T), \\ u = 0 & \text{on } \partial\Omega \times (0,T), \\ u = u_0 & \text{on } \bar{\Omega} \times \{0\}, \end{cases} \tag{5.16}$$

where $u_0 \in C(\bar{\Omega})$ is the initial time condition and T is the (finite) end-time. As before, we assume that $\mathcal{A}$ is compact and that the coefficients

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

are uniformly continuous in $\bar{\Omega}$. In addition, we assume that the coefficients are time-independent, $f^\alpha \geq 0$, $c^\alpha \leq 0$, $u_0 \geq 0$, and, most importantly, the elliptic operator $\mathcal{L}^\alpha$ is isotropic, that is,

$$\tilde{\mathcal{L}}^\alpha u(x) = a^\alpha(x)\Delta u(x) + \mathbf{b}^\alpha(x) \cdot Du(x) + c^\alpha(x)u(x),$$

with $a^\alpha \geq 0$. Note that, even in the isotropic case, each elliptic operator is in non-divergence form.

Before discussing the finite element method let us first extend the definition of viscosity solutions to the parabolic setting (compare Definition 2.48).

Definition 5.18 (viscosity solution).

(a) A function $u_\star \in USC(\bar{\Omega} \times [0,T])$ is a viscosity subsolution to (5.16) if $u_\star \leq 0$ on $\partial\Omega \times (0,T)$, $u_\star \leq u_0$ on $\bar{\Omega} \times \{0\}$, and if, whenever

$$(x_0,t_0) \in \Omega \times (0,T), \quad \varphi \in C^2((0,T) \times \Omega)$$

and $u_\star - \varphi$ has a local maximum at (x_0,t_0), we have that

$$\frac{\partial \varphi}{\partial t}(x_0,t_0) - \sup_{\alpha \in \mathcal{A}}(\tilde{\mathcal{L}}^\alpha(x_0)\varphi(x_0,t_0) - f^\alpha(x_0)) \leq 0.$$

(b) A function $u^\star \in LSC(\bar{\Omega} \times [0,T])$ is a viscosity supersolution to (5.16) if $u^\star \geq 0$ on $\partial\Omega \times (0,T)$, $u^\star \geq u_0$ on $\bar{\Omega} \times \{0\}$, and if, whenever

$$(x_0,t_0) \in \Omega \times (0,T), \quad \varphi \in C^2((0,T) \times \Omega)$$

and $u_\star - \varphi$ has a local minimum at (x_0,t_0), we have that

$$\frac{\partial \varphi}{\partial t}(x_0,t_0) - \sup_{\alpha \in \mathcal{A}}(\tilde{\mathcal{L}}^\alpha(x_0)\varphi(x_0,t_0) - f^\alpha(x_0)) \geq 0.$$

(c) A function $u \in C(\bar{\Omega} \times [0,T])$ is a viscosity solution to (5.16) if it is both a sub- and supersolution.

The discretization of the non-divergence part of $\tilde{\mathcal{L}}^\alpha$ proposed by Jensen and Smears (2013*b*) is based on a 'freezing the coefficients' strategy. Let $\{\phi_i\} \subset X^l_{0,h}$ be the normalized hat functions defined in Section 3.5, where we recall that $X^l_{0,h}$ is the space of piecewise linear polynomials with vanishing trace (see (3.35)). Then, since ϕ_i is essentially a regularized Dirac delta distribution, if a point $x \in \Omega$ is close to a node $z_i \in \Omega^I_h$, then we have, at least formally,

$$a^\alpha(x)\Delta u(x) \approx -a^\alpha(z_i) \int_\Omega Du \cdot D\phi_i = a^\alpha(z_i)\Delta_h I^{ep}_h u(z_i),$$

where the finite element Laplacian Δ_h is given by (3.39), and I^{ep}_h is the elliptic projection. This heuristic approximation motivates the following semi-discrete finite element method: find $u_h : [0,T] \to X^l_{0,h}$ such that

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

$u_h(0) = I_h^{fe} u_0$ and

$$\frac{\partial u_h}{\partial t} - \sup_{\alpha \in \mathcal{A}} (\tilde{\mathcal{L}}_h^\alpha u_h - f_h^\alpha) = 0 \quad \text{in } \Omega_h^I \times (0,T), \tag{5.17}$$

where $I_h^{fe} : C^0(\bar{\Omega}) \to X_h^l$ is the nodal interpolant onto X_h^l, and the discrete operator and discrete source function are given respectively by

$$\tilde{\mathcal{L}}_h^\alpha u_h(z_i) = a^\alpha(z_i) \Delta_h u_h(z_i) + \int_\Omega (\mathbf{b}^\alpha \cdot Du_h + c^\alpha u_h)\phi_i,$$
$$f_h^\alpha(z_i) = \int_\Omega f^\alpha \phi_i.$$

The fully discrete method proposed in Jensen and Smears (2013*b*) applies a one-step explicit–implicit ODE solver to a regularized version of (5.17). Let $\tau \in (0,1)$ be the (uniform) time step size and assume that $T/\tau =: M \in \mathbb{N}$. For a sequence of discrete functions $\{v_h^k\}_{k=0}^M \subset X_{0,h}^l$, we define the backward difference quotient as

$$d_\tau v_h^{k+1} := \frac{1}{\tau}(v_h^{k+1} - v_h^k) \in X_{0,h}^l.$$

For each $\alpha \in \mathcal{A}$, the discrete operator $\tilde{\mathcal{L}}_h^\alpha$ is approximately split into an explicit and implicit part,

$$\tilde{\mathcal{L}}_h^\alpha \approx \mathcal{E}_h^\alpha + \mathcal{I}_h^\alpha, \tag{5.18}$$

with

$$\mathcal{E}_h^\alpha u_h(z_i) = a_e^\alpha(z_i) \Delta_h u_h(z_i) + \int_\Omega (\mathbf{b}_e^\alpha \cdot Du_h + c_e u_h)\phi_i,$$
$$\mathcal{I}_h^\alpha u_h(z_i) = a_i^\alpha(z_i) \Delta_h u_h(z_i) + \int_\Omega (\mathbf{b}_i^\alpha \cdot Du_h + c_i u_h)\phi_i.$$

We now consider the following fully discrete method: find the sequence $\{u_h^k\}_{k=0}^M \subset X_{0,h}^l$ with $u_h^0 = I_h^{fe} u_0$ and

$$d_\tau u_h^{k+1} - \sup_{\alpha \in \mathcal{A}} (\mathcal{E}_h^\alpha u_h^k + \mathcal{I}_h^\alpha u_h^{k+1} - f_h^\alpha) \quad \text{in } \Omega_h^I. \tag{5.19}$$

To show the well-posedness and to analyse method (5.19), we make several assumptions. First we quantify the approximation in (5.18) and make assumptions of the coefficients in the explicit and implicit operators.

Assumption 5.19 (coefficients). We assume that the coefficients satisfy

$$\lim_{h \to 0} \sup_{\alpha \in \mathcal{A}} \Big(\sup_{z \in \Omega_h^I} \|a^\alpha - (a_e^\alpha(z) + a_i^\alpha(z))\|_{L^\infty(\omega_z)} + \|\mathbf{b}^\alpha - (\mathbf{b}_e^\alpha + \mathbf{b}_i^\alpha)\|_{L^\infty(\Omega)} + \|c^\alpha - (c_e^\alpha + c_i^\alpha)\|_{L^\infty(\Omega)} \Big) = 0.$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

In addition, when $\mathcal{A}$ is not finite, we also assume that the mappings $\alpha \mapsto (a_e^\alpha, \mathbf{b}_e^\alpha, c_i^\alpha)$ and $\alpha \mapsto (a_i^\alpha, \mathbf{b}_i^\alpha, c_i^\alpha)$ are continuous, $a_e, a_i \geq 0$, and $c_e, c_i \leq 0$.

Next we make assumptions regarding the monotonicity properties of the operators.

Assumption 5.20 (monotonicity). The splitting (5.18) has the following properties.

(a) The time-step explicit operators satisfy

$$\delta_{i,j} + \tau \mathcal{E}_h^\alpha(\tilde{\phi}_j)(z_i) \geq 0, \quad \text{for all } \alpha \in \mathcal{A}, \tag{5.20}$$

where $\{\tilde{\phi}_j\} \subset X_{0,h}^l$ are the (unnormalized) hat functions satisfying $\tilde{\phi}_j(z_i) = \delta_{i,j}$.

(b) The operator $\mathcal{I}_h^\alpha$ is monotone (in the sense of Definition 3.5) for each $\alpha \in \mathcal{A}$.

Condition (5.20) is essentially a time-step restriction and a monotonicity condition (if $\mathcal{E}_h^\alpha \not\equiv 0$ for all $\alpha \in \mathcal{A}$). For example, if $c_e^\alpha \equiv 0$, $\mathbf{b}_e^\alpha \equiv 0$, and $a_e^\alpha > 0$ for all $\alpha \in \mathcal{A}$, then the condition reads

$$\delta_{i,j} - \tau a_e^\alpha(z_i) \int_\Omega D\tilde{\phi}_j \cdot D\phi_i \geq 0.$$

Therefore, in this setting, (5.20) is satisfied if and only if (3.42) holds (so that Δ_h is monotone) and if

$$\tau \leq \left(a_e^\alpha(z_i) \int_\Omega D\phi_i \cdot D\tilde{\phi}_i \right)^{-1} = O(h^2).$$

The monotonicity of $\mathcal{I}_h^\alpha$ implies that the inverse of $(I - \tau \mathcal{I}_h^\alpha)$ is sign-preserving for any $\tau > 0$, that is, if $(I - \tau \mathcal{I}_h^\alpha) v_h \leq 0$ (≥ 0), then $v_h \leq 0$ (≥ 0). To see this, suppose that $v_h \in X_{0,h}^l$ satisfies $(I - \tau \mathcal{I}_h^\alpha) v_h \leq 0$ and v_h has a positive maximum at z_i. Then the monotonicity of $\mathcal{I}_h^\alpha$ implies $\mathcal{I}_h^\alpha v_h(z_i) \leq 0$ (see Definition 3.5), leading to $(I - \tau \mathcal{I}_h^\alpha) v_h(z_i) \geq v_h(z_i) > 0$, a contradiction. This argument also shows that $(I - \tau \mathcal{I}_h^\alpha)$ is invertible, and therefore the following problem is well-posed. For fixed $\alpha \in \mathcal{A}$, find $\{u_{h,\alpha}^k\}_{k=0}^M \subset X_{0,h}^l$ such that $u_{h,\alpha}^0 = I_h^{fe} u_0$ and $(k = 0, 1, \ldots, M-1)$

$$d_\tau u_{h,\alpha}^{k+1} - (\mathcal{E}_h^\alpha u_{h,\alpha}^k + \mathcal{I}_h^\alpha u_{h,\alpha}^{k+1} - f_h) \ \text{ in } \Omega_h^I, \quad u_{h,\alpha}^0 = I_h^{fe} u_0. \tag{5.21}$$

The construction of operators $\mathcal{E}_h^\alpha$, $\mathcal{I}_h^\alpha$ satisfying Assumptions 5.19–5.20 (on weakly acute triangulations) using the method of artificial diffusion can be found in Jensen and Smears (2013*b*, Section 8) and Jensen and Smears (2013*a*).

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Theorem 5.21 (existence). Suppose that Assumptions 5.19 and 5.20 are satisfied. Then for each $k \in \{0, 1, \dots, M-1\}$ there exists a unique solution $u_h^{k+1} \in X_{0,h}^l$ to (5.19). Moreover, if we let $\{u_{h,\alpha}^k\} \subset X_{0,h}^l$ denote the solution to (5.21), then $0 \le u_h^k \le u_{h,\alpha}^k$.

The proof of existence and uniqueness follows from the convergence results of Howard's method, which is discussed in the next section. The monotonicity and continuity properties of $\mathcal{E}_h^\alpha$, $\mathcal{I}_h^\alpha$ ensure that the hypotheses of Theorem 5.31 below are satisfied (see Jensen and Smears 2013*b*, Theorem 3.1).

The next result establishes the stability of the method. Its proof essentially follows from Assumption 5.20 and the non-positivity of c_e^α and c_i^α; see Jensen and Smears (2013*b*, Lemma 3.2 and Corollary 3.3) for details.

Lemma 5.22 (stability). Suppose that the assumptions of Theorem 5.21 hold, and let $\{u_{h,\alpha}^k\} \subset X_{0,h}^l$ denote the solution to (5.21) for a fixed $\alpha \in \mathcal{A}$. Then we have

$$\|\{u_{h,\alpha}^k\}_{k=0}^M\|_{\ell^\infty(\mathbb{N}\cap[0,M],L^\infty(\Omega))} \le \|I_h^{fe} u_0\|_{L^\infty(\Omega)} + T\|f^\alpha\|_{L^\infty(\Omega)}.$$

Therefore, by Theorem 5.21,

$$\|\{u_h^k\}_{k=0}^M\|_{\ell^\infty(\mathbb{N}\cap[0,M],L^\infty(\Omega))} \le \|I_h^{fe} u_0\|_{L^\infty(\Omega)} + T \inf_{\alpha\in\mathcal{A}} \|f^\alpha\|_{L^\infty(\Omega)}.$$

Next, to apply the Barles–Souganidis theory, we look at the consistency of the method. Essentially this result follows from Lemma 3.41 and the stability of the elliptic projection.

Lemma 5.23 (consistency). Let $I_h^{ep} : H^1(\Omega) \to X_h^l$ be the elliptic projection onto X_h^l. Let $\{(z_h, t_\tau)\}_{h>0,\tau>0}$ with $(z_h, t_\tau) \in \bar{\Omega}_h \times \mathbb{N} \cap [0, M]$ and $(z_h, t_\tau) \to (z_0, t_0) \in \bar{\Omega}\times[0,T]$ as $h, \tau \to 0^+$. Then, for any $\phi \in C^2(\bar{\Omega}\times[0,T])$,

$$\lim_{h,\tau\to 0^+} \big(d_\tau I_h^{ep}\phi(t_\tau+\tau) - (\mathcal{E}_h^\alpha I_h^{ep}(\phi(t_\tau)) + \mathcal{I}_h^\alpha I_h^{ep}(\phi(t_\tau+\tau)) - f_h^\alpha)\big)(z_h)$$
$$= \frac{\partial\phi}{\partial t}(z_0,t_0) - (\tilde{\mathcal{L}}^\alpha\phi(z_0,t_0) - f^\alpha(z_0)), \quad \text{for all } \alpha \in \mathcal{A},$$

where $\phi(t) := \phi(\cdot, t)$, and the convergence is uniform with respect to $\alpha \in \mathcal{A}$.

Remark 5.24 (anisotropy). Note that the consistency result given in Lemma 5.23 does not extend to operators with anisotropic diffusion.

Finally to apply Theorem 3.8, the following assumption is made, which ensures that the limiting solution satisfies the boundary conditions in a classical sense.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Assumption 5.25 (limiting boundary values). Define $\bar{u}_\alpha \in USC(\bar{\Omega} \times [0,T])$:

$$\bar{u}_\alpha(x,t) := \limsup_{\substack{(z,k\tau)\to(x,t)\\ h,\tau\to 0^+}} u_{h,\alpha}^k(z),$$

where $u_{h,\alpha}^k$ is the solution to (5.21). Then

$$\inf_{\alpha\in\mathcal{A}} \bar{u}_\alpha(x,t) = 0, \quad \text{for all } (x,t) \in \partial\Omega \times [0,T].$$

Finally, combining Lemmas 5.23 and 5.22 with Theorem 3.8 yields the following convergence result; see Jensen and Smears (2013*b*, Theorem 6.2).

Theorem 5.26 (convergence). Suppose that Assumptions 5.19, 5.20 and 5.25 are satisfied. Suppose further that the HJB problem (5.16) satisfies a comparison principle (see Definition 2.49). Then

$$\lim_{h,\tau\to 0^+} \max_k \|u(\cdot,k\tau) - u_h^k\|_{L^\infty(\Omega)} = 0.$$

5.3. Solution of the discrete problems

To finalize the discussion concerning the approximation of convex equations, we must, at least briefly, discuss how to solve the system of nonlinear equations that arises after discretization, whether by finite differences or finite elements. We will present one of the most popular methods – known as policy iterations or Howard's algorithm – and discuss its convergence properties.

Once (5.1) is discretized by any of the methods described in the previous sections, we end up with the following system of nonlinear equations: find $\mathbf{x} \in \mathbb{R}^N$ that satisfies

$$\mathbf{F}(\mathbf{x}) = \sup_{\alpha\in\mathcal{A}} [\mathbf{K}^\alpha \mathbf{x} - \mathbf{f}^\alpha] = \mathbf{0}, \tag{5.22}$$

where the supremum is taken component-wise, N is the number of degrees of freedom in the discretization, $\{\mathbf{K}^\alpha\}_{\alpha\in\mathcal{A}}$ and $\{\mathbf{f}^\alpha\}_{\alpha\in\mathcal{A}}$ are discretizations of $\{\tilde{\mathcal{L}}^\alpha\}_{\alpha\in\mathcal{A}}$ and $\{f^\alpha\}_{\alpha\in\mathcal{A}}$, respectively, or as in (5.19), come from the implicit-explicit splitting of the operators (5.18).

In order to define Howard's algorithm we must introduce the following operations on $\mathcal{A}$, the matrices $\mathbf{K}^\alpha$ and vectors $\mathbf{f}^\alpha$. Given $\mathbf{y} \in \mathbb{R}^N$ and $i \in \{1,\dots,N\}$ we define the element $\alpha(\mathbf{y},i) \in \mathcal{A}$ as the element that realizes the supremum in (5.22) when applied to $\mathbf{y}$ at component i, that is,

$$[\mathbf{K}^{\alpha(\mathbf{y},i)}\mathbf{y} - \mathbf{f}^{\alpha(\mathbf{y},i)}]_i = \sup_{\alpha\in\mathcal{A}} [\mathbf{K}^\alpha \mathbf{y} - \mathbf{f}^\alpha]_i.$$

We define $\boldsymbol{\alpha}(\mathbf{y}) \in \mathcal{A}^N$ with components $\boldsymbol{\alpha}(\mathbf{y})_i = \boldsymbol{\alpha}(\mathbf{y},i)$. Given $\boldsymbol{\alpha}(\mathbf{y})$ we

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

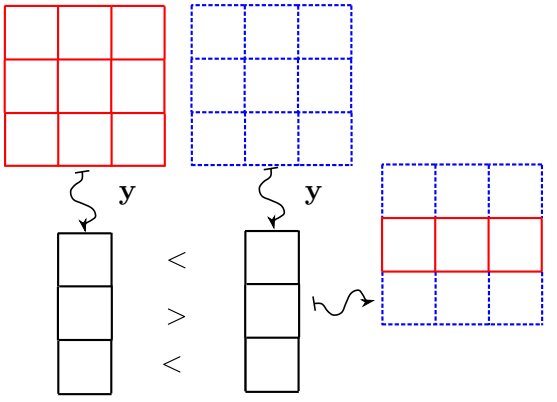


Figure 5.1. An illustration of the selection procedure defined in (5.23), for the simplified case of $\mathbf{f}^{\alpha} = \mathbf{0}$ for all $\alpha \in \mathcal{A}$ and $\#\mathcal{A} = 2$. We multiply each matrix (`solid red` and `dashed blue` in the figure) by the vector $\mathbf{y}$ and compare the components of the results. The matrix $\mathbf{K}^{\boldsymbol{\alpha}(\mathbf{y})}$ is constructed by choosing the row of the matrix that gives the largest result.

finally define the matrix $\mathbf{K}^{\boldsymbol{\alpha}(\mathbf{y})}$ and vector $\mathbf{f}^{\boldsymbol{\alpha}(\mathbf{y})}$ as follows:

$$\mathbf{K}_{i,j}^{\boldsymbol{\alpha}(\mathbf{y})} = \mathbf{K}_{i,j}^{\boldsymbol{\alpha}(\mathbf{y},i)}, \quad \mathbf{f}_{i}^{\boldsymbol{\alpha}(\mathbf{y})} = \mathbf{f}_{i}^{\boldsymbol{\alpha}(\mathbf{y},i)}. \tag{5.23}$$

An illustration of this selection procedure for $\#\mathcal{A} = 2$ is shown in Figure 5.1. With these operations at hand, Howard's algorithm, as presented in Bokanowski, Maroso and Zidani (2009), is described in Algorithm 5.1.

Note that at each step Algorithm 5.1 requires the solution to the linear system (5.24). Because of the way the matrices are constructed, it is necessary to ensure that they are invertible. This is guaranteed by the following condition.

Assumption 5.27 (monotonicity). For every $\mathbf{y} \in \mathbb{R}^N$ the matrix $\mathbf{K}^{\boldsymbol{\alpha}(\mathbf{y})}$ is monotone in the sense of Definition 3.5.

It is natural to wonder when such a condition is satisfied. As shown by Witte and Reisinger (2012), if, for every $\alpha \in \mathcal{A}$, the matrix $\mathbf{K}^{\alpha}$ is strictly diagonally dominant and monotone and its off-diagonal entries are non-positive, then this condition is satisfied. We comment finally that a monotone matrix is non-singular.

Having guaranteed that Howard's algorithm can proceed, we focus our attention on its convergence. While many references have studied it, we will follow Bokanowski *et al.* (2009) and Hintermüller, Ito and Kunisch (2002), and draw a connection between this method and active set strategies which, in turn, can be analysed as semismooth Newton methods. We begin with the definition of slant derivative.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Algorithm 5.1: Howard's algorithm.

Input : Set $\mathcal{A}$.
Matrices $\{\mathbf{K}^\alpha\}_{\alpha\in\mathcal{A}} \subset \mathbb{R}^{N\times N}$.
right-hand sides $\{\mathbf{f}^\alpha\}_{\alpha\in\mathcal{A}} \subset \mathbb{R}^N$.
Output: Vector $\mathbf{x} \in \mathbb{R}^N$, solution of (5.22).

Initialization: Choose $\mathbf{x}_{-1} \in \mathbb{R}^N$;
for $k \geq 0$ **do**
  Set $\boldsymbol{\alpha}_k = \boldsymbol{\alpha}(\mathbf{x}_{k-1})$;
  Compute $\mathbf{K}^{\boldsymbol{\alpha}_k} \in \mathbb{R}^{N\times N}$ and $\mathbf{f}^{\boldsymbol{\alpha}_k} \in \mathbb{R}^N$;
  Find: $\mathbf{x}_k \in \mathbb{R}^N$ that solves

$$\mathbf{K}^{\boldsymbol{\alpha}_k}\mathbf{x}_k = \mathbf{f}^{\boldsymbol{\alpha}_k}; \tag{5.24}$$

  if $k \geq 1$ **and** $\mathbf{x}_k = \mathbf{x}_{k-1}$ **then**
    return $\mathbf{x}_k$;
  end
end

Definition 5.28 (slant derivative). Let X and Z be Banach spaces and let $D \subset X$ be open. The mapping $F : D \to Z$ is called *slantly differentiable* in the open subset $U \subset D$ if there is a family of mappings $G : U \to \mathfrak{L}(X, Z)$ such that, for every $x \in U$,

$$\lim_{h\to 0} \frac{1}{\|h\|_X} \|F(x + h) - F(x) - G(x + h)h\|_Z = 0.$$

We call G a *slanting function* for F.

To approximate $x^\star \in X$, the solution of $F(x) = 0$, one can then apply the semismooth Newton method which, starting from $x_0 \in D$, computes iterates via

$$x_{k+1} = x_k - G(x_k)^{-1}F(x_k). \tag{5.25}$$

The convergence of (5.25) is given in the following result, whose proof can be found in Hintermüller *et al.* (2002, Theorem 1.1).

Theorem 5.29 (convergence). Assume that F is slantly differentiable in a neighbourhood of $x^\star$. If $G(x)$ is non-singular for all $x \in U$ and $\{\|G(x)^{-1}\| : x \in U\}$ is bounded, then the semismooth Newton method (5.25) converges superlinearly provided x_0 is sufficiently close to $x^\star$.

The connection between Howard's algorithm and semismooth Newton methods is given by the following result; see Hintermüller *et al.* (2002, Lemma 3.1).

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Lemma 5.30 (slant derivative of max). The mapping

$$\mathbf{M} : \mathbb{R}^N \ni \mathbf{y} \mapsto \max\{\mathbf{0}, \mathbf{y}\} \in \mathbb{R}^N$$

is slantly differentiable on $\mathbb{R}^N$ and a slanting function is

$$\mathbf{G}_{\mathbf{M}}(\mathbf{y})_{i,j} = g(\mathbf{y}_j)\delta_{i,j}, \quad g(z) = \begin{cases} 0 & z \leq 0, \\ 1 & z > 0. \end{cases}$$

Let us explain how this result relates to Howard's algorithm in the (perhaps overly) simplistic case that $N = 1$, $\mathcal{A} = \{1, 2\}$ and $\mathbf{f}^1 = \mathbf{f}^2 = \mathbf{f}$. In this setting, we are looking for $\mathbf{x} \in \mathbb{R}$ that solves

$$\mathbf{0} = \mathbf{F}(\mathbf{x}) = \max\{\mathbf{K}^1\mathbf{x}, \mathbf{K}^2\mathbf{x}\} - \mathbf{f} = \mathbf{K}^1\mathbf{x} + \max\{\mathbf{0}, (\mathbf{K}^2 - \mathbf{K}^1)\mathbf{x}\} - \mathbf{f}.$$

By Lemma 5.30 a slanting function for $\mathbf{F}$ at $\mathbf{y}$ is

$$\mathbf{G}(\mathbf{y}) = \mathbf{K}^1 + \mathbf{G}_{\mathbf{M}}((\mathbf{K}^2 - \mathbf{K}^1)\mathbf{y})(\mathbf{K}^2 - \mathbf{K}^1).$$

If $(\mathbf{K}^2 - \mathbf{K}^1)\mathbf{y} \geq 0$, then

$$\mathbf{G}(\mathbf{y}) = \mathbf{K}^1 + (\mathbf{K}^2 - \mathbf{K}^1) = \mathbf{K}^2,$$

and, similarly, if $(\mathbf{K}^2 - \mathbf{K}^1)\mathbf{y} \leq 0$ we get $\mathbf{G}(\mathbf{y}) = \mathbf{K}^1$. In other words, $\mathbf{G}(\mathbf{y})$ always coincides with the coefficient that gives the largest result. Having made this observation, we can now establish convergence for Howard's algorithm (Bokanowski *et al.* 2009).

Theorem 5.31 (convergence). If Assumption 5.27 is satisfied, then the sequence $\{\mathbf{x}_k\}_{k\geq 0}$, generated by Algorithm 5.1, satisfies $\mathbf{x}_k \geq \mathbf{x}_{k+1}$ for every $k \geq 0$ and converges to $\mathbf{x}$, the solution of (5.22). Moreover,

(1) If $\#\mathcal{A}$ is finite, the sequence converges in at most $(\#\mathcal{A})^N$ steps.

(2) If $\mathcal{A}$ is infinite, the convergence is asymptotically superlinear.

Proof. To draw a connection between Algorithm 5.1 and the semismooth Newton method (5.25), let us sketch the proof. Recall that we are looking for a solution of (5.22), which can be understood as looking for a zero of the map $\mathbf{F}$. Now, following Bokanowski *et al.* (2009, Theorem 3.8), let us show that for $\mathbf{y} \in \mathbb{R}^N$ the matrix $\mathbf{K}^{\boldsymbol{\alpha}(\mathbf{y})}$ is a slanting function for $\mathbf{F}$ at the point $\mathbf{y}$. For if that is the case, then (5.25) can be rewritten as

$$\mathbf{K}^{\boldsymbol{\alpha}(\mathbf{x}_k)}(\mathbf{x}_{k+1} - \mathbf{x}_k) = -\mathbf{F}(\mathbf{x}_k) = -\mathbf{K}^{\boldsymbol{\alpha}(\mathbf{x}_k)}\mathbf{x}_k + \mathbf{f}^{\boldsymbol{\alpha}(\mathbf{x}_k)},$$

which is clearly (5.24). By invoking Theorem 5.29 this will yield the superlinear convergence of $\{\mathbf{x}_k\}_{k\geq 0}$.

Now, to show the slant-differentiability of $\mathbf{F}$, let $\mathbf{y}, \mathbf{h} \in \mathbb{R}^N$ and note that

$$\begin{aligned} \mathbf{F}(\mathbf{y}) + \mathbf{K}^{\boldsymbol{\alpha}(\mathbf{y}+\mathbf{h})}\mathbf{h} &\geq \mathbf{K}^{\boldsymbol{\alpha}(\mathbf{y}+\mathbf{h})}\mathbf{y} - \mathbf{f}^{\boldsymbol{\alpha}(\mathbf{y}+\mathbf{h})} + \mathbf{K}^{\boldsymbol{\alpha}(\mathbf{y}+\mathbf{h})}\mathbf{h} = \mathbf{F}(\mathbf{y}+\mathbf{h}) \\ &\geq \mathbf{K}^{\boldsymbol{\alpha}(\mathbf{y})}(\mathbf{y}+\mathbf{h}) - \mathbf{f}^{\boldsymbol{\alpha}(\mathbf{y})} = \mathbf{F}(\mathbf{y}) + \mathbf{K}^{\boldsymbol{\alpha}(\mathbf{y})}\mathbf{h}. \end{aligned}$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Consequently,

$$\begin{aligned}\mathbf{0} &\geq \mathbf{F}(\mathbf{y}+\mathbf{h}) - \mathbf{F}(\mathbf{y}) - \mathbf{K}^{\boldsymbol{\alpha}(\mathbf{y}+\mathbf{h})}\mathbf{h} \geq [\mathbf{K}^{\boldsymbol{\alpha}(\mathbf{y})} - \mathbf{K}^{\boldsymbol{\alpha}(\mathbf{y}+\mathbf{h})}]\mathbf{h} \\ &\geq -|\mathbf{K}^{\boldsymbol{\alpha}(\mathbf{y})} - \mathbf{K}^{\boldsymbol{\alpha}(\mathbf{y}+\mathbf{h})}|_\infty |\mathbf{h}|_\infty.\end{aligned}$$

Since Lemma 3.2 of Bokanowski *et al.* (2009) shows that $\lim_{|\mathbf{h}|_\infty \to 0} |\mathbf{K}^{\boldsymbol{\alpha}(\mathbf{y})} - \mathbf{K}^{\boldsymbol{\alpha}(\mathbf{y}+\mathbf{h})}|_\infty = 0$, the result follows.

When $\#\mathcal{A}$ is finite we remark that convergence in a finite number of steps follows from the monotonicity of the iterates. To show the monotonicity, we exploit the construction of the parameters $\boldsymbol{\alpha}_k$. Indeed, since

$$\mathbf{K}^{\boldsymbol{\alpha}_{k+1}}\mathbf{x}_k - \mathbf{f}^{\boldsymbol{\alpha}_{k+1}} = \mathbf{F}(\mathbf{x}_k) \geq \mathbf{K}^{\boldsymbol{\alpha}_k}\mathbf{x}_k - \mathbf{f}^{\boldsymbol{\alpha}_k} = \mathbf{0} = \mathbf{K}^{\boldsymbol{\alpha}_{k+1}}\mathbf{x}_{k+1} - \mathbf{f}^{\boldsymbol{\alpha}_{k+1}},$$

we conclude that $\mathbf{K}^{\boldsymbol{\alpha}_{k+1}}(\mathbf{x}_k - \mathbf{x}_{k+1}) \geq 0$. From Assumption 5.27 we have that $\mathbf{K}^{\boldsymbol{\alpha}_{k+1}}$ is monotone and therefore $\mathbf{x}_k \geq \mathbf{x}_{k+1}$. This concludes the proof. □

We conclude by commenting that in Algorithm 5.1 it is necessary to solve the linear system of equations (5.24) at every step, which, for large N, can be very time-consuming. We refer to Hoppe (1986), who describes a multilevel technique to solve this problem and shows its global convergence. Variations of this multigrid strategy have been explored by Alla, Falcone and Kalise (2015) and Han and Wan (2013). Numerical methods of penalty type have been studied by Witte and Reisinger (2012) and Lions and Mercier (1980, Algorithme III).

6. The Monge–Ampère equation

The focus of this section is the study of discretizations for a particular convex equation, the Monge–Ampère equation with Dirichlet boundary conditions:

$$\begin{cases}\det(D^2 u) = f & \text{in } \Omega, \\ u = g & \text{on } \partial\Omega.\end{cases} \tag{6.1}$$

The Monge–Ampère equation finds relevance in several areas of mathematics and its applications, including differential geometry, calculus of variations, economics, meteorology and the optimal mass transportation problem (Gutiérrez 2001, Trudinger and Wang 2005, Villani 2003, Benamou and Brenier 1998, Benamou and Brenier 2000, Stojanovic 2005). As we have already mentioned (see also Remark 6.12 below), the Monge–Ampère problem fits into the framework of the previous section. However, due to its many applications, and correspondingly its many discretizations and convergence results, we find it appropriate to devote an entire section to a sample of detailed results. Before discussing these methods, let us give a brief synopsis of some discretizations for the Monge–Ampère problem that we do not cover in detail.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

- *Monotone, wide-stencil finite difference schemes* (see Oberman 2008, Froese and Oberman 2011, Froese 2016 and Benamou *et al.* 2014). The basis of these methods is to use determinant identities (such as Hadamard's inequality) to construct consistent and monotone finite difference schemes for the Monge–Ampère problem. Such schemes conform to the Barles–Souganidis theory under certain assumptions and convergence to the viscosity solution is immediate. We refer the reader to Nochetto and Zhang (2017*b*), who have recently derived rates of convergence for these methods.

- *Regularization.* Feng and Neilan (2009, 2011) and Neilan (2010) first approximate the Monge–Ampère equation at the continuous level by a fourth-order quasi-linear PDE. In this framework, solutions are defined via variational principles, and therefore Galerkin methods are readily constructed. Error estimates between the finite element approximation and the regularized solution are derived, but a general convergence theory for viscosity solutions is an open problem.

- *Variational methods.* Classical Galerkin methodologies are used by Böhmer (2008), Brenner, Gudi, Neilan and Sung (2011), Brenner and Neilan (2012), Neilan (2013) and Awanou (2015*a*, 2015*b*, 2015*c*), for example, to construct finite element approximations for strong or classical solutions of the Monge–Ampère equation. Assuming that the solution to (6.1) is sufficiently regular, error estimates are typically derived in H^1- or H^2-type norms.

- *Optimal control.* Dean and Glowinski (2006*a*, 2006*b*) treat the Monge–Ampère equation as a constraint of a minimization problem, leading to a saddle-point reformulation. Computationally efficient algorithms are proposed, but convergence to the viscosity solution is not shown.

We refer to the review paper by Feng, Glowinski and Neilan (2013) for a summary of these discretization types.

6.1. Solution concepts

The Monge–Ampère equation has a very unique structure and, because of this, the solution to problem (6.1) can be understood not only with the concepts described in Section 2 but with some additional ones. To discuss these issues, we first make the (sometimes necessary) assumption that the domain Ω is convex. With this assumption, and with an abuse of notation, we (re)define the convex envelope of a function v with $v = 0$ on $\partial\Omega$ as

$$\Gamma(v)(x) := \sup\{L(x) : L(z) \leq v(z)\ \forall z \in \Omega,\ L \in \mathbb{P}_1\}.$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

As we previously described, in contrast to the problems discussed thus far, the Monge–Ampère operator $F(x, r, \mathbf{p}, M) := \det(M) - f(x)$ is *not* elliptic on $\Omega \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{S}^d$ (see Definition 2.10 and Example 2.14). Indeed, we have

$$\det(M) \leq \det(N), \quad \text{for all } N \geq M$$

if and only if $M \geq 0$, and consequently F is only elliptic when restricted to the class of semi-positive definite matrices. In particular, if a classical solution u to (6.1) exists, then F is elliptic at u if and only if $D^2u \geq 0$ in Ω, implying that u is a convex function. Further note that, besides ensuring ellipticity, the convexity assumption is required to have any hope of uniqueness, since, when d is even,

$$\det(D^2(-u)) = \det(D^2u).$$

Let us now discuss the existence and uniqueness of different types of solutions of the Monge–Ampère problem (6.1) and the required conditions on the domain Ω and data, f and g, to guarantee that such solutions exist. First we have the following result regarding classical solutions (Wang 1996, Trudinger and Wang 2008).

Theorem 6.1 (classical solutions). Suppose that Ω is an open, bounded and strictly convex domain with boundary $\partial\Omega \in C^3$. Suppose that $g \in C^3(\bar{\Omega})$, $f \in C^\alpha(\Omega)$ for some $\alpha \in (0, 1)$, and that $\min_{\bar{\Omega}} f > 0$. Then there exists a unique (classical) convex solution $u \in C^{2,\alpha}(\Omega)$ satisfying (6.1).

Remark 6.2 (optimality). The conditions on $\partial\Omega$, g and f stated in Theorem 6.1 are optimal (Trudinger and Wang 2008).

Next we turn our attention to viscosity solutions to (6.1). Since the operator F is only elliptic on the class of convex functions, we must modify the definition of viscosity solutions to reflect this fact.

Definition 6.3 (viscosity solution). A function $u \in C(\bar{\Omega})$ is called a viscosity subsolution (resp. supersolution) of (6.1) on the set of convex functions if u is convex and if, for all convex $\varphi \in C^2(\Omega)$ such that $u - \varphi$ has a local maximum (resp. minimum) at $x_0 \in \Omega$, we have $\det(D^2\varphi(x_0)) \geq f(x_0)$ (resp. $\det(D^2\varphi(x_0)) \leq f(x_0)$). A function $u \in C(\bar{\Omega})$ is a viscosity solution to (6.1) on the class of convex functions if it is simultaneously a viscosity subsolution and supersolution on the set of convex functions.

Remark 6.4 (f is non-negative). Note that the definition of viscosity solution implicitly requires the source function f to be continuous and non-negative.

Before stating existence and uniqueness results for viscosity solutions, let us give an example of a non-classical viscosity solution, following Froese and Oberman (2011). Note that this example shows that $f \in C^\alpha(\Omega)$ and smooth Ω is not sufficient to guarantee the existence of a classical solution.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Example 6.5 (viscosity solution). Let $\Omega = B_2(0) \subset \mathbb{R}^2$, and consider the function

$$u(x) = \begin{cases} 0 & \text{in } |x| \leq 1, \\ \frac{1}{2}(|x| - 1)^2 & \text{in } 1 \leq |x| \leq 2. \end{cases} \tag{6.2}$$

Then u is a viscosity solution of (6.1) with $g = 1/2$ and

$$f(x) = \begin{cases} 0 & \text{in } |x| \leq 1, \\ 1 - |x|^{-1} & \text{in } 1 \leq |x| \leq 2. \end{cases}$$

Theorem 6.6 (existence). Assume that Ω is convex. Assume further that $g \in C(\partial\Omega)$, $f \in C(\bar{\Omega})$ and $f \geq 0$. Then there exists a unique viscosity solution $u \in C(\bar{\Omega})$ to (6.1) in the class of convex functions.

Apart from the viscosity solution, another type of weak solution of the Monge–Ampère equation is based on geometric arguments, which we now describe. To motivate it, suppose for the moment that f is uniformly positive in Ω and that $u \in C^2(\bar{\Omega})$ is a strictly convex function satisfying (6.1) pointwise. Let $\partial u(x)$ denote the subdifferential of u at the point $x \in \Omega$ given by Definition 2.58, that is, $\partial u(x)$ is the set of slopes of the supporting hyperplanes at x. From the convexity and smoothness assumptions we infer that $\partial u(x) = \{Du(x)\}$ and that the subdifferential, viewed as a map from Ω to $\mathbb{R}^d$, is injective. Consequently, a change of variables reveals that

$$\int_D f = \int_D \det(D^2 u) = \int_{\partial u(D)} = |\partial u(D)| \quad \text{for all Borel sets } D \subset \Omega,$$

where we recall that $\partial u(D) = \cup_{x \in D} \partial u(x)$ and $|\partial u(D)|$ is the d-dimensional Lebesgue measure of $\partial u(D)$. Since ∂u is well-defined for non-smooth convex functions, the above identity allows us to widen the class of admissible solutions.

Definition 6.7 (Alexandrov solution). A convex function $u \in C(\bar{\Omega})$ is an *Alexandrov solution* to (6.1) if $u|_{\partial\Omega} = g$ and

$$|\partial u(D)| = \int_D f \tag{6.3}$$

for all Borel sets $D \subset \Omega$.

Note that the continuity of the source term f is no longer required for (6.3) to be well-defined. The following example illustrates this feature.

Example 6.8 (Alexandrov solution). Let $\Omega = B_1(0) \subset \mathbb{R}^2$. Then the function

$$u(x) = |x| - 1$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

is an Alexandrov solution of the Monge–Ampère equation

$$\det(D^2 u)(x) = \pi \delta_{(x=0)},$$

where $\delta_{(x=0)}$ is the Dirac measure at origin. However, u is not a viscosity solution because the right-hand side is not a (continuous) function.

The existence and uniqueness of Alexandrov solutions is summarized in the next theorem: see Gutiérrez (2001, Theorem 1.6.2).

Theorem 6.9 (existence of Alexandrov solutions). Let $\Omega \subset \mathbb{R}^d$ be an open, bounded and strictly convex domain. Suppose that the data satisfies $g \in C(\partial\Omega)$ and $\int_\Omega f < \infty$. Then there exists a unique Alexandrov solution $u \in C(\bar{\Omega})$ to (6.1) in the class of convex functions.

The next result states situations where viscosity solutions and Alexandrov solutions coincide (Gutiérrez 2001, Propositions 1.3.4 and 1.7.1). The result also shows that the notion of Alexandrov solution is strictly weaker than that of a viscosity solution.

Proposition 6.10 (equivalence). Suppose that $u \in C(\bar{\Omega})$ is an Alexandrov solution to (6.1). Then if f is continuous, u is a viscosity solution to (6.1). Conversely, if u is a viscosity solution to (6.1) and if $f \in C(\bar{\Omega})$ with $f > 0$ in $\bar{\Omega}$, then u is an Alexandrov solution.

The rest of our presentation will deal separately with numerical methods for (6.1) depending on the type of solution we aim to approximate: viscosity or Alexandrov solutions. Before we discuss them, it is useful to review a property of the determinant. For convenience of notation we define the set of symmetric non-negative definite matrices as

$$\mathbb{S}^d_+ := \{A \in \mathbb{S}^d : A \geq 0\}.$$

Proposition 6.11 (determinant inequality). The following inequality holds:

$$\det(A)^{1/d} \det(B)^{1/d} \leq \frac{1}{d} \operatorname{tr} AB = \frac{1}{d}(A : B), \quad \text{for all } A, B \in \mathbb{S}^d_+$$

with equality holding if and only if $B^{1/2} A B^{1/2} = cI$ for some scalar $c \geq 0$.

Proof. Let $C = B^{1/2} A B^{1/2}$, and note that $C \in \mathbb{S}^d_+$. The arithmetic–geometric inequality yields

$$\det(C)^{1/d} \leq \frac{1}{d} \operatorname{tr} C,$$

with equality holding if and only if $B^{1/2} A B^{1/2} = C = cI$. Since $A, B \in \mathbb{S}^d_+$, we find that $c \geq 0$. The inequality follows from the algebraic identities $\det C = \det(AB) = \det A \det B$ and $\operatorname{tr} C = \operatorname{tr} AB = A : B$. □

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Remark 6.12 (Monge–Ampère is concave). The determinant inequality of Proposition 6.11 implies that for any positive definite matrix A,

$$\det(A)^{1/d} = \inf_{B \in \mathbb{S}^d_+} \left\{ \frac{1}{d} \operatorname{tr} AB : \det B = 1 \right\}.$$

Therefore, for any symmetric positive definite A and $B \in \mathbb{S}^d_+$, and for $0 < t \leq 1$, we have

$$\begin{aligned} \det(tA + (1-t)B)^{1/d} &= \inf_{K \in \mathbb{S}^d_+} \left\{ \frac{1}{d} \operatorname{tr}((tA + (1-t)B)K) : \det K = 1 \right\} \\ &\geq t \inf_{K \in \mathbb{S}^d_+} \left\{ \frac{1}{d} \operatorname{tr}(AK) : \det K = 1 \right\} \\ &\qquad + (1-t) \inf_{K \in \mathbb{S}^d_+} \left\{ \frac{1}{d} \operatorname{tr}(BK) : \det K = 1 \right\} \\ &= t \det(A)^{1/d} + (1-t) \det(B)^{1/d}. \end{aligned}$$

The inequality

$$\det(tA + (1-t)B)^{1/d} \geq t \det(A)^{1/d} + (1-t) \det(B)^{1/d}$$

is also satisfied if $t = 0$, and if $\det(A) = \det(B) = 0$. Thus the function $A \to \det A^{1/d}$ is concave over $\mathbb{S}^d_+$.

6.2. Approximation of viscosity solutions: Hamilton–Jacobi–Bellman reformulation

In this section we summarize the results of Feng and Jensen (2017), where a numerical method based on an HJB reformation of the Monge–Ampère problem is developed. Before stating the method we first note that Remark 6.12 implicitly gives such an HJB reformation, that is, if u satisfies the first equation in (6.1), then formally

$$f(x)^{1/d} = (\det(D^2 u)(x))^{1/d} = \frac{1}{d} \inf_{\substack{B \in \mathbb{S}^d_+ \\ \det(B)=1}} D^2 u(x) : B.$$

Note that the constraint in the control set is nonlinear, and furthermore, the control set is not compact. These two features make the PDE and numerical constructions less obvious. Instead, the method proposed in Feng and Jensen (2017) is based on the following result (Krylov 1987).

Proposition 6.13 (HJB reformulation of MA). Define

$$\mathbb{S}^d_1 = \{A \in \mathbb{S}^d_+ : \operatorname{tr} A = 1\},$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

and let $f \in \mathbb{R}$ with $f \geq 0$. Then a matrix $A \in \mathbb{S}^d$ satisfies

$$H(A, f) := \sup_{B \in \mathbb{S}^d_1} \left(-\frac{1}{d} B : A + f^{1/d} \det(B)^{1/d} \right) = 0 \tag{6.4}$$

if and only if $\det(A) = f$ and $A \in \mathbb{S}^d_+$. Moreover, there exists a maximizer $B_* \in \mathbb{S}^d_1$ that commutes with A.

Proof. If $A \in \mathbb{S}^d_+$ and $\det(A) = f$, then Proposition 6.11 implies

$$-\frac{1}{d} B : A + f^{1/d} \det(B)^{1/d} \leq (f^{1/d} - \det(A)^{1/d}) \det(B)^{1/d} = 0$$

and equality holds if and only if $B^{1/2} A B^{1/2} = cI$ for some $c \geq 0$. If $A > 0$ then we take $B = A^{-1}/(\operatorname{tr} A^{-1})$. Otherwise, if $\det(A) = 0$, we use the eigendecomposition of a matrix to construct $B \in \mathbb{S}^d_+$ that commutes with A, $\operatorname{tr} B = 1$, and $B^{1/2} A B^{1/2} = 0$. In either case we conclude that A satisfies (6.4).

Now our goal is to show that (6.4) implies that $\det(A) = f$ and $A \in \mathbb{S}^d_+$. We prove the statement in two cases.

Case I. If $A > 0$, then take $B = cA^{-1}$ with $c = 1/(\operatorname{tr} A^{-1})$ in (6.4). Using the identities

$$\frac{1}{d} B : A = c \quad \text{and} \quad \det(B) = c^d \det(A)^{-1},$$

we find that

$$c(f^{1/d} - \det(A)^{1/d}) \det(A)^{-1/d} = \frac{-1}{d} B : A + f^{1/d} \det(B)^{1/d} \leq 0$$

and therefore $\det(A) \geq f$. On the other hand Proposition 6.11 implies that

$$0 = \sup_{B \in \mathbb{S}^d_1} \left(-\frac{1}{d} B : A + f^{1/d} \det(B)^{1/d} \right) \leq \sup_{B \in \mathbb{S}^d_1} (f^{1/d} - \det(A)^{1/d}) \det(B)^{1/d}.$$

Now if $\det(A) > f$, then we must have $\det(B) = 0$ for the matrix B to attain the supremum. However, since $A > 0$, Proposition 6.11 implies that the inequality above is strict if $B \neq cA^{-1}$. This leads to a contradiction. Therefore $\det(A) = f$. This proves the first case.

Case II. If A is not strictly positive definite, then without loss of generality, we assume that A is of the diagonal form

$$A = \lambda_1 \mathbf{v}_1 \otimes \mathbf{v}_1 + \cdots + \lambda_d \mathbf{v}_d \otimes \mathbf{v}_d \quad \lambda_1 \leq \cdots \leq \lambda_d,$$

where $\{\mathbf{v}_i\}_{i=1}^d$ are the eigenvectors of A with unit length, and $\lambda_1 \leq 0$.

Consider the matrix

$$B = (1 - \epsilon) \mathbf{v}_1 \otimes \mathbf{v}_1 + \frac{\epsilon}{d-1} \sum_{i=2}^{d} \mathbf{v}_i \otimes \mathbf{v}_i \in \mathbb{S}^d_1,$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

for some parameter $\epsilon \in (0,1)$. We then have

$$\begin{aligned}-\frac{1}{d}B : A + f^{1/d}\det(B)^{1/d} &= -\frac{1}{d}\left((1-\epsilon)\lambda_1 + \frac{\epsilon}{d-1}\sum_{i=2}^{d}\lambda_i\right)\\ &\quad + (1-\epsilon)^{1/d}\frac{\epsilon^{(d-1)/d}}{(d-1)^{(d-1)/d}}f^{1/d}\\ &\geq \frac{-1}{d}((1-\epsilon)\lambda_1 + \epsilon\lambda_d) + \left[\frac{(1-\epsilon)\epsilon^{(d-1)}}{(d-1)^{(d-1)}}f\right]^{1/d}.\end{aligned}$$

Note that if $f > 0$ and for sufficiently small ϵ, the dominating term is

$$\left[\frac{(1-\epsilon)\epsilon^{(d-1)}}{(d-1)^{(d-1)}}f\right]^{1/d}.$$

Thus, if $f > 0$, then we deduce

$$-\frac{1}{d}B : A + f^{1/d}\det(B)^{1/d} > 0,$$

which contradicts (6.4). If $f = 0$, then

$$-\frac{1}{d}B : A + f^{1/d}\det(B)^{1/d} = -\frac{1}{d}\left((1-\epsilon)\lambda_1 + \frac{\epsilon}{d-1}\sum_{i=2}^{d}\lambda_i\right).$$

Equation (6.4) holds for any $\epsilon > 0$ if and only if $\lambda_1 = 0$. Thus, $A \geq 0$ and $\det(A) = f = 0$. This completes the proof. □

Lemma 6.13 leads to the following result, showing that viscosity solutions to the Monge–Ampère equation are viscosity solutions of an HJB problem. We refer the reader to Feng and Jensen (2017, Theorem 3.3) for a proof.

Theorem 6.14 (equivalence of viscosity solutions). Let $f \in C(\Omega)$ be a non-negative function, and let u be a viscosity subsolution (resp. supersolution) of the Monge–Ampère problem (6.1) on the set of convex functions. Then u is a viscosity subsolution (resp. supersolution) to the HJB problem

$$\begin{cases} H(D^2u, f) = 0 & \text{in } \Omega,\\ u = g & \text{on } \partial\Omega. \end{cases} \tag{6.5}$$

Remark 6.15 (convexity of solution). We emphasize that the convexity of the solution to (6.5) is not assumed *a priori*; it occurs implicitly from the structure of the HJB problem.

Remark 6.16 ($\mathbb{S}_1^d$ is not Cordes). It is worth mentioning that matrices in $\mathbb{S}_1^d$ do not necessarily satisfy the Cordes condition if $d \geq 3$ (see Definition 2.32). To see this, consider $B \in \mathbb{S}_1^3$ with eigenvalues $\lambda_1 = 1 - \tau$,

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

$\lambda_2 = \lambda_3 = \tau/2$ for some $\tau \in (0, 1/3]$. We then find

$$\frac{|B|^2}{(\operatorname{tr} B)^2} = \lambda_1^2 + \lambda_2^2 + \lambda_3^2 = \frac{3}{2}\tau^2 - 2\tau + 1 \geq \frac{1}{2},$$

and thus B does not satisfy the Cordes condition.

6.2.1. Discretization of the HJB-reformulation

To describe a discretization method for problem (6.5), we first note that if $B \in \mathbb{S}_1^d$, then $B = YDY^\intercal$, where $D \in \mathbb{S}_1^d$ is a diagonal matrix and $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_d] \in \mathbb{SO}^d$, where $\mathbb{SO}^d$ is the special orthogonal group. Thus, letting $\mathbb{D}_1^d$ denote the space of diagonal non-negative definite matrices with unit trace, we find that

$$H(A, f) = \sup_{\substack{Y \in \mathbb{SO}^d \\ D \in \mathbb{D}_1^d}} \left(-\frac{1}{d} YDY^\intercal : A + f^{1/d} \det(D)^{1/d} \right).$$

Let $\mathbf{y}_i \in \mathbb{R}^d$ denote the ith column of $Y \in \mathbb{SO}^d$ and let $\lambda_i = D_{i,i}$. Then the solution to the Monge–Ampèreequation satisfies

$$\sup_{\substack{Y \in \mathbb{SO}^d \\ D \in \mathbb{D}_1^d}} \left(-\frac{1}{d} \sum_{i=1}^d \lambda_i \mathbf{y}_i \otimes \mathbf{y}_i : D^2 u(x) + f(x)^{1/d} \left(\prod_{i=1}^d \lambda_i \right)^{1/d} \right) = 0$$

in the viscosity sense.

Let $\mathcal{T}_h$ be a quasi-uniform and shape-regular mesh of Ω and let X_h^l be the continuous piecewise linear polynomials over $\mathcal{T}_h$ defined by (3.34). We let Ω_h^I denote the set of interior vertices of $\mathcal{T}_h$, and let Ω_h^B be the set of boundary vertices. Let $v_h \in X_h^l$. Recall that, for a discretization parameter $k > 0$ and vector y, the second-order difference operator is defined by

$$\delta_{y,k}^2 v_h(z) = \frac{1}{k^2}(v_h(z + ky) - 2v_h(z) + v_h(z - ky))$$

provided that z is sufficiently far from the boundary, namely, $z \pm ky \in \bar{\Omega}$. Otherwise we set

$$\delta_{y,k}^2 v_h(z) = \frac{2}{k^+ + k^-} \left(\frac{v_h(z + k^+ y) - v_h(z)}{k^+} + \frac{v_h(z - k^- y) - v_h(z)}{k^-} \right),$$

where $k^\pm$ is the only element in $(0, k]$ such that $z \pm k^\pm y \in \partial\Omega$.

With these definitions at hand we introduce, on X_h^l, the following approximation of the HJB reformulation:

$$H_h^k[u_h, f](z) = \sup_{\substack{Y \in \mathbb{SO}^d \\ D \in \mathbb{D}_1^d}} \left(\frac{-1}{d} \sum_{i=1}^d \lambda_i \delta_{y,k}^2 u_h(z) + f(z)^{1/d} \det(D)^{1/d} \right),$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

where $z \in \Omega_h^I$.

The numerical method based on the HJB reformulation seeks a piecewise linear function $u_h \in X_h^l$ such that

$$\begin{cases} H_h^k[u_h, f] = 0 & \text{in } \Omega_h^I, \\ u_h = g & \text{on } \Omega_h^B. \end{cases} \tag{6.6}$$

Note that the boundary conditions uniquely determine u_h on $\partial\Omega$, and therefore the problem is well-defined. The scheme (6.6) is a semi-Lagrangian method because the points $z \pm ky$, which are used to evaluate $\delta_{y,h}^2 u_h(z)$, may not belong to $\bar{\Omega}_h$. Thus, additional effort is needed to evaluate $u_h(z \pm ky)$.

It is straightforward to check that $\delta_{y,k}^2 u_h(z)$ is monotone for any direction y, implying that operator $H_h^k[u_h, f]$ is monotone. The monotonicity also leads to stability of the method (Feng and Jensen 2017, Lemmas 6.2 and 6.4).

Lemma 6.17 (stability). Problem (6.6) is stable in the sense that there exists a unique solution $u_h \in X_h^l$ to (6.6) and an h-independent constant $C > 0$ such that $\|u_h\|_{L^\infty(\Omega)} \le C$. Moreover, if we set

$$\bar{u}(x) := \limsup_{\substack{y \to x \\ h \to 0^+}} u_h(y), \quad \underline{u}(x) := \liminf_{\substack{y \to x \\ h \to 0^+}} u_h(y), \quad x \in \bar{\Omega},$$

then $\bar{u}(x) = \underline{u}(x) = g(x)$ for all $x \in \partial\Omega$ provided that Ω is strictly convex.

To show the consistency of the method, we recall that $I_h^{fe} : C(\bar{\Omega}) \to X_h^l$ denotes the nodal interpolant. We then have the following result.

Lemma 6.18 (consistency). Let $\phi \in C^{2,\alpha}(\Omega)$. Then there is a constant C such that for every $z \in \Omega_h^I$ we have

$$\left| H(D^2\phi(z), f(z)) - H_h^k[I_h^{fe}\phi, f](z) \right| \le C\left(k^\alpha + \frac{h^2}{k_{\min}^2} \right),$$

where $k_{\min} = \min\{k^+, k^-\}$. Consequently, the method is consistent if $k \to 0$ and $h/k_{\min} \to 0$.

Proof. It suffices to show that

$$|\delta_{y,k}^2 I_h^{fe}\phi(z) - (y \otimes y) : D^2\phi(z)| \le C\left(k^\alpha + \frac{h^2}{k_{\min}^2} \right).$$

First, by Taylor's theorem, we have

$$|\delta_{y,k}^2 \phi(z) - (y \otimes y) : D^2\phi(z)| \le Ck^\alpha \|\phi\|_{C^{2,\alpha}(\Omega)}. \tag{6.7}$$

Recalling that

$$\|\phi - I_h^{fe}\phi\|_{L^\infty(\Omega)} \le Ch^2 \|\phi\|_{W^{2,\infty}(\Omega)}$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

(Ciarlet 2002), we have

$$|\delta^2_{y,k}\phi(z) - \delta^2_{y,k} I_h^{fe}\phi(z)| \leq C\frac{h^2}{k^2_{\min}}\|\phi\|_{W^{2,\infty}(\Omega)}. \tag{6.8}$$

The desired result now follows from (6.7)–(6.8) and the triangle inequality. □

Remark 6.19 (discrete controls). To implement the method (6.6), it remains to specify a discrete set $\mathbb{S}^d_{1,h} \subset \mathbb{S}^d_1$ of symmetric positive definite matrices with unit trace. To ensure consistency of the method, we require the discrete set $\mathbb{S}^d_{1,h}$ to be dense as $h \to 0$, that is, for any $B \in \mathbb{S}^d_1$, there exists $B_h \in \mathbb{S}^d_{1,h}$ such that $B_h \to B$ as $h \to 0$.

Since the method is monotone and consistent, the convergence of the numerical solution now follows from the Barles–Souganidis theory; see Theorem 3.8 and Feng and Jensen (2017, Theorem 6.5).

Theorem 6.20 (convergence). Let $\Omega \subset \mathbb{R}^d$ be a strictly convex domain. Assume that $f \in C(\Omega)$ with $f \geq 0$ and $g \in C(\partial\Omega)$. Then as $h \to 0$, $h/k_{\min} \to 0$, the solutions $u_h \in X^l_h$ of (6.6) converge uniformly to the unique viscosity solution on the set of convex functions of the Monge–Ampère problem (6.1).

We conclude the discussion on semi-Lagrangian schemes by commenting that rates of convergence for a general semi-Lagrangian scheme for (5.1) have been obtained by Debrabant and Jakobsen (2013, Corollary 7.3).

6.3. Approximation of Alexandrov solutions

Now we discuss numerical methods based on the Alexandrov solution concept presented in Definition 6.7. Essentially, numerical methods of this class are finite-dimensional analogues of (6.3).

Let $\{\omega_i\}_{i=1}^N$ be an open, disjoint partition of the domain, that is, $\omega_i \cap \omega_j = \emptyset$ for $i \neq j$ and $\bar{\Omega} = \cup_{i=1}^N \bar{\omega}_i$. Let $\Omega^I_h := \{z_i\}_{i=1}^N$ be a collection of points with the property that $z_i \in \omega_j$ if and only if $i = j$, and let $\Omega^B_h := \{z_i\}_{i=N+1}^{M+N}$ be a set of distinct points on $\partial\Omega$. As before, we let $\bar{\Omega}_h = \Omega^I_h \cup \Omega^B_h$ and we will call its elements nodes or grid points.

Recall that for a nodal function $v_h \in X^{fd}_h$, its subdifferential at a grid point $z \in \bar{\Omega}_h$ is given by (3.20). Now, a natural generalization of (6.3), and the discrete problem proposed in Oliker and Prussner (1988) reads as follows: find a convex nodal function $u_h \in X^{fd}_h$ satisfying

$$\begin{cases} |\partial u_h(z_i)| = \displaystyle\int_{\omega_i} f & \text{for all } z_i \in \Omega^I_h, \\ u_h(z_i) = g(z_i) & \text{for all } z_i \in \Omega^B_h. \end{cases} \tag{6.9}$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Note that since the partition $\{\omega_i\}_{i=1}^N$ is non-overlapping, for all Borel sets $D \subset \Omega$ we have

$$|\partial u_h(D)| = \sum_{z_i \in D} f_i, \quad f_i = \int_{\omega_i} f.$$

Thus, the scheme is obtained by replacing f in (6.3) with a family of Dirac measures supported at the nodes and by replacing g with its nodal interpolant on the boundary.

One special case of this method is when the interior nodal set is a lattice, that is, for some basis $\{\tilde{\mathbf{e}}_j\}_{j=1}^d$ of $\mathbb{R}^d$ we have

$$\Omega_h^I = \left\{ z = h \sum_{j=1}^d z^j \tilde{\mathbf{e}}_j : z^j \in \mathbb{Z} \right\} \cap \Omega.$$

We remark that this property applies to interior nodes only. For the boundary nodes, we only require that their spacing is of order h, namely, $\partial\Omega \subset \cup_{z \in \Omega_h^B} B_{h/2}(z)$. In this case, the partition $\{\omega_i\}$ of the domain can be taken as parallelotopes

$$\omega_i = \left\{ z_i + \sum_{j=1}^d h^j \tilde{\mathbf{e}}_j : h^j \in \mathbb{R},\ |h^j| \leq \frac{h}{2} \right\} \cap \Omega. \tag{6.10}$$

The length of the coordinate vectors $\{\tilde{\mathbf{e}}_j\}_{j=1}^d$ is such that the parallelotope ω_i is inside the ball $B_h(z_i)$ centred at z_i and of radius h. Note that, by construction, $\omega_j = z_j - z_i + \omega_i$; consequently, the radius of the largest ball inscribed in ω_i does not depend on i and we denote it by ρ. We define the shape-regularity of the nodal set as

$$\sigma = h\rho^{-1}. \tag{6.11}$$

Remark 6.21 (meshless nodal function). It is worth mentioning again that the solution u_h is only defined at the nodes $\bar{\Omega}_h$. Its convex envelope induces a triangulation of the domain Ω and a piecewise linear function. However, this triangulation is not known *a priori*. See Remark 3.18 and Examples 3.44–3.45.

In view of Example 3.45, if the solution of the Monge–Ampère equation is nearly degenerate, wide stencils are needed to compute the subdifferential. Thus, method (6.9) may be regarded as a wide-stencil finite difference scheme when the solution is nearly degenerate. On the other hand, if the solution to (6.1) is strictly convex, that is, for $0 < \lambda \leq \Lambda$ we have $\lambda I \leq D^2 u \leq \Lambda I$, then the subdifferential of u_h at node z depends only on the values of u_h at the adjacent nodes of z. To make this last statement rigorous, we state a definition.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Definition 6.22 (adjacent set). Let $u_h \in X_h^{fd}$ and $z \in \Omega_h^I$. The *adjacent set* A_z of u_h at z is the collection of nodes $z_i \in \bar{\Omega}_h$ such that there exists a supporting hyperplane ℓ of u_h at z and $\ell(z_i) = u_h(z_i)$.

Note that A_z is the set of nodes of a star of z which is induced by the discrete convex envelope of u_h; see Figure 3.2. In particular, we have that the subdifferential $\partial u_h(z)$ is determined by the values $u(z_i)$ for $z_i \in A_z$.

Let us now estimate the size of A_z.

Lemma 6.23 (size of A_z). Let $v \in C^2(\bar{\Omega})$ be a strictly convex function with $\lambda I \leq D^2 v(x) \leq \Lambda I$ for some constants $0 < \lambda \leq \Lambda$. Assume that $E := \{\tilde{\mathbf{e}}_j\}_{j=1}^d$ is a basis of $\mathbb{R}^d$ such that Ω_h^I is a lattice spanned by E with shape-regularity constant σ defined in (6.11). Let the nodal function $v_h \in X_h^{fd}$ be defined by $v_h = I_h^{fd} v$. Then, if A_z is the adjacent set of v_h at $z \in \Omega_h^I$,

$$A_z \subset B_{Rh}(z),$$

where

$$R \geq \bar{R} := \frac{\Lambda}{\lambda} \sigma^2 \left| \sum_{j=1}^d \tilde{\mathbf{e}}_j \right|^2. \tag{6.12}$$

Proof. Without loss of generality, we may assume that $z = 0$, $v(z) = 0$ and $Dv(z) = \mathbf{0}$. Let $\hat{z} \in \Omega_h^I \cap \partial \operatorname{conv}(A_z)$ and let ω be the parallelotope defined as in (6.10) with centre $z = 0$. By convexity of ω, there is a $c \in (0,1)$ such that $c\hat{z} \in \partial\omega$. Thus, we can express $\hat{z}$ as a multiple of a convex combination of $\{\zeta_j\}_{j=1}^{2^d}$, the vertices of ω. In other words, for $R = 1/c$ we have

$$\hat{z} = R \sum_{j=1}^{2^d} k_j \zeta_j, \quad k_j \geq 0, \quad \sum_{j=1}^{2^d} k_j = 1.$$

This representation shows that $|\hat{z}| \leq Rh$; thus, to obtain the result, it remains to estimate R.

Since $c\hat{z} \in \partial\omega$ we have, using (6.11), that $|c\hat{z}| \geq \rho = h\sigma^{-1}$ which can be rewritten as $|\hat{z}| \geq Rh\sigma^{-1}$. Using that $D^2 v \geq \lambda I$ we estimate

$$v_h(\hat{z}) = v(\hat{z}) \geq \frac{1}{2} \lambda R^2 \sigma^{-2} h^2.$$

Let us now obtain an upper bound for $v_h(\hat{z})$. To do so, let us introduce $\hat{\omega}$ as the (unique) smallest parallelotope with vertices $\{\bar{z}_m\}_{m=1}^{2^d} \subset \Omega_h^I$ such that $z \in \{\bar{z}_m\}_{m=1}^{2^d}$ is a vertex and $c\hat{z} \in \hat{\omega}$. This parallelotope can be thought of as belonging to the *dual mesh*. We now invoke Caratheodory's theorem (Dacorogna 2008, Theorem 2.13) to conclude that there is a subset of $\{\bar{z}_m\}_{m=1}^{2^d}$, of cardinality $d+1$, for which $c\hat{z}$ can be expressed as a convex combination of these vertices. In other words, up to a permutation in

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

$\{1,\dots,2^d\}$, we have

$$\hat{z} = R\sum_{m=1}^{d+1}\alpha_m\bar{z}_m, \quad \alpha_m \geq 0, \quad \sum_{m=1}^{d+1}\alpha_m = 1. \tag{6.13}$$

We now invoke that $\hat{z} \in A_z$. This implies that there is an affine function ℓ that verifies

$$\ell(z) = v_h(z) = 0, \quad \ell(\hat{z}) = v_h(\hat{z}), \quad \ell(\bar{z}_m) \leq v_h(\bar{z}_m),\ m = 1,\dots,d+1.$$

Using representation (6.13) of $\hat{z}$ and that $D^2v \leq \Lambda I$ we then obtain

$$v_h(\hat{z}) = R\sum_{m=1}^{d+1}\alpha_m\ell(\bar{z}_m) \leq \frac{1}{2}\Lambda R\sum_{m=1}^{d+1}\alpha_m|\bar{z}_m|^2.$$

It remains to observe now that

$$\bar{z}_m = z + \sum_{j=1}^{d}\epsilon_j h\tilde{\mathbf{e}}_j$$

with $\epsilon_j \in \{-1,0,1\}$, and therefore

$$|\bar{z}_m| \leq h\left|\sum_{j=1}^{d}\tilde{\mathbf{e}}_j\right|.$$

A combination of the obtained upper and lower bounds for $v_h(\hat{z})$ yields

$$\frac{1}{2}\lambda R^2h^2\sigma^{-2} \leq \frac{1}{2}\Lambda Rh^2\left|\sum_{j=1}^{d}\tilde{\mathbf{e}}_j\right|^2,$$

from which (6.12) follows. □

Remark 6.24 (Cartesian lattice). In the setting of Lemma 6.23, if E is the canonical basis of $\mathbb{R}^d$, then Mirebeau (2015) and Benamou, Collino and Mirebeau (2016) have improved estimate (6.12) to $\bar{R} = (\Lambda/\lambda)\sigma^2$.

If Ω_h^I is a lattice, then we are able to show consistency of the method (6.9) in the following sense.

Lemma 6.25 (consistency). Let $E = \{\tilde{\mathbf{e}}_j\}_{j=1}^d$ be a basis of $\mathbb{R}^d$ and Ω_h^I be a lattice spanned by E. Let p be a strictly convex quadratic polynomial with $\lambda I \leq D^2p \leq \Lambda I$. If $z \in \Omega_h^I$ is such that $\operatorname{dist}(z,\partial\Omega) \geq \bar{R}h$, with $\bar{R}$ as in (6.12), then we have

$$|\partial I_h^{fd}p(z)| = \det(D^2p)|\omega_z|, \tag{6.14}$$

where ω_z is the parallelotope defined by (6.10).

Proof. We divide the proof into two steps.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Step 1. We first show that (6.14) holds when the domain is $\Omega = \mathbb{R}^d$. Without loss of generality, we assume that $z = 0$ and $p(x) = \frac{1}{2}x \cdot Mx$ for some $M \geq \lambda I$. For a vector $\mathbf{q} \in \mathbb{R}^d$ we define the norm $|\mathbf{q}|_M^2 := \mathbf{q} \cdot M\mathbf{q}$, and define the set

$$\begin{aligned} V &:= \left\{\mathbf{q} \in \mathbb{R}^d : |\mathbf{q}|_M \leq |\mathbf{q} - z_j|_M,\ \forall z_j \in \bar{\Omega}_h\right\} \\ &= \left\{\mathbf{q} \in \mathbb{R}^d : \mathbf{q} \cdot Mz_j \leq \frac{1}{2} z_j \cdot Mz_j,\ \forall z_j \in \bar{\Omega}_h\right\}. \end{aligned}$$

It can be shown (see Lemma 2.3 of Mirebeau 2015 for details) that translations of V tile $\mathbb{R}^d$ and that $|V| = |\omega_z|$. Moreover, by a simple algebraic manipulation,

$$V = \left\{M^{-1}\mathbf{q} \in \mathbb{R}^d : \mathbf{q} \cdot z_j \leq I_h^{fd} p(z_j),\ \forall z_j \in \bar{\Omega}_h\right\}.$$

Thus, $V = M^{-1}[\partial I_h^{fd} p(0)]$, that is, it is the image of subdifferential $\partial I_h^{fd} p(0)$ under the linear map M^{-1}. Taking measures on both sides yields

$$|\omega_z| = |V| = \det(M)^{-1} |\partial I_h^{fd} p(0)|.$$

The proof of Step 1 is now completed by rearranging terms.

Step 2. We now consider a bounded domain and show that (6.14) holds for nodes sufficiently far away from the boundary, that is, $\operatorname{dist}(z, \partial\Omega) \geq \bar{R}h$. To do so, we observe that the subdifferential $\partial I_h^{fd} p(z)$ is determined only by the function values in the adjacent set A_z. Since, as shown in Lemma 6.23, $A_z \subset B_{\bar{R}h}(z)$, we deduce that if the node z is bounded away from the boundary with $\operatorname{dist}(z, \partial\Omega) \geq \bar{R}h$, then $A_z \subset \Omega$. Thus, (6.14) holds.

This concludes the proof. □

6.3.1. A truncated version

When Ω_h^I is a Cartesian lattice, scheme (6.9) is is closely related to the finite difference method of Mirebeau (2015) and Benamou *et al.* (2016) which we now describe. For simplicity, suppose that $d = 2$ and that the interior grid points are given by

$$\Omega_h^I = \Omega \cap \mathbb{Z}_h^2.$$

Let $S \subset \mathbb{Z}^2 \backslash \{0\}$ denote a stencil. For any $y \in S$ and $z \in \Omega_h^I$ sufficiently far from $\partial\Omega$, we recall that the second-order difference operator in the direction y is given by

$$\delta_{y,h}^2 v(z) = \frac{v(z + hy) - 2v(z) + v(z - hy)}{h^2}.$$

When $z \in \Omega_h^I$ is close to $\partial\Omega$, the point $z \pm hy$ may not belong to $\bar{\Omega}_h$. In

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

such cases, we define

$$\delta^2_{y,h} v(z) := \frac{2}{h^+ + h^-}\left(\frac{v(z+h^+y)-v(z)}{h^+} + \frac{v(z-h^-y)-v(z)}{h^-}\right),$$

where $h^\pm$ is the only element in $[0,h]$ such that $z \pm h^\pm y \in \partial\Omega$. This construction implicitly defines the set of boundary points Ω_h^B.

We define the set of *superbases* of S as

$$Y_h := \big\{(y_0,y_1,y_2) \in S^3 : |\det(y_0,y_1,y_2)| = 1,\ y_0+y_1+y_2 = \mathbf{0}\big\}.$$

Note that for $z \in \Omega_h^I$ and $\mathbf{y} = (y_0,y_1,y_2) \in Y_h$, the convex hull $\mathcal{H}_{z,\mathbf{y}} = \operatorname{conv}\{z \pm hy_i\}_{i=0}^3$ is a hexagon. Given a nodal function v_h, superbasis $\mathbf{y} = (y_0,y_1,y_2) \in Y_h$ and a point $z \in \Omega_h^I$, we let $\Gamma_{z,\mathbf{y}}(v_h)$ denote the maximal convex map bounded above by v_h at the points z and $\{z \pm hy_i\}_{i=0}^3$. As before, $\Gamma_{z,\mathbf{y}}(v_h)$, restricted to $\mathcal{H}_{z,\mathbf{y}}$, is a piecewise linear function with respect to some triangulation of $\mathcal{H}_{z,\mathbf{y}}$, which depends on the values of v_h on the extreme points of $\mathcal{H}_{z,\mathbf{y}}$ and z.

We define the discrete Monge–Ampère operator

$$\gamma(\Delta^+_{y_0} v_h(z), \Delta^+_{y_1} v_h(z), \Delta^+_{y_2} v_h(z)),$$

with $\Delta^+_y v(z) = \max\{\delta^2_{y,h} v_h(z), 0\}$ and

$$\gamma(\delta_0,\delta_1,\delta_2) := \begin{cases} \delta_{i+1}\delta_{i+2} & \text{if } \delta_i \geq \delta_{i+1} + \delta_{i+2},\ i = 0,\dots,2 \mod 3, \\ \gamma_1(\delta_0,\delta_1,\delta_2) & \text{otherwise,} \end{cases}$$

with

$$\gamma_1(\delta_0,\delta_1,\delta_2) := \frac{1}{2}(\delta_0\delta_1 + \delta_1\delta_2 + \delta_0\delta_2) - \frac{1}{4}(\delta_0^2 + \delta_1^2 + \delta_2^2).$$

As shown by Benamou *et al.* (2016, Remark 1.10), from the definition of subdifferential and some geometric arguments it follows that

$$\gamma(\Delta^+_{y_0} v_h(z), \Delta^+_{y_1} v_h(z), \Delta^+_{y_2} v_h(z)) = h^2 |\partial\Gamma_{z,\mathbf{y}}(v_h)(z)|.$$

The scheme proposed in Benamou *et al.* (2016) reads as follows: find the nodal function $u_h \in X_h^{fd}$ such that

$$\begin{cases} \min\limits_{\mathbf{y}\in Y_h} \gamma(\Delta^+_{y_0} u_h(z), \Delta^+_{y_1} u_h(z), \Delta^+_{y_2} u_h(z)) = f(z) & \text{for all } z \in \Omega_h^I, \\ u_h(z) = g(z) & \text{for all } z \in \Omega_h^B. \end{cases} \tag{6.15}$$

Lemma 6.26 (consistency). Let M be a positive definite matrix and $p(x) = \frac{1}{2} x \cdot Mx$ be a convex quadratic polynomial. Then

$$\min_{\mathbf{y}\in Y_h} |\partial\Gamma_{z,\mathbf{y}}(I_h^{fd} p)(z)| = \det(M)$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

if and only if there is an M-obtuse superbasis (y_0, y_1, y_2), that is,

$$y_i \cdot My_j \leq 0, \quad \text{for all } i \neq j.$$

Moreover, if $B_R \subset \operatorname{conv} S$ with $R^2 = 2|M|\,|M^{-1}|$, then such an M-obtuse basis exists in S.

We refer to Benamou *et al.* (2016, Propositions 1.12 and 2.2) for a proof. Note that the previous result shows that if the matrix M is anisotropic, that is, $|M|\,|M^{-1}|$ is large, then a wide stencil is needed to ensure the existence of an M-obtuse superbasis.

Remark 6.27 (three dimensions). In three space dimensions, to the best of our knowledge, there is no explicit formula to compute $|\partial\Gamma_{z,\mathbf{y}}(v_h)(z)|$. As shown in Mirebeau (2015), if $p(x) = \frac{1}{2}x \cdot Mx$ and the stencil S is such that $B_R \subset \operatorname{conv} S$, for some R that depends on $|M|\,|M^{-1}|$, then we have that

$$\det(M) = h^3|\partial\Gamma_{z,\mathbf{y}}(I_h^{fd}p)(z)|.$$

This result is consistent with Lemma 6.25.

6.3.2. Stability of (6.9)

While the convergence of monotone and consistent schemes, like (6.9), can be obtained using the framework described in Section 3.1, few results are known on the rate of convergence of such approximations. Here, following Nochetto and Zhang (2017*b*), we discuss some recent results on the L^∞-rate of convergence of scheme (6.9).

In addition to the discrete Alexandrov estimates of Lemmas 3.24 and 3.46, the derivation of these error estimates involves a discrete barrier argument as in Section 4.3 and the Brunn–Minkowski inequality, which we now state.

Let D and E be two non-empty compact subsets of $\mathbb{R}^d$. We define their (Minkowski) sum as

$$D + E := \{\mathbf{v} + \mathbf{w} \in \mathbb{R}^d : \mathbf{v} \in D,\, \mathbf{w} \in E\}.$$

Proposition 6.28 (Brunn–Minkowski). Let A and B be two non-empty compact subsets of $\mathbb{R}^d$. Then the following inequality holds:

$$|A + B|^{1/d} \geq |A|^{1/d} + |B|^{1/d}$$

where $|\cdot|$ denotes the Lebesgue measure on $\mathbb{R}^d$.

Remark 6.29 (concavity). The Brunn–Minkowski inequality implies that the function $D \to |D|^{1/d}$ is concave, in the sense that for $0 \leq t \leq 1$,

$$|tA + (1-t)B|^{1/d} \geq t|A|^{1/d} + (1-t)|B|^{1/d}.$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

The discrete Alexandrov estimate Lemma 3.24 shows that the L^∞-norm of nodal or piecewise linear function v_h is controlled by the measure of the subdifferential $|\partial v_h|$. Now suppose u_h and w_h are two nodal functions. The following stability estimate shows that the difference $v_h - w_h$ measured in the L^∞-norm is controlled by the difference of the measure of their subdifferentials. This can be recast as a stability estimate for scheme (6.9).

Proposition 6.30 (stability). Let v_h and w_h be two nodal functions with $v_h \geq w_h$ on Ω_h^B. Then

$$\sup_{\bar\Omega_h}(v_h - w_h)^- \leq C\Bigg(\sum_{z\in\mathcal{C}_h^-(v_h-w_h)} \big(|\partial v_h(z)|^{1/d} - |\partial w_h(z)|^{1/d}\big)^d\Bigg)^{1/d}.$$

Proof. Consider the discrete convex envelope of the difference $v_h - w_h$, which we denote by $\Gamma_h(v_h - w_h)$ and its lower nodal contact set

$$\mathcal{C}_h^-(v_h - w_h) = \{z \in \Omega_h^I : \Gamma_h(v_h - w_h)(z) = (u_h - w_h)(z)\}.$$

By Lemma 3.24 (finite difference Alexandrov estimate), we have

$$\sup_{\bar\Omega_h}(v_h - w_h)^- \leq C\Bigg(\sum_{z\in\mathcal{C}_h(v_h-w_h)^-} |\partial\Gamma_h(v_h - w_h)(z)|\Bigg)^{1/d}. \tag{6.16}$$

Thus, we only need to estimate $|\partial\Gamma_h(v_h - w_h)(z)|$ for all z in the contact set, which we do as follows. We first note that Lemma 3.21 implies that

$$\partial w_h(z) + \partial\Gamma_h(v_h - w_h)(z) \subset \partial v_h(z). \tag{6.17}$$

From this, and the Brunn–Minkowski inequality (Proposition 6.28), we obtain

$$\begin{aligned} &|\partial w_h(z)|^{1/d} + |\partial\Gamma_h(v_h - w_h)(z)|^{1/d} \\ &\quad\leq |\partial w_h(z) + \partial\Gamma_h(v_h - w_h)(z)|^{1/d} \leq |\partial v_h(z)|^{1/d}, \end{aligned} \tag{6.18}$$

which clearly implies that

$$|\partial\Gamma_h(v_h - w_h)(z)| \leq \big(|\partial v_h(z)|^{1/d} - |\partial w_h(z)|^{1/d}\big)^d.$$

This is the desired estimate for $|\partial\Gamma_h(u_h - w_h)(z)|$. Inserting it into (6.16) yields the claimed result. □

A direct consequence of this stability result is a maximum principle for nodal functions, which we state below.

Corollary 6.31 (maximum principle). Let v_h and w_h be two nodal functions associated with $\bar\Omega_h$. If $v_h \geq w_h$ on Ω_h^B and $|\partial v_h(z)| \leq |\partial w_h(z)|$ for all $z \in \Omega_h^I$, then

$$w_h(z) \leq v_h(z), \quad \text{for all } z \in \bar\Omega_h.$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Proof. By (6.17) and (6.18), we have $|\partial w_h(z)| \leq |\partial v_h(z)|$ for any $z \in \mathcal{C}_h^-(v_h - w_h)$. Since $|\partial v_h(z)| \leq |\partial w_h(z)|$ by assumption, we have $|\partial v_h(z)| = |\partial w_h(z)|$ at contact points. Thus, by Proposition 6.30, we get

$$\sup_{\bar{\Omega}_h}(v_h - w_h)^- = 0.$$

Consequently $v_h - w_h \geq 0$, which proves the result. □

6.3.3. Error estimates for (6.9)

Let us now derive rates of convergence in the L^∞-norm for method (6.9). To do so, we will use all the tools we have developed in previous sections, namely the discrete Alexandrov estimate of Proposition 3.24, the Brunn–Minkowski inequality of Proposition 6.28 and the stability result of Proposition 6.30.

Owing to Proposition 6.30 we only need to study the consistency error. Since the method is consistent for convex quadratic polynomials at nodes bounded away from the boundary $\partial\Omega$ (see Lemma 6.25), if we expect that u can be well approximated by quadratic polynomials, then the consistency error will also be small. Let us make this intuition rigorous.

Lemma 6.32 (interior consistency). Let $E = \{\tilde{\mathbf{e}}_j\}_{j=1}^d$ be a basis of $\mathbb{R}^d$ and let Ω_h^I be a lattice spanned by E. Given u, strictly convex, let $z \in \Omega_h^I$ with $\operatorname{dist}(z, \partial\Omega) \geq \bar{R}h$, where $\bar{R}$ is defined in (6.12), and set $\bar{B} = B_{\bar{R}h}(z)$. If $u \in C^{2,\alpha}(\bar{B})$, then we have

$$\left| |\partial I_h^{fd} u(z)| - \int_{\omega_z} \det(D^2 u) \right| \leq C h^\alpha |\omega_z|,$$

where the constant C depends on $|u|_{C^{2,\alpha}(\bar{B})}$, and ω_z is defined in (6.10).

Proof. Let us show that

$$|\partial I_h^{fd} u(z)| \leq \int_{\omega_z} \det(D^2 u) + C h^\alpha |\omega_z|.$$

The other inequality can be obtained in a similar fashion.

Since $u \in C^{2,\alpha}(\bar{B})$, there is a convex quadratic polynomial $p \in \mathbb{P}_2$ that satisfies $p(z) = u(z)$, $Dp(z) = Du(z)$, $D^2 p = D^2 u(z)$ and, moreover,

$$u(x) \leq p(x) + |u|_{C^{2,\alpha}(\bar{B})} h^{2+\alpha}, \quad \text{for all } x \in \bar{B}.$$

Define $q(x) = p(x) + h^\alpha |u|_{C^{2,\alpha}(\bar{B})} |x - z|^2$ and note that, by construction, $q(z) = u(z)$ and for all nodes $z_j \in \bar{B} \cap \Omega_h^I$ we have $u(z_j) \leq q(z_j)$. Since q is convex, its nodal interpolant $I_h^{fd} q$ is also convex (see Definition 3.16). Thus, we can apply Lemma 3.20 (monotonicity), to get $|\partial I_h^{fd} u(z)| \leq |\partial I_h^{fd} q(z)|$. From these considerations we see that it is sufficient to show that

$$|\partial I_h^{fd} q(z)| \leq \int_{\omega_z} \det(D^2 u) + C h^\alpha |\omega_z|.$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Since $\lambda + Ch^\alpha \leq D^2 q \leq \Lambda + Ch^\alpha$ and

$$\frac{\Lambda + Ch^\alpha}{\lambda + Ch^\alpha} \leq \frac{\Lambda}{\lambda},$$

we invoke the consistency result of Lemma 6.25 and the regularity $u \in C^{2,\alpha}(\bar{B})$ to obtain

$$\begin{aligned} |\partial I_h^{fd} q(z)| &= \det(D^2 q)|\omega_z| \leq (\det(D^2 p) + Ch^\alpha)|\omega_z| \\ &\leq \int_{\omega_z} \det(D^2 u) + Ch^\alpha |\omega_z|. \end{aligned}$$

This concludes the proof. □

The previous result establishes a consistency error at nodes bounded away from the boundary. For nodes close to the boundary, we have the following estimate.

Lemma 6.33 (boundary consistency). Let $E = \{\tilde{\mathbf{e}}_j\}_{j=1}^d$ be a basis of $\mathbb{R}^d$ and let Ω_h^I be a lattice spanned by E. Given u, strictly convex, let $z \in \Omega_h^I$ satisfy $\operatorname{dist}(z, \partial\Omega) \leq \bar{R}h$, where $\bar{R}$ is defined in (6.12) and set $\bar{B} = B_{\bar{R}h}(z) \cap \Omega$. If $u \in C^{1,1}(\bar{B})$, then

$$\left| |\partial I_h^{fd} u(z)| - \int_{\omega_z} \det D^2 u \right| \leq C|\omega_z|, \tag{6.19}$$

where the constant C depends only on $|u|_{C^{1,1}(\bar{B})}$.

Proof. As in Lemma 6.32, it suffices to show that

$$|\partial I_h^{fd} u(z)| \leq \int_{\omega_z} \det D^2 u + C|\omega_z|.$$

Since $\lambda I \leq D^2 u \leq \Lambda I$, Lemma 6.23 yields $A_z \subset B_{\bar{R}h}(z) \cap \Omega$. Recall now that $\Gamma(I_h^{fd} u)$ is piecewise linear with respect to a triangulation that has as nodes $\tilde{\Omega}_h$. The $C^{1,1}$-regularity assumption of u implies that, if $K \subset \omega_z \subset \bar{B}$ is an element of this triangulation, we have

$$D\Gamma(I_h^{fd} u)|_K = Du(z) + \mathbf{v}_K \quad |\mathbf{v}_K| \leq Ch|u|_{C^{1,1}(\bar{B})}.$$

Thus, by Lemma 3.22 (characterization of subdifferential), we deduce that the piecewise gradient $D\Gamma(I_h^{fd} u)|_K$ is contained in a ball centred at $Du(z)$ and with radius $Ch|u|_{C^{1,1}(\bar{B})}$. Thus, we have

$$|\partial I_h^{fd} u(z)| \leq \int_{\omega_z} \det D^2 u + C|u|^d_{C^{1,1}(\bar{B})}|\omega_z|.$$

This completes the proof. □

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

To control the L^∞-error caused by the $O(1)$ error near the boundary, we construct a discrete barrier function below. We refer to Nochetto and Zhang (2017*b*) for a proof.

Lemma 6.34 (discrete barrier). Let Ω be uniformly convex and let Ω_h^I be a translation-invariant nodal set in Ω. Given a constant $M > 0$, for each node $z \in \Omega_h^I$ with $\operatorname{dist}(z, \partial\Omega) \leq \bar{R}h$, there exists a convex nodal function $p_z \in X_h^{fd}$ such that $|\partial p_z(z_i)| \geq M|\omega_z|$ for all $z_i \in \bar{\Omega}_h$, $p_z(z_i) \leq 0$ on $z_i \in \Omega_h^B$ and

$$|p_z(z)| \leq CRM^{1/d}h,$$

for sufficiently small h.

Now we are ready to prove the L^∞-error estimate.

Theorem 6.35 (rate of convergence I). Assume that Ω is uniformly convex, and let u be the strictly convex (Alexandrov) solution of Monge–Ampère equation (6.1) with $f \geq \lambda^d > 0$. Suppose that the nodes Ω_h^I are translation-invariant, and let $u_h \in X_h^{fd}$ be the solution of (6.9). If $\lambda I \leq D^2 u \leq \Lambda I$ and $u \in C^{2,\alpha}(\Omega)$, then

$$\|u - \Gamma(u_h)\|_{L^\infty(\Omega)} \leq Ch^\alpha,$$

where the constant

$$C = C(d, \Omega, \lambda, \Lambda)(|u|_{C^{2,\alpha}(\bar{\Omega})} + |u|_{C^{1,1}(\bar{\Omega})}).$$

Proof. We begin by constructing a piecewise linear approximation of u. Recall that $\Gamma(I_h^{fd}u)$, the convex envelope of the nodal function $I_h^{fd}u$, is a piecewise linear function over a mesh that has $\bar{\Omega}_h$ as nodes. Thus, classic interpolation theory yields

$$\|\Gamma(I_h^{fd}u) - u\|_{L^\infty(\Omega)} \leq Ch^2|u|_{C^{1,1}(\bar{\Omega})}.$$

Therefore, we only need to prove that

$$\sup_{\bar{\Omega}_h}(I_h^{fd}u - u_h)^- \leq Ch^\alpha. \tag{6.20}$$

A similar inequality, which controls the positive part of $I_h^{fd}u - u_h$, can be derived in an analogous fashion.

Step 1. We first show that for all $z \in \bar{\Omega}_h$ such that $\operatorname{dist}(z, \partial\Omega) \leq \bar{R}h$,

$$(I_h^{fd}u - u_h)(z) \geq -Ch|u|_{C^{1,1}(\bar{\Omega})}. \tag{6.21}$$

Let p_z be the discrete barrier defined in Lemma 6.34 with free parameter M and consider the function $u_h + p_z$. Since Lemma 3.21 (addition inequality) implies

$$\partial u_h(z_i) + \partial p_z(z_i) \subset \partial(u_h + p_z)(z_i),$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

by Proposition 6.28 (Brunn–Minkowski inequality), we obtain

$$|\partial(u_h + p_z)(z_i)| \geq \big(|\partial u_h(z_i)|^{1/d} + |\partial p_z(z_i)|^{1/d}\big)^d.$$

Therefore, by Lemmas 6.33 and 6.34, we have

$$\begin{aligned}|\partial(u_h + p_z)(z_i)| &\geq \left(\left(\int_{\omega_{z_i}} \det(D^2 u)\right)^{1/d} + (M|\omega_z|)^{1/d}\right)^d \\ &\geq |\partial I_h^{fd} u(z_i)|, \quad \text{for all } z_i \in \Omega_h^I\end{aligned}$$

provided that $M = C|u|^d_{C^{1,1}(\bar{B})}$ Since $p_z \leq 0$ on Ω_h^B, we have $u_h + p_z \leq I_h^{fd} u$ on Ω_h^B. Moreover, because $|\partial(u_h + p_z)(z_i)| \geq |\partial I_h^{fd} u(z_i)|$ for all $z_i \in \Omega_h^I$, we have, by the maximum principle of Corollary 6.31,

$$u_h(z_i) + p_z(z_i) \leq I_h^{fd} u(z_i), \quad \text{for all } z_i \in \bar{\Omega}_h.$$

The estimate on the discrete barrier function, given in Lemma 6.34, yields

$$u_h(z) - Ch|u|_{C^{1,1}(\bar{\Omega})} \leq u_h(z) + p_z(z) \leq I_h^{fd} u(z), \tag{6.22}$$

thus proving (6.21).

Step 2. For all nodes z with $\operatorname{dist}(z, \partial\Omega) \geq \bar{R}h$, thanks to the consistency of the method, Lemma 6.32, we have

$$\big||\partial I_h^{fd} u(z)| - |\partial u_h(z)|\big| \leq Ch^\alpha |u|_{C^{2,\alpha}(\bar{\Omega})} |\omega_z|. \tag{6.23}$$

We apply the stability result of Proposition 6.30 on a smaller domain,

$$\bar{\Omega}_{*,h} = \{z \in \Omega_h^I : \operatorname{dist}(z, \partial\Omega) \geq \bar{R}h\}, \tag{6.24}$$

and on the nodal functions $I_h^{fd} u$ and $u_h - Ch|u|_{C^{1,1}(\bar{\Omega})}$, where C is the constant in (6.21) so that $I_h^{fd} u \geq u_h - Ch|u|_{C^{1,1}(\bar{\Omega})}$ on the boundary nodes of $\bar{\Omega}_{*,h}$. Letting $\mathcal{C}^-_{h,*}$ denote the contact set of $I_h^{fd} u - (u_h - Ch|u|_{C^{1,1}(\bar{\Omega})})$ with respect to $\bar{\Omega}_{h,*}$, we get

$$\begin{aligned}&\sup_{\bar{\Omega}_{*,h}} \big(I_h^{fd} u - (u_h - Ch|u|_{C^{1,1}(\bar{\Omega})})\big)^- \\ &\qquad \leq C\left(\sum_{z \in \mathcal{C}^-_{h,*}} \big(|\partial I_h^{fd} u(z)|^{1/d} - |\partial u_h(z)|^{1/d}\big)^d\right)^{1/d}.\end{aligned} \tag{6.25}$$

Note now that $t \to t^{1/d}$ is a concave function and therefore

$$(t + \delta)^{1/d} \leq t^{1/d} + \frac{1}{d} t^{(1-d)/d} \delta.$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Thus, by setting $t = |\partial u_h(z)|$, $\delta = |\partial I_h^{fd} u(z)| - |\partial u_h(z)|$ and applying (6.23), we find

$$\begin{aligned}|\partial I_h^{fd} u(z)|^{1/d} - |\partial u_h(z)|^{1/d} &\leq \frac{1}{d}|\partial u_h(z)|^{(1-d)/d}(|\partial u_h(z)| - |\partial I_h^{fd} u(z)|) \\ &\leq Ch^\alpha |u|_{C^{2,\alpha}(\bar\Omega)} |\omega_z| \left(\int_{\omega_z} f \right)^{(1-d)/d} \\ &\leq Ch^\alpha |u|_{C^{2,\alpha}(\bar\Omega)} |\omega_z|^{1/d}\end{aligned}$$

because $f \geq \lambda^d > 0$. Inserting this estimate into (6.25) yields

$$\begin{aligned}\sup_{\bar\Omega_{h,*}} (I_h^{fd} u - (u_h - Ch|u|_{C^{1,1}(\bar\Omega)}))^- &\leq Ch^\alpha |u|_{C^{2,\alpha}(\bar\Omega)} \left(\sum_{z \in \mathcal{C}_{h,*}^-} |\omega_z| \right)^{1/d} \\ &\leq Ch^\alpha |u|_{C^{2,\alpha}(\bar\Omega)} |\Omega|^{1/d},\end{aligned}$$

or equivalently

$$\sup_{\bar\Omega_{h,*}} (I_h^{fd} u - u_h)^- \leq Ch|u|_{C^{1,1}(\bar\Omega)} + Ch^\alpha |u|_{C^{2,\alpha}(\bar\Omega)}.$$

This inequality, together with (6.22), proves the lower bound for $I_h^{fd} u - u_h$. The upper bound can be proved in a similar fashion. □

Note that Proposition 6.30 controls the L^∞-error by the L^d-norm of the consistency error. Thus, if large consistency errors occur only in regions with small measure, we may still derive a rate of convergence. This observation may be used for solutions that are not $C^2(\Omega)$-regular.

To state the result, we first introduce the Minkowski–Bouligand dimension. Let U be a subset of Ω. Let $\{\omega_i\}_{z_i \in \Omega_h^I}$ be a translation-invariant partition covering Ω where ω_i is as in (6.10). Define $m = m(h)$ to be the number of elements of $\{\omega_{z_i}\}$ required to cover U. We define the (Minkowski–Bouligand) dimension of U as

$$\dim U = -\lim_{h \to 0} \frac{\log m(h)}{\log h}.$$

For example, it is easy to check that ∂B_1, the discontinuity set of $D^2 u$ in Example 6.5, is of dimension one. Note that in this example the solution $u \in C^{1,1}(\bar\Omega) \setminus C^2(\bar\Omega)$.

The next result addresses the rate of convergence for piecewise smooth solutions such that the discontinuity set of $D^2 u$ is of low dimension.

Theorem 6.36 (rate of convergence II). Let $u \in C^{1,1}(\bar\Omega)$ be strictly convex with $\lambda I \leq D^2 u \leq \Lambda I$, and solve (6.1). Assume that $D^2 u$ is piecewise Hölder-continuous (with exponent $\alpha > 0$) and its discontinuity set U has dimension $n < d$. Let u_h be the solution of (6.9), over a lattice Ω_h^I of

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

translation-invariant nodes. Then

$$\| u - \Gamma(u_h) \|_{L^\infty(\Omega)} \le Ch^\alpha |u|_{C^{2,\alpha}(\bar\Omega\setminus U)} + Ch^{(d-n)/d} |u|_{C^{1,1}(\bar\Omega)}.$$

Proof. Following the estimate of Theorem 6.35, we first note that for $z \in \bar\Omega_h$ with $\operatorname{dist}(z, \partial\Omega) \le \bar R h$,

$$|(u_h - I_h^{fd} u)(z)| \le Ch. \tag{6.26}$$

Since $D^2 u$ is Hölder-continuous except on the set U and the aspect ratio Λ/λ of $D^2 u$ is bounded, by Lemmas 6.32 and 6.33 we have that

$$|\partial I_h^{fd} u(z)| \le \begin{cases} \displaystyle\int_{\omega_z} \det D^2 u + Ch^\alpha |\omega_z|\, |u|_{C^{2,\alpha}(\bar\Omega\setminus U)} & z \in \bar\Omega_{h,*} \setminus U_{\bar R h}, \\ \displaystyle\int_{\omega_z} \det D^2 u + C|\omega_z|\, |u|_{C^{1,1}(\bar\Omega)} & z \in U_{\bar R h}, \end{cases}$$

where $\bar\Omega_{h,*}$ is given by (6.24) and

$$U_{\bar R h} = \{ z \in \bar\Omega_{h,*} : \operatorname{dist}(z, U) \le \bar R h \}.$$

Now, the stability estimate of Proposition 6.30 yields

$$\begin{aligned} &\sup_{\bar\Omega_{h,*}} \big(I_h^{fd} u - (u_h - Ch|u|_{C^{1,1}(\bar\Omega)}) \big)^- \\ &\qquad \le C \Bigg(\sum_{z \in \mathcal{C}_{h,*}^-} \big(|\partial I_h^{fd} u(z)|^{1/d} - |\partial u_h(z)|^{1/d} \big)^d \Bigg)^{1/d}, \end{aligned} \tag{6.27}$$

where $C > 0$ is the constant in (6.26).

For $z \in \bar\Omega_{h,*} \setminus U_{\bar R h}$, we apply the same arguments as in the proof of Theorem 6.35, to get

$$|\partial I_h^{fd} u(z)|^{1/d} - |\partial u_h(z)|^{1/d} \le Ch^\alpha |u|_{C^{2,\alpha}(\bar\Omega\setminus U)} |\omega_z|^{1/d}.$$

For $z \in U_{\bar R h}$ similar arguments (essentially take $\alpha = 0$) yield

$$|\partial I_h^{fd} u(z)|^{1/d} - |\partial u_h(z)|^{1/d} \le C|u|_{C^{1,1}(\bar\Omega)} |\omega_z|^{1/d}.$$

Inserting these estimates into (6.27), we deduce that

$$\begin{aligned} \sup_{\bar\Omega_{h,*}} (I_h^{fd} u - u_h)^- &\le C \Bigg(h^{\alpha d} |u|^d_{C^{2,\alpha}(\bar\Omega\setminus U)} |\Omega| + |u|^d_{C^{1,1}(\bar\Omega)} \sum_{z \in U_{\bar R h}} |\omega_z| \Bigg)^{1/d} \\ &\le Ch^\alpha |u|_{C^{2,\alpha}(\bar\Omega\setminus U)} + C|u|_{C^{1,1}(\bar\Omega)} \Bigg(\sum_{z \in U_{\bar R h}} |\omega_z| \Bigg)^{1/d}. \end{aligned}$$

Since $\dim U = n < d$, we have

$$\sum_{z \in U_{\bar R h}} |\omega_z| = N|\omega_z| \le CNh^d$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

with $N \leq Ch^{-n}$. Thus, we conclude that

$$\sup_{\bar{\Omega}_{h,*}}(I_h^{fd}u - u_h)^- \leq C(h^\alpha |u|_{C^{2,\alpha}(\bar{\Omega}\setminus U)} + h^{(d-n)/d}|u|_{C^{1,1}(\bar{\Omega})}).$$

This completes the proof. □

Remark 6.37 (extensions). The developments of this section have found the following extensions.

(1) Nochetto *et al.* (2017) have recently applied the error analysis developed for method (6.9) to the wide-stencil schemes by Oberman (2008) and Froese and Oberman (2011).

(2) The error estimate stated in Theorems 6.35 and 6.36 applies only to structured nodes (lattices). It remains an open problem how to extend the analysis to unstructured nodes and to the degenerate case, that is, when $f = 0$ in some region.

(3) The error analysis of monotone schemes for the Monge–Ampère equation in other norms, such as the H^1-norm, remains an open problem.

7. Discretizations of non-convex second-order elliptic equations

This section is a continuation of the developments of Section 5, where we apply the discretizations and results for uniformly elliptic linear PDEs to fully nonlinear problems. However, in contrast to Section 5, we shall not assume convexity (or concavity) of the differential operator, but rather, only assume it is uniformly elliptic. As explained in Section 1.3, it suffices to consider numerical approximations of the Isaacs equation given in Example 2.13. For simplicity and to communicate the essential points in the discussion, we shall assume that the nonlinear problem does not have lower-order terms, and in addition has homogeneous Dirichlet boundary conditions. Thus we consider numerical approximations for the problem

$$\begin{cases} F(x, D^2u) := \inf_{\beta\in\mathcal{B}} \sup_{\alpha\in\mathcal{A}} [\mathcal{L}^{\alpha,\beta}u(x) - f^{\alpha,\beta}(x)] = 0 & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{7.1}$$

with $\mathcal{L}^{\alpha,\beta}u(x) = A^{\alpha,\beta}(x) : D^2u(x)$, and the coefficient matrices satisfy

$$\lambda I \leq A^{\alpha,\beta} \leq \Lambda I, \quad \text{for all } \alpha \in \mathcal{A},\ \beta \in \mathcal{B},$$

so that F is uniformly elliptic. We assume that for each $x \in \Omega$, the mapping $M \mapsto F(x, M)$ is continuous and locally Lipschitz-continuous on $\mathbb{S}^d$, and that $f^{\alpha,\beta} \in C(\bar{\Omega})$ for each $\alpha \in \mathcal{A}$ and $\beta \in \mathcal{B}$. Further assumptions may be made in subsequent developments.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

As in the convex case, we can roughly classify numerical schemes as finite difference, finite element and semi-Lagrangian methods. The construction and analysis of finite difference schemes as detailed in Section 7.1, in principle, follow the convex case of Section 5 and the Barles–Souganidis theory as presented in Theorems 3.7 and 3.8, but are clouded by numerous technicalities that have for many years prevented researchers from obtaining rates of convergence. In fact – as described in the introduction of Kuo and Trudinger (1992) – this was considered an important unsolved open problem for several years.

The heart of the issue can be captured by examining the proof of Theorem 5.3. An important step there is the construction of a smooth subsolution to the equation (scheme) which, in the convex case, can be obtained by mollification of a subsolution to a perturbed equation. Convexity of the operator allows us to claim that this is a subsolution to the original problem and can thus be used to carry out the programme detailed at the beginning of Section 5.1. However, without convexity, it is not clear how to construct a smooth approximation to the solution of (7.1), which can be used to invoke the consistency of the scheme. This is particularly important in the non-convex case since, as shown in Example 2.70, one cannot assume smoothness of the solution to (7.1).

For many years, all of the available results were rather specialized. For instance, Jakobsen (2004) considers a one-dimensional problem and clearly shows that the arguments do not extend to more dimensions. A particular case of an Isaacs equation – an obstacle problem for an HJB equation – is discussed in Jakobsen (2006), where this special structure is exploited.

The derivation of rates of convergence for general schemes remained an unsolved problem until Caffarelli and Souganidis (2008) showed how to obtain a rate of convergence, within the Barles–Souganidis framework, for approximations of (7.1) in the case that, for all $\alpha \in \mathcal{A}$ and $\beta \in \mathcal{B}$, the matrices $A^{\alpha,\beta}$ do not depend on x. We detail these results in Section 7.2, where we also comment on extensions and variations to these estimates. Simply put, the estimates assert that there exists an algebraic rate of convergence. An explicit rate, however, is not available at the moment.

Quite recently, Salgado and Zhang (2016) have extended the results of the finite element method in Section 4.3 to the case of (7.1), and obtained an algebraic rate of convergence which, as in the finite difference case, is not explicit. These developments are detailed in Section 7.3.

We also comment that, as of the time of writing, no rates of convergence are available for semi-Lagrangian schemes. The only known result is convergence, as obtained by Debrabant and Jakobsen (2013).

We conclude our discussion on non-convex equations in Section 7.4 by describing how to solve the nonlinear system of equations that results after discretization, by any of the schemes discussed before.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

7.1. Finite difference methods

Here, following the framework given in Section 3.2 and in Kuo and Trudinger (1992), we construct finite difference approximations to the fully nonlinear problem (7.1), and study the stability and convergence of these discretizations. We consider the following problem: find $u_h \in X_h^{fd}$ satisfying

$$\begin{cases} F_h[u_h] = 0 & \text{in } \Omega_h^I, \\ u_h = 0 & \text{on } \Omega_h^B, \end{cases} \tag{7.2}$$

where the interior and boundary nodes are as in Definition 3.14. We assume that F_h is consistent and that the scheme is of the form $F_h[u_h](z) = \mathcal{F}_h(z, \delta_h^2 u_h(z))$ with $\delta_h^2 u_h(z) = \{\delta_{y,h}^2 u_h(z) : y \in S\}$. We further assume that $\mathcal{F}_h$ is of positive type in the sense of Definition 3.12, so that F_h is monotone, and that

$$\frac{\partial \mathcal{F}_h}{\partial s_y} \leq \Lambda_0,$$

for some $\Lambda_0 > 0$. For example, schemes that satisfy these properties, and the ones that we have in mind, are

$$F_h[u_h] = \inf_{\beta \in \mathcal{B}} \sup_{\alpha \in \mathcal{A}} (\mathcal{L}_h^{\alpha,\beta} u_h - f^{\alpha,\beta}),$$

where each (linear) discrete operator is given by

$$\mathcal{L}_h^{\alpha,\beta} u_h(z) = \sum_{y \in S} a_y^{\alpha,\beta}(z) \delta_{y,h}^2 u_h(z),$$

and is of positive type and consistent with $\mathcal{L}^{\alpha,\beta}$; Section 3.4 describes how to construct linear operators with these properties.

Before discussing the solvability of the finite difference scheme (7.2), let us show first that solutions to the scheme are uniformly bounded. Let v_h, w_h be two grid functions, and consider the linearization

$$\begin{aligned} F_h[v_h](z) - F_h[w_h](z) &= \mathcal{L}_h(v_h - w_h)(z) \\ &:= \sum_{y \in S} a_y(z) \delta_{y,h}^2 (v_h - w_h)(z), \end{aligned} \tag{7.3}$$

with

$$a_y(z) = \int_0^1 \frac{\partial \mathcal{F}_h}{\partial s_y}(z, \delta_{h,y}^2 q_t(z))\, \mathrm{d}t, \quad q_t = t v_h + (1-t) w_h.$$

In particular, if we set $v_h = u_h$, $w_h = 0$, and $f = -F(x, 0)$, then the solution to (7.2) satisfies

$$\mathcal{L}_{h,u_h} u_h = f \quad \text{in } \Omega_h^I, \tag{7.4}$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

where the coefficients in the operator $\mathcal{L}_{h,u_h}$ depend on u_h. Since $\mathcal{L}_{h,u_h}$ is a positive operator, Theorem 3.25 yields the follow stability result.

Theorem 7.1 (uniqueness and stability). In this setting, solutions to (7.2) are unique, and satisfy

$$\|u_h\|_{L^\infty(\bar{\Omega}_h)} \leq C\left(\sum_{z\in\Omega_h^I} h^d |f(z)|^d\right)^{1/d} \leq C.$$

In addition to uniqueness and stability, these *a priori* estimates also imply the existence of solutions.

Theorem 7.2 (existence). Under these conditions, problem (7.2) has a unique solution.

Proof. Consider the map $Q_h : X_h^{fd} \to X_h^{fd}$ satisfying

$$\begin{cases} \mathcal{L}_{h,v_h} Q_h(v_h) = f & \text{in } \Omega_h^I, \\ v_h = 0 & \text{on } \Omega_h^B. \end{cases}$$

Theorem 3.25 ensures that Q_h is well-defined and that $\|Q_h(v_h)\|_{L^\infty(\bar{\Omega}_h)} \leq C$. Brouwer's fixed-point theorem shows that Q_h has a fixed point $u_h \in X_h^{fd}$, which is a solution to (7.4), and hence (7.2). □

Remark 7.3 (other proofs). Other (constructive) proofs of existence of solutions can be found in Kuo and Trudinger (1992, Section 4) and in Section 7.4.

Finally we apply Theorem 3.30 to deduce that solutions to (7.2) are Hölder-continuous.

Proposition 7.4 (discrete Hölder continuity). Suppose that $u_h \in X_h^{fd}$ solves (7.2). Then there exists $\eta \in (0,1)$ and $C > 0$ depending on the data, such that, for $z_1, z_2 \in \bar{\Omega}_h$,

$$|u_h(z_1) - u_h(z_2)| \leq C|z_1 - z_2|^\eta.$$

Proposition 7.4 implies that the sequence of solutions $\{u_h\}_{h>0}$ is equicontinuous. Thus, applying Theorem 3.8 we obtain convergence to the viscosity solution.

Theorem 7.5 (convergence). The solutions u_h to (7.2) converge locally uniformly to the viscosity solution of (7.1).

7.2. Rates of convergence for finite difference schemes with constant coefficients

Here we summarize the results regarding rates of convergence of finite difference schemes for fully nonlinear, non-convex PDEs, as originally shown

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

by Caffarelli and Souganidis (2008). As a starting point, we focus on the case where the operator in (7.1) is such that, for every $\alpha \in \mathcal{A}$ and $\beta \in \mathcal{B}$, the matrices $A^{\alpha,\beta}$ are independent of x, and, moreover, we have $f^{\alpha,\beta} = f \in C^{0,1}(\Omega)$. Under these conditions, problem (7.1) satisfies a comparison principle (see Theorem 2.52), and there exists a unique viscosity solution with regularity $u \in C^{1,s}(\bar{\Omega})$ for some universal constant $s > 0$; see Theorem 2.71.

In order to derive rates, we shall, with an abuse of notation, extend the domain of F_h to continuous functions and to all

$$z \in \Omega^I := \{x \in \Omega : \operatorname{dist}\{x, \partial\Omega\} \geq mh\} \supset \Omega_h^I,$$

where m is the stencil size. We also assume, in this section, the following (strengthened) consistency criterion:

$$|F(z, D^2\phi(z)) - F_h[\phi](z)| \leq Ch\|D^3\phi\|_{L^\infty(\Omega)}, \tag{7.5}$$

for all smooth ϕ and $z \in \Omega_h^I$.

Now, as in the convex case, the key idea to obtain rates of convergence is to construct a smooth function u_ϵ that is a subsolution to (7.1) and then apply the consistency and monotonicity of the finite difference operator to get a one-sided bound of the error; see the proof of Theorem 5.3 for details. However, unlike the convex case, it is not immediate how to construct a sufficiently smooth subsolution to carry out this programme. For example, the standard mollification $u^\epsilon = u_\epsilon \star \rho_\epsilon$ used in Theorem 5.3 is no longer a subsolution due to the lack of convexity of F (and F_h).

Instead, we employ the so-called *sup-* (*inf-*) *convolutions* of the viscosity solution u.

Definition 7.6 (sup-convolution). Let $u \in C(\bar{\Omega})$ and $\tau > 0$, The *sup-* (*inf-*) *convolution* u_τ^+ of u is

$$u_\tau^+(x) = \sup_{y\in\bar{\Omega}}\left[u(y) - \frac{1}{h^\tau}|x-y|^2\right],$$
$$u_\tau^-(x) = \inf_{y\in\bar{\Omega}}\left[u(y) + \frac{1}{h^\tau}|x-y|^2\right].$$

Remark 7.7 (alternative definition). Note that the definition we give here is tied to a mesh size h. We do so because this is the scale that suits our needs. In general the literature defines the sup-convolution of a function by

$$u_\varrho^+ = \sup_{y\in\bar{\Omega}}\left[u(y) - \frac{1}{2\varrho}|x-y|^2\right], \quad \varrho > 0.$$

The change of variables $\varrho = \frac{1}{2}h^\tau$ shows the relation between these two. Similar reasoning can be used for the inf-convolution.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

We show two examples of sup-convolution of functions to motivate the introduction of this useful concept.

Example 7.8 (sup-convolution I). Let $u = |x|^2$ be defined in $\mathbb{R}^d$ and consider its sup-convolution

$$u_\tau^+(x) = \sup_{y\in\mathbb{R}^d}\left[|y|^2 - \frac{1}{h^\tau}|x-y|^2\right].$$

Let $y_* = y_*(x)$ be a point where the supremum is attained. Then a simple calculation shows that

$$2y_* + 2h^{-\tau}(x - y_*) = 0,$$

which implies that $y_* = (1-h^\tau)^{-1}x$. Inserting y_* into the definition of $u_\tau^+(x)$, we obtain

$$u_\tau^+(x) = \left(\frac{1}{1-h^\tau}\right)^2|x|^2 - h^{-\tau}\left(\frac{h^\tau}{1-h^\tau}\right)^2|x|^2 = \frac{|x|^2}{(1-h^\tau)}.$$

See Figure 7.1 and note that $D^2u_\tau^+ > D^2u$.

Next, we consider a function that is less smooth.

Example 7.9 (sup-convolution II). Let $u = -|x|$ be defined in $\mathbb{R}^d$. The maximum is attained at $y_* = y_*(x)$ if and only if

$$\mathbf{0} \in -\partial|y_*| + 2h^{-\tau}(x - y_*)$$

or, equivalently,

$$x \in y_* + \frac{h^\tau}{2}\partial|y_*|,$$

where $\partial|y_*|$ denotes the subdifferential of $|\cdot|$ at y_*. We note that if $|x| \le h^\tau/2$, then $y_* = \mathbf{0}$ because $\partial|\mathbf{0}| = B_1(0)$. Otherwise, we have

$$x = y_* + \frac{h^\tau y_*}{2|y_*|}$$

because $\partial|y_*| = \{y_*/|y_*|\}$ for $|y_*| > 0$. Therefore, we conclude that

$$u_\tau^+(x) = \begin{cases} -h^{-\tau}|x|^2 & |x| \le \dfrac{h^\tau}{2}, \\ -|x| + \dfrac{h^\tau}{4} & \text{otherwise.} \end{cases}$$

Note that u_τ^+ is $C^{1,1}$, while u is only Lipschitz. Moreover, near the singularity $x = 0$, u_τ^+ behaves like a paraboloid with $D^2u_\tau^+ = -2h^{-\tau}I$; see Figure 7.2.

Example 7.9 shows that, intuitively speaking, the sup-convolution u_τ^+ 'opens up' the kinks of u.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

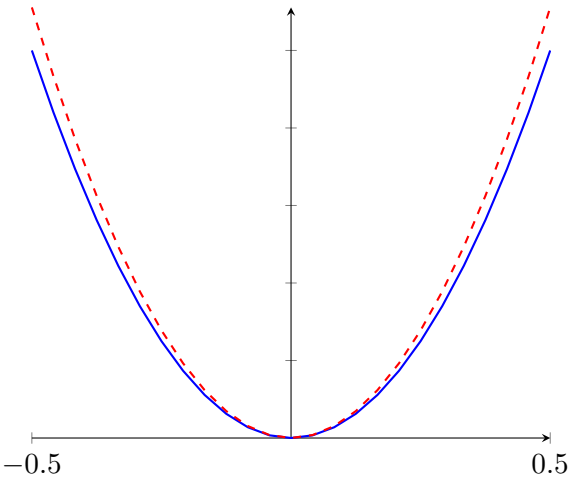


Figure 7.1. The graph of the function $u(x) = x^2$ (solid blue) and its sup-convolution (dashed red) with parameters $h = 0.1$ and $\tau = 1$.

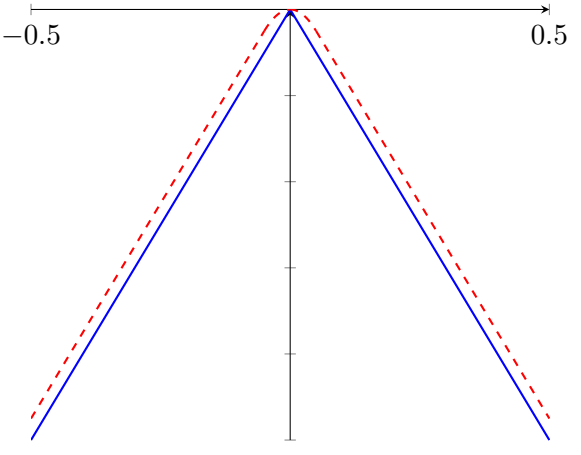


Figure 7.2. The graph of the function $u(x) = -|x|$ (solid blue) and its sup-convolution (dashed red) with parameters $h = 0.1$ and $\tau = 1$.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Figures 7.1 and 7.2 illustrate some of the general properties of sup- and inf-convolutions. To state them concisely, we begin with a definition.

Definition 7.10 (opening t). We say that $p \in \mathbb{P}_2$ is a paraboloid of *opening* t if, for some $\ell \in \mathbb{P}_1$, we have

$$p(x) = \ell(x) \pm \frac{t}{2}|x|^2.$$

The main properties of sup- and inf-convolutions are as follows. To simplify the presentation, we let $\bar{\partial}u(x)$ denote the superdifferential of u at the point x, that is,

$$\mathbf{v} \in \bar{\partial}u(x) \Leftrightarrow -\mathbf{v} \in \partial(-u)(x).$$

Proposition 7.11 (properties of u_τ^+). Let $u \in C^{0,1}(\bar{\Omega})$ and, for $x \in \Omega$, let $y_* = y_*(x)$ denote the point where the supremum in the definition of u_τ^+ is attained. The following statements hold:

(i) $x = y_* + \frac{1}{2}h^\tau \mathbf{v}$ for some vector $\mathbf{v} \in \bar{\partial}u(y_*)$ and, therefore, $|x - y_*| \le Ch^\tau$,

(ii) $\|u - u_\tau^+\|_{L^\infty(\Omega)} \le Ch^\tau$,

(iii) there exists a paraboloid of opening $2h^{-\tau}$, that touches u_τ^+ (resp. u_τ^-) from below (resp. above) at x,

(iv) if u is the viscosity solution to (7.1), then $|x_1 - x_2| \le C|y_*(x_1) - y_*(x_2)|$.

Proof. Let us prove each statement separately.

(i) For any fixed x, if $u(y) - h^{-\tau}|x - y|^2$ attains its maximum at y_*, then

$$\mathbf{0} \in \bar{\partial}u(y_*) - 2h^{-\tau}(x - y_*).$$

Thus, we have $x = y_* + \frac{1}{2}h^\tau \mathbf{v}$ for some $\mathbf{v} \in \bar{\partial}u(y_*)$.

(ii) By definition $u_\tau^+ \ge u$. Moreover, property (i) implies that

$$0 \le u_\tau^+(x) - u(x) = u(y_*) - u(x) - h^{-\tau}\left|\frac{1}{2}h^\tau \mathbf{v}\right|^2 \le |u(x) - u(y_*)| + \frac{1}{4}h^\tau |\mathbf{v}|^2$$

for some $\mathbf{v} \in \bar{\partial}u(y_*)$. Since u is Lipschitz-continuous, we conclude that, for every $x \in \Omega$,

$$|u(x) - u_\tau^+(x)| \le Ch^\tau.$$

(iii) Since, for any $x \in \Omega$, $y_* = y_*(x)$ is the point where the supremum is attained, let us define the paraboloid

$$p(z) = u(y_*) - \frac{1}{h^\tau}|z - y_*|^2,$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

and note that $p(x) = u_\tau^+(x)$. Moreover,

$$p(z) \leq \sup_{y\in\bar{\Omega}}\left[u(y) - \frac{1}{h^\tau}|z-y|^2\right] =: u_\tau^+(z), \quad \text{for all } z \in \bar{\Omega}.$$

Thus, the paraboloid p touches the graph of u_τ^+ at point x from below.

(iv) Since, by definition,

$$u_\tau^+(x) = u(y_*(x)) - \frac{1}{h^\tau}|x - y_*(x)|^2,$$

upon defining the paraboloid

$$p(z) = u_\tau^+(x) + \frac{1}{h^\tau}|x-z|^2$$

we have that

$$p(y_*(x)) = u(y_*(x)), \quad p(z) \geq u(z), \quad \text{for all } z \in \bar{\Omega}.$$

In other words, at $y_*(x)$, the function u is touched from below by a paraboloid with opening $2h^{-\tau}$. On the other hand, since u solves (7.1), the Harnack inequality of Theorem 2.63 implies that, at $y_*(x)$, the function u can be touched from above by a paraboloid of opening $Ch^{-\tau}$ with C depending only on d, λ and Λ. Invoking Caffarelli and Cabré (1995, Proposition 1.2) we conclude that for every $x_1, x_2 \in \bar{\Omega}$ the function u is differentiable at $y_*(x_i)$, $i = 1, 2$ and, moreover,

$$|Du(y_*(x_1)) - Du(y_*(x_2))| \leq Ch^{-\tau}|y_*(x_1) - y_*(x_2)|.$$

We now invoke (i) to obtain that

$$|x_1 - x_2| = \left|y_*(x_1) - y_*(x_2) + \frac{1}{2}h^\tau Du(y_*(x_1)) - \frac{1}{2}h^\tau Du(y_*(x_2))\right|.$$

The previous two inequalities yield the claim.

This completes the proof. □

Next we establish a lower bound for $F_h[u_\tau^+](\cdot)$. We define

$$\Omega_h^{I,\tau} = \{z \in \Omega_h^I : \operatorname{dist}(z, \partial\Omega) \geq Ch^\tau\},$$

where the constant $C > 0$ depends on the stencil size m and is chosen so that $F_h[u](x)$ is well-defined for x satisfying $|x-z| \leq Ch^\tau$ for some $z \in \Omega_h^{I,\tau}$; see Figure 7.3.

Proposition 7.12 (bound on $F_h[u_\tau^+](\cdot)$). Assume that $u \in C^{0,1}(\bar{\Omega})$ and that $z \in \Omega_h^{I,\tau}$. Let $y_* = y_*(z) \in \Omega$ be chosen so that

$$u_\tau^+(z) = u(y_*) - \frac{1}{h^\tau}|z - y_*|^2. \tag{7.6}$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

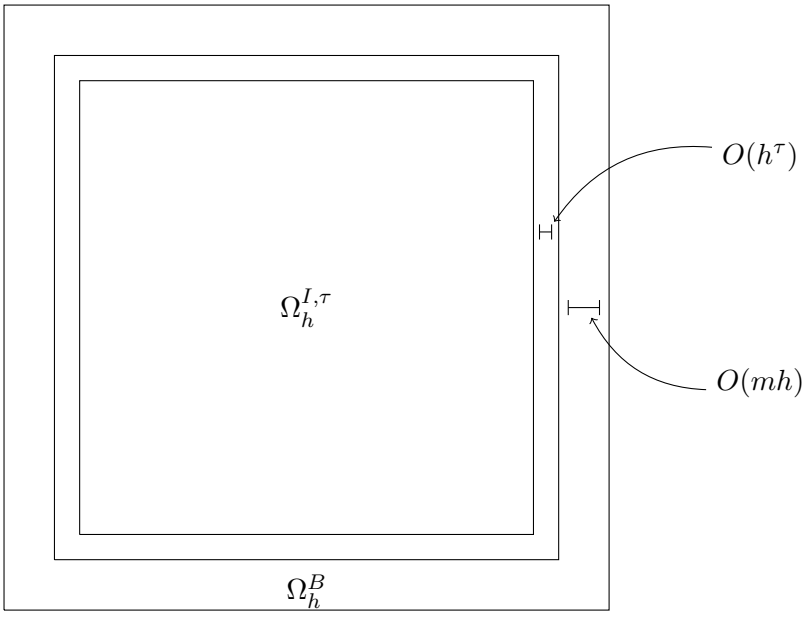


Figure 7.3. A pictorial description of the set $\Omega_h^{I,\tau}$. The values $F_h[u](x)$ are well-defined for x within a distance of Ch^τ of this set.

We then have

$$F_h[u_\tau^+](z) \geq F_h[u](y_*) + (f(y_*) - f(z)).$$

Moreover, for all $z \in \Omega_h^I$, we have

$$F_h[u_\tau^+](z) \geq -Ch^{-\tau}. \tag{7.7}$$

Proof. First note that, by Proposition 7.11(i) we have that $|z - y_*| \leq Ch^\tau$, and therefore $F_h[u](y_*)$ is well-defined. Now, owing to the monotonicity of F_h, to show the first estimate it suffices to show that

$$\delta_{y,h}^2 u_\tau^+(z) \geq \delta_{y,h}^2 u(y_*) \quad y \in S.$$

Define the function

$$v(x) = u(x - z + y_*) - \frac{1}{h^\tau}|z - y_*|^2,$$

and note that

$$v(z) = u(y_*) - \frac{1}{h^\tau}|z - y_*|^2 = u_\tau^+(z).$$

The change of variables $\zeta = x - z + y_*$ reveals that

$$v(x) = u(\zeta) - \frac{1}{h^\tau}|\zeta - x|^2 \leq u_\tau^+(x),$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

and therefore we conclude that, for any $y \in S$, $\delta^2_{y,h} v(z) \leq \delta^2_{y,h} u^+_\tau(z)$. This, together with the fact that, for any $y \in S$, $\delta^2_{y,h} v(z) = \delta^2_{y,h} u(y_*)$, yields

$$\delta^2_{y,h} u(y_*) = \delta^2_{y,h} v(z) \leq \delta^2_{y,h} u^+_\tau(z).$$

Since f is independent of α and β, this proves the estimate $F_h[u^+_\tau](z) + f(z) \geq F_h[u](y_*) + f(y_*)$.

To prove (7.7), let

$$w(x) = u(y_*) - \frac{1}{h^\tau}|x - y_*|^2.$$

Clearly $w(z) = u^+_\tau(z)$ and $w(x) \leq u^+_\tau(x)$. Therefore, monotonicity and the identity

$$\delta^2_{y,h} w(z) = -\frac{2}{h^\tau}|y|^2$$

yields

$$F_h[u^+_\tau](z) \geq F_h[w](z) \geq -Ch^{-\tau},$$

which concludes the proof. □

Our goal now is to compare u^+_τ with the numerical solution u_h through $F_h[u_h]$ and $F_h[u^+_\tau]$ and a comparison principle for the discrete operator F_h. From Proposition 7.2 it follows that $F_h[u^+_\tau]$ at z depends on the consistency error of $F_h[u]$ at $y_* := y_*(z)$. The consistency of F_h, as stated in (7.5), implies that

$$F_h[p](y_*) = F(y_*, D^2 p), \quad \text{for all } p \in \mathbb{P}_2.$$

Consequently, the consistency error $F_h[u](y_*)$ depends on the quality of the approximation of the solution u by a quadratic polynomial at y_*. This, in turn, depends on the regularity of u.

The next result asserts that in our setting, outside sets of arbitrarily small measure, viscosity solutions to (7.1) have pointwise second-order Taylor expansions with an error that is controlled by the size of the singular set. For a proof, the reader is referred to Caffarelli and Souganidis (2008, Theorem A).

Theorem 7.13 (quadratic approximation). Assume that F has constant coefficients, is uniformly elliptic with ellipticity constants $\lambda, \Lambda > 0$, and that $f \in C^{0,1}(\bar{\Omega})$. Let u be a Lipschitz-continuous solution of $F(x, D^2 u) = 0$ in $B_{2r}(x_0)$. There exist positive constants σ, t_0 and C that depend on λ, Λ and d such that, for all $t > t_0$, it is possible to find an open set $A_t \subset \Omega$ such that if $x \in A_t$, there exists a quadratic polynomial $p_x \in \mathbb{P}_2$ of opening less than t that satisfies

$$F(x, D^2 p_x(x)) = 0.$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Moreover,

$$|u(y) - u(x) - p_x(y-x)| \le Ct|x-y|^3, \quad \text{for all } y \in B_{2r}(x_0), \tag{7.8}$$

and

$$|\Omega \backslash A_t| \le Ct^{-\sigma}.$$

This theorem essentially says that, for every t, except in a singular set of measure $t^{-\sigma}$, which we denote by $\mathcal{S}_t$, the function u has a second-order Taylor expansion. We define the regular set $\mathcal{R}_t := \bar{\Omega} \backslash \mathcal{S}_t$.

Let us now explain how Theorem 7.13 can be used to obtain a rate of convergence. By consistency, (7.5), we have that, for every $p \in \mathbb{P}_2$,

$$F_h[p](z) = F(z, D^2 p(z)), \quad \text{for all } z \in \Omega.$$

Therefore, if $t > 0$, $z \in \Omega$ and $B_h(z) \cap \mathcal{R}_t \neq \emptyset$, we choose $y \in B_h(z) \cap \mathcal{R}_t$ and set $p = p_y$, the second-order expansion of u at y from Theorem 7.13. We then estimate the consistency error as follows:

$$|F_h[u](z)| \le |F_h[u](z) - F_h[p_y](z)| + |F_h[p_y](z) - F_h[p_y](y)| + |F_h[p_y](y)|.$$

From estimate (7.8) and the fact that $F_h[p_y](z) - F_h[p_y](y) = f(y) - f(z)$ and $F_h[p_y](y) = 0$, it follows that

$$|F_h[u](z)| \le Cht + |f(z) - f(y)| \le C_1 ht + C_2 h$$

where $t > 1$, C_1 depends on the constant in (7.8) and stencil size m, and C_2 depends on the Lipschitz constant of f. Thus, the consistency error is of the order $O(th)$ on Ω_h^I except for a set that has measure $O(t^{-\sigma})$. To be more precise, we introduce the discrete regular and singular sets

$$\begin{aligned} \mathcal{R}_{t,h} &:= \{z \in \Omega_h^{I,\tau} : B_h(y_*(z)) \cap \mathcal{R}_t \neq \emptyset\}, \\ \mathcal{S}_{t,h} &:= \{z \in \Omega_h^{I,\tau} : z \notin \mathcal{R}_{t,h}\}. \end{aligned} \tag{7.9}$$

We now estimate the size of the discrete singular set.

Lemma 7.14 (cardinality of $\mathcal{S}_{t,h}$). Let $\mathcal{S}_{t,h}$ be defined by (7.9). Then

$$\#\mathcal{S}_{t,h} \le C(h^{-d} t^{-\sigma} + h^{1-d}).$$

Proof. Note that

$$Ch^d(\#\mathcal{S}_{t,h}) \le \sum_{z \in \mathcal{S}_{t,h}} |B_h(y_*(z))| \le C \left| \bigcup_{z \in \mathcal{S}_{t,h}} B_h(z) \right| \le C(t^{-\sigma} + h).$$

Indeed, the first inequality is obvious. The second one follows from Proposition 7.11(iv), which shows that, for $z_1, z_2 \in \bar{\Omega}_h$ with $z_1 \neq z_2$, we have $|y_*(z_1) - y_*(z_2)| \ge Ch$, so the overlap of the balls can be controlled independently of h. Finally, the last inequality follows from Theorem 7.13. Rearranging terms in this last inequality yields the result. □

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

We now estimate the consistency error of $F_h[u_\tau^+]$ and have the following corollary of Proposition 7.2.

Corollary 7.15 (consistency). For every $z \in \Omega_h^{I,\tau}$ we have

$$F_h[u_\tau^+](z) \geq -C \begin{cases} (ht + h^\tau) & z \in \mathcal{R}_{t,h}, \\ h^{-\tau} & z \in \mathcal{S}_{t,h}. \end{cases}$$

Proof. For $z \in \mathcal{R}_{t,h}$, since $|z - y_*| \leq Ch^\tau$ and f is Lipschitz-continuous, we have by Proposition 7.2

$$F_h[u_\tau^+](z) \geq F_h[u](y_*) - Ch^\tau \geq -C(ht + h^\tau).$$

The estimate on the singular set $\mathcal{S}_{t,h}$ was already established in (7.7). □

The previous result controls the consistency error in the interior of the domain. To obtain a rate of convergence for the scheme, we also need to control this error near the boundary. This is accomplished with the help of the following discrete barrier function.

Lemma 7.16 (discrete barrier). Let the domain Ω satisfy an exterior ball condition. If h is sufficiently small, then for any point $z \in \Omega_h^I$ with $\operatorname{dist}(z, \partial\Omega) \leq \delta$, there exists a discrete function $b_{z,h} \in X_h^{fd}$ such that

$$F_h[b_{z,h}](\zeta) \leq 0, \quad \text{for all } \zeta \in \Omega_h^I, \qquad b_{z,h}(\zeta) \geq 0, \quad \text{for all } \zeta \in \Omega_h^B.$$

Moreover, we have

$$|b_{z,h}(z)| \leq C\delta,$$

where C depends only on λ, Λ, d and Ω.

Proof. Let $z_* \in \partial\Omega$ such that $|z - z_*| = \operatorname{dist}(z, \partial\Omega)$. Since Ω satisfies an exterior ball condition, there exists a open ball $B_r(z')$ centred at some point z' such that

$$z_* \in \partial B_r(z') \quad \text{and} \quad B_r(z') \cap \Omega = \emptyset.$$

We construct a barrier function

$$b_z(x) = B(r^{-q} - |x - z'|^{-q}),$$

where B and q are some positive constant to be determined later. We observe that

$$\begin{aligned} Db_z(x) =& qB|x - z'|^{-q-2}(x - z'), \\ D^2 b_z(x) =& qB|x - z'|^{-q-2}\left(I - (2+q)\frac{(x-z')}{|x-z'|} \otimes \frac{(x-z')}{|x-z'|}\right). \end{aligned}$$

Note that $D^2 b_z(x)$ is a diagonal perturbation of a rank-one matrix that has eigenvalues 0 (with multiplicity $d - 1$) and $-q(2 + q)B|x - z'|^{-q-2}$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

(with multiplicity 1). Therefore $D^2 b_z$ has eigenvalues $qB|x-z'|^{-q-2}$ (with multiplicity $d-1$) and $-q(1+q)B/|x-z'|^{q+2}$ (with multiplicity 1). In particular the smallest eigenvalue of $D^2 b_z(x)$ is $-q(1+q)B/|x-z'|^{q+2}$ and the largest is $qB|x-z'|^{-q-2}$.

Since $\lambda I \le A^{\alpha,\beta} \le \Lambda I$ we have

$$\inf_{\beta\in\mathcal{B}} \sup_{\alpha\in\mathcal{A}} \big[A^{\alpha,\beta} : D^2 b_z(x)\big] \le \frac{Bq}{|x-z'|^{q+2}}[(d-1)\Lambda - (1+q)\lambda].$$

Taking

$$q = 2(d-1)\frac{\Lambda}{\lambda} - 1,$$

we obtain that

$$\begin{aligned}\inf_{\beta\in\mathcal{B}} \sup_{\alpha\in\mathcal{A}} \big[A^{\alpha,\beta} : D^2 b_z(x)\big] &\le \frac{Bq}{|x-z'|^{q+2}}(1-d)\Lambda \\ &\le \frac{Bq}{|r+\operatorname{diam}(\Omega)|^{q+2}}(1-d)\Lambda.\end{aligned}$$

Since the right-hand side of the last inequality is negative, we conclude that, for B sufficiently large,

$$\inf_{\beta\in\mathcal{B}} \sup_{\alpha\in\mathcal{A}} \big[A^{\alpha,\beta} : D^2 b_z(x)\big] \le -2\|f\|_{L^\infty(\Omega)}.$$

The discrete barrier function is then defined as $b_{h,z} = I_h^{fd} b_z$. It is easy to check that

$$b_{h,z}(\zeta) \ge 0, \quad \text{for all } \zeta \in \Omega_h^B, \quad |b_{h,z}(z)| \le C\delta,$$

so that it remains to verify that $F_h[b_h](\zeta) \le 0$ for all $\zeta \in \Omega_h^I$.

Since b_z is a smooth function, the consistency condition (7.5) implies

$$F_h[b_{h,z}](\zeta) - F(\zeta, D^2 b_z)(\zeta) = O(h).$$

Finally, because $F(\zeta, D^2 b_z)(\zeta) \le -\|f\|_{L^\infty(\Omega)}$ we deduce that, for h sufficiently small,

$$F_h[b_{h,z}](\zeta) \le 0, \quad \text{for all } \zeta \in \Omega_h^I.$$

This completes the proof. □

Remark 7.17 (bound on u_h). Corollary 7.15 implies that for any $z \in \Omega_h^I$ with $\operatorname{dist}(z, \partial\Omega) \le \delta$, we have

$$|u_h(z)| \le C\delta.$$

Indeed, since $b_{z,h} \ge u_h = 0$ on Ω_h^B and $F_h[b_{z,h}](z) \le F_h[u_h](z)$ for all $z \in \Omega_h^I$, invoking a comparison principle for F_h, we obtain

$$u_h(z) \le b_{z,h}(z) \le C\delta.$$

Similarly we can show that $u_h(z) \ge -C\delta$.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Combining Corollary 7.15 and Lemma 7.14, we obtain a rate of convergence for the finite difference scheme (7.2).

Theorem 7.18 (rate of convergence). Let the domain Ω satisfy an exterior ball condition and $f \in C^{0,1}(\bar{\Omega})$. Let $u_h \in X_h^{fd}$ be the solution to (7.2) and $u \in C^{0,1}(\bar{\Omega})$ the viscosity solution to (7.1). Then we obtain

$$\|I_h^{fd} u - u_h\|_{L^\infty(\bar{\Omega}_h)} \leq C h^{\sigma/(2d+\sigma)},$$

where σ is the exponent in Theorem 7.13.

Proof. By Proposition 7.11(ii) we have that $\|u - u_\tau^+\|_{L^\infty(\Omega)} \leq C h^\tau$. Thus it is sufficient to bound $v_h = u_h - I_h^{fd} u_\tau^+$. We shall show that $\sup_{\bar{\Omega}_h} v_h^- \leq C h^{\sigma/(2d+\sigma)}$; the proof of the complementary estimate is similar. We divide the proof into two steps.

Step 1 (*boundary estimate*). We first show that, for any $z \in \bar{\Omega}_h \backslash \Omega_h^{I,\tau}$,

$$|v_h(z)| \leq C h^\tau$$

for some constant $C > 0$. By the definition of v_h, we have

$$\begin{aligned} |v_h(z)| &\leq |u_h(z) - u(z)| + |u(z) - I_h^{fd} u_\tau^+(z)| \\ &\leq |u_h(z)| + |u(z)| + |u(z) - I_h^{fd} u_\tau^+(z)|. \end{aligned}$$

Since u is Lipschitz-continuous and $u = 0$ on $\partial\Omega$, we have that $|u(z)| \leq C h^\tau$. In addition, owing to Remark 7.17 and Proposition 7.11(ii), we conclude that

$$|v_h(z)| \leq C h^\tau.$$

Step 2 (*interior estimate*). Now

$$-F_h[u_\tau^+](z) = F_h[u_h](z) - F_h[u_\tau^+](z) \geq \mathcal{L}_h(u_h - I_h^{fd} u_\tau^+)(z) = \mathcal{L}_h v_h(z),$$

where $\mathcal{L}_h$ is the linear(ized) operator (7.3) with coefficients that depend on u_τ^+ and u_h. Thus, setting $t = h^{\tau-1}$, we have by Corollary 7.15

$$\mathcal{L}_h v_h(z) \leq C \begin{cases} h^\tau & z \in \mathcal{R}_{h^{\tau-1},h}, \\ h^{-\tau} & z \in \mathcal{S}_{h^{\tau-1},h}, \end{cases} \tag{7.10}$$

and, by Lemma 7.14,

$$\#\mathcal{S}_{h^{\tau-1},h} \leq C(h^{-d+\sigma(1-\tau)} + h^{1-d}) \leq C h^{-d+\sigma(1-\tau)}.$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Applying the ABP estimate given in Theorem 3.25 to (7.10) (with Ω_h^I replaced with $\Omega_h^{I,\tau}$) yields

$$\sup_{\Omega_h^{I,\tau}} v_h^- \le C\Bigg(\sum_{z\in\mathcal{R}_{h^{\tau-1},h}} h^{d+d\tau} + \sum_{z\in\mathcal{S}_{h^{\tau-1},h}} h^{d-d\tau}\Bigg)^{1/d} \le C\big(h^{d\tau} + h^{-d\tau+\sigma(1-\tau)}\big)^{1/d}.$$

Thus, setting τ so that $d\tau = -d\tau + \sigma(1-\tau)$, that is, $\tau = \sigma/(2d+\sigma)$, we obtain

$$\sup_{\Omega_h^{I,\tau}} v_h^- \le Ch^{\sigma/(2d+\sigma)},$$

and therefore

$$\sup_{\Omega_h^{I,\tau}} (u_h - I_h^{fd} u_\tau^+)^- \le Ch^{\sigma/(2d+\sigma)}.$$

Gathering the obtained bounds implies the result. □

Remark 7.19 (extensions). The works by Krylov (2015) and Turanova (2015) have obtained rates of convergence for finite difference schemes for non-convex PDEs with variable coefficients and low-order terms.

7.3. Finite element methods

Now, following Salgado and Zhang (2016), let us focus on the development of a finite element scheme for (7.1) based on the integro-differential approximation presented in Section 4.3. The idea is that after choosing an $\epsilon > 0$, we replace each of the operators $\mathcal{L}^{\alpha,\beta}$ with their integro-differential approximations $\mathcal{L}_\epsilon^{\alpha,\beta}$, as defined in (4.27). In so doing we obtain – according to Caffarelli and Silvestre (2010*b*) – a smooth approximation u^ϵ of u, the solution to (7.1). Moreover, the difference $u - u^\epsilon$ can be controlled in terms of ϵ. We can now proceed to approximate u^ϵ by simply taking the inf-sup over discrete problems of the form (4.28), thus obtaining a discrete function $u_h^\epsilon \in X_{0,h}^l$. Using the regularity of u^ϵ, we can control the difference $u^\epsilon - u_h^\epsilon$ in terms of the mesh size and, possibly, negative powers of ϵ. Optimizing with respect to ϵ yields a rate of convergence.

Let us now proceed with the details. To set ideas we will assume that for every $\alpha \in \mathcal{A}$ and $\beta \in \mathcal{B}$ we have $f^{\alpha,\beta} = f \in C^{0,1}(\bar{\Omega})$ and that the matrices $A^{\alpha,\beta}$ are constant. The approximate problem is then

$$\begin{cases} \dfrac{\lambda}{2}\Delta u^\epsilon + \inf\limits_{\beta\in\mathcal{B}} \sup\limits_{\alpha\in\mathcal{A}} I_\epsilon^{\alpha,\beta} u^\epsilon = f & \text{in } \Omega_\epsilon, \\ u = 0 & \text{on } \omega_\epsilon, \end{cases} \tag{7.11}$$

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

where, in analogy to (4.26), the integral operators are defined by

$$I_\epsilon^{\alpha,\beta} w(x) = \frac{1}{\epsilon^{d+2} \det A_\lambda^{\alpha,\beta}} \int_{\mathbb{R}^d} |y|^2 \delta_{\theta y,\theta}^2 w(x) \varphi\left(\frac{1}{\epsilon}(A_\lambda^{\alpha,\beta})^{-1} y\right),$$

with

$$A_\lambda^{\alpha,\beta} = \left(A^{\alpha,\beta} - \frac{\lambda}{2} I\right)^{1/2},$$

and the domains Ω_ϵ and ω_ϵ are given by (4.25).

Let us now state the existence, uniqueness and smoothness properties of u^ϵ and its rate of convergence to u. For a proof see Theorem 3.5, Theorem 4.8 and Theorem 6.1 of Caffarelli and Silvestre (2010*b*).

Proposition 7.20 (properties of u^ϵ). Problem (7.11) has a unique classical solution u^ϵ. Moreover, there exists $s \in (0,1)$ that depends on λ, Λ and d but not on ϵ such that, for every $\omega \Subset \Omega$, we have

$$\|u^\epsilon\|_{C^{1,s}(\omega)} + \|u^\epsilon\|_{C^{0,1}(\bar{\Omega})} \leq C\big(\|u^\epsilon\|_{L^\infty(\Omega)} + \|f\|_{L^\infty(\Omega)}\big),$$

where the constant $C > 0$ depends on the distance between ω and $\partial\Omega$. Additionally, there exists a $\gamma > 0$ such that

$$\|u - u^\epsilon\|_{L^\infty(\Omega)} \leq C\epsilon^\gamma \|f\|_{C^{0,1}(\Omega)},$$

where the constant $C > 0$ depends only on λ, Λ, d and Ω.

With the properties of u^ϵ at hand we introduce a finite element scheme to approximate it. Namely, given a quasi-uniform triangulation for which (3.42) holds, we seek $u_h^\varepsilon \in X_{0,h}^l$ such that (compare with (4.28))

$$F_h^\epsilon[u_h^\epsilon](z_i) := \frac{\lambda}{2}\Delta_h u_h^\epsilon(z_i) + \inf_{\beta\in\mathcal{B}} \sup_{\alpha\in\mathcal{A}} I_\epsilon^{\alpha,\beta} u_h^\epsilon(z_i) - f_i = 0, \quad \text{for all } z_i \in \Omega_h^I, \tag{7.12}$$

where, as in (4.28), $f_i = \int_\Omega f\phi_i$ and ϕ_i are the normalized hat functions.

Let us now address the existence, uniqueness and approximation properties of u_h^ϵ. We begin by noting that, since each one of the operators

$$w_h \mapsto \frac{\lambda}{2}\Delta_h w_h + I_\epsilon^{\alpha,\beta} w_h$$

is monotone (see Lemma 4.19), the same holds for F_h^ϵ. This is the content of the next result.

Corollary 7.21 (monotonicity). Assume that $\mathcal{T}_h$ is such that Δ_h is monotone. Then the operator F_h^ϵ, defined in (7.12) is monotone.

Remark 7.22 (flexibility). Note that, in Corollary 7.21, all that is necessary is a comparison principle *for the finite element Laplacian* Δ_h. A sufficient condition for this is given in Lemma 3.42. This is in stark contrast

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

with the methods of Section 5.2, where either a restrictive mesh condition is needed (see Remark 5.8) or the coefficients are assumed to be isotropic as in (5.16).

In addition to monotonicity we must also consider the consistency of the scheme which, owing to the consistency of Δ_h (see Lemma 3.41), reduces to the study of the consistency of the inf-sup of integral transforms when applied to Lipschitz functions (recall that $u^\epsilon \in C^{0,1}(\bar{\Omega})$ uniformly in ϵ). To measure this we define, for $z_i \in \Omega_h^I$,

$$\mathcal{R}_{h,\epsilon}[w](z_i) = \int_\Omega \left[\inf_{\beta\in\mathcal{B}} \sup_{\alpha\in\mathcal{A}} I_\epsilon^{\alpha,\beta} I_h^{ep} w(z_i) - \inf_{\beta\in\mathcal{B}} \sup_{\alpha\in\mathcal{A}} I_\epsilon^{\alpha,\beta} w(x) \right] \phi_i(x).$$

The consistency of the scheme is then encoded in the following result, whose proof mainly follows Lemma 4.18 and Theorem 4.23 but must take into account the reduced regularity of the functions. This is the reason for the reduced rate.

Lemma 7.23 (consistency). Let $w \in C^{0,1}(\bar{\Omega})$. Then for every $z \in \Omega_h^I$ we have

$$|\mathcal{R}_{h,\epsilon}[w](z)| \le C \frac{h}{\epsilon^2} |\log h| \|w\|_{C^{0,1}(\bar{\Omega})},$$

where the constant $C > 0$ is independent of z, h, ϵ and w.

Monotonicity and consistency allow us to conclude existence and uniqueness of solutions. The proof of the following result is either a corollary of Proposition 7.27 below or follows from a discrete version of Perron's method.

Theorem 7.24 (existence and uniqueness). Let the family of meshes be such that Δ_h is monotone. Then, for every $h > 0$ and $\epsilon > 0$, the scheme (7.12) has a unique solution.

To conclude, we study the rates of convergence. Since Proposition 7.20 provides a rate for $\|u - u^\epsilon\|_{L^\infty(\Omega)}$ and Proposition 3.40 provides one for $\|u^\epsilon - I_h^{ep} u^\epsilon\|_{L^\infty(\Omega)}$, it remains to compare the discrete solution u_h^ϵ to the Galerkin projection $I_h^{ep} u^\epsilon$. Upon denoting $e_h = I_h^{ep} u^\epsilon - u_h^\epsilon \in X_{0,h}^l$, and after some manipulations, it turns out that the error e_h satisfies, for every $z \in \Omega_h^I$,

$$\frac{\lambda}{2} \Delta_h e_h(z) + \inf_{\beta\in\mathcal{B}} \sup_{\alpha\in\mathcal{A}} I_\epsilon^{\alpha,\beta} I_h^{ep} u^\epsilon(z) - \inf_{\beta\in\mathcal{B}} \sup_{\alpha\in\mathcal{A}} I_\epsilon^{\alpha,\beta} u_h^\epsilon(z) = \mathcal{R}_{h,\epsilon}[u^\epsilon](z). \quad (7.13)$$

With this equation at hand we are ready to establish a rate of convergence.

Theorem 7.25 (rate of convergence). Let $u \in C^{1,s}(\Omega) \cap C^{0,1}(\bar{\Omega})$ be the solution of problem (7.1) and let $u_h^\epsilon \in X_{0,h}^l$ be the solution to scheme

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

(7.12) with $\epsilon \geq Ch^{1/2}\,|\log h|$. There exists $\gamma > 0$ such that

$$\|u - u_h^\epsilon\|_{L^\infty(\Omega)} \leq C\left(\epsilon^\gamma + \frac{h}{\epsilon^2}\,|\log h|\right)\|f\|_{C^{0,1}(\bar{\Omega})},$$

where the constant $C > 0$ depends only on λ, Λ, Ω and d.

Proof. As already discussed, the rates of convergence reduce to estimating the difference e_h. To do so we employ the error equation (7.13) and note that, if z belongs to $\mathcal{C}_h^-(e_h)$, the contact set of e_h, then we must have

$$\inf_{\beta\in\mathcal{B}}\sup_{\alpha\in\mathcal{A}} I_\epsilon^{\alpha,\beta} I_h^{ep} u^\epsilon(z) \geq \inf_{\beta\in\mathcal{B}}\sup_{\alpha\in\mathcal{A}} I_h^{\alpha,\beta} u_h^\epsilon(z).$$

In other words, for $z \in \mathcal{C}_h^-(e_h)$, the error e_h satisfies the inequality

$$\frac{\lambda}{2}\Delta_h e_h(z) \leq \mathcal{R}_{h,\epsilon}[u^\epsilon](z).$$

The discrete ABP estimate of Theorem 4.20 then implies that

$$\sup_\Omega e_h^- \leq C\left(\sum_{z\in\mathcal{C}_h^-(e_h)} |\omega_z|\,|\mathcal{R}_{h,\epsilon}[u^\epsilon](z)|^d\right)^{1/d}.$$

We must now invoke the consistency estimates of Lemma 7.23 to obtain an upper bound for e_h^-. A similar argument will yield the reverse bound. The theorem is thus proved. □

Remark 7.26 (algebraic rate). If the actual value of $\gamma > 0$ is known, then one can choose $\epsilon^{\gamma+2} = h\,|\log h|$ to obtain

$$\|u - u_h^\epsilon\|_{L^\infty(\Omega)} \leq Ch^{\gamma/(\gamma+2)}\,|\log h|^{\gamma/(\gamma+2)},$$

which in the best scenario, $\gamma = 1$, would yield $O(h^{1/3}\,|\log h|^{1/3})$.

7.4. Solution of the discrete problems

Having discussed discretization schemes for Isaacs equations (7.1) and detailed their convergence properties, to finalize our discussion we will focus on how to solve the ensuing nonlinear systems of equations. We will see – as should by now be clear to the reader – that the approaches are extensions of the convex case, which we discussed in Section 5.3, but the results are much more modest.

After discretization, in complete analogy to (5.22), we must solve the following problem: find $\mathbf{x} \in \mathbb{R}^N$, such that

$$\mathbf{F}(\mathbf{x}) = \inf_{\beta\in\mathcal{B}}\sup_{\alpha\in\mathcal{A}}\left[\mathbf{K}^{\alpha,\beta}\mathbf{x} - \mathbf{f}^{\alpha,\beta}\right] = \mathbf{0}, \tag{7.14}$$

where the inf-sup is computed component-wise, $\{\mathbf{K}^{\alpha,\beta} : \alpha \in \mathcal{A}, \beta \in \mathcal{B}\}$ and

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

$\{\mathbf{f}^{\alpha,\beta} : \alpha \in \mathcal{A}, \beta \in \mathcal{B}\}$ are, respectively, discretizations of $\{\mathcal{L}^{\alpha,\beta} : \alpha \in \mathcal{A}, \beta \in \mathcal{B}\}$ and $\{f^{\alpha,\beta} : \alpha \in \mathcal{A}, \beta \in \mathcal{B}\}$. Once again, the dimension N equals the number of degrees of freedom in the discretization.

One would be tempted, following Algorithm 5.1, to define, for $\mathbf{y} \in \mathbb{R}^N$, the indices $\boldsymbol{\alpha}(\mathbf{y})$ and $\boldsymbol{\beta}(\mathbf{y})$ by conditions similar to those that led to (5.23) and use them in each iteration of a possible extension of Howard's algorithm. However, this will not produce a convergent method; see, for instance, Bokanowski *et al.* (2009, Remark 5.8). The reason for this, simply put, is that the map $\mathbf{y} \mapsto \min\{\mathbf{g}, \max\{\mathbf{y}, \mathbf{h}\}\}$ is not slantly differentiable and, moreover, since the map $\mathbf{F}$, defined in (7.14), is neither convex nor concave, any generalized notion of derivative for this function will not possess any monotonicity properties.

For the reasons outlined above, we now describe, following Bokanowski *et al.* (2009, Section 5), a convergent scheme that cannot be cast as a Newton-type method but can be understood as a two-level Howard's algorithm. We begin by defining, for $\mathbf{y} \in \mathbb{R}^N$ and $i \in \{1, \dots, N\}$, the element $\beta(\mathbf{y}, i) \in \mathcal{B}$ by the condition

$$\mathbf{F}(\mathbf{y})_i = \sup_{\alpha \in \mathcal{A}} \left[\mathbf{K}^{\alpha,\beta(\mathbf{y},i)}\mathbf{y} - \mathbf{f}^{\alpha,\beta(\mathbf{y},i)}\right]_i. \tag{7.15}$$

With this at hand, we define $\boldsymbol{\beta}(\mathbf{y}) \in \mathcal{B}^N$ by $\boldsymbol{\beta}(\mathbf{y})_i = \boldsymbol{\beta}(\mathbf{y}, i)$, and the mapping $\mathbf{F}^{\boldsymbol{\beta}(\mathbf{y})} : \mathbb{R}^N \to \mathbb{R}^N$ is such that

$$\mathbf{F}^{\boldsymbol{\beta}(\mathbf{y})}(\mathbf{w})_i = \sup_{\alpha \in \mathcal{A}} \left[\mathbf{K}^{\alpha,\beta(\mathbf{y},i)}\mathbf{w} - \mathbf{f}^{\alpha,\beta(\mathbf{y},i)}\right]_i, \quad i = 1, \dots, N. \tag{7.16}$$

The generalization of Howard's method is described in Algorithm 7.1.

Note that, owing to the definition of $\mathbf{F}^{\boldsymbol{\beta}(\mathbf{y})}$ given in (7.15)–(7.16), every step of Algorithm 7.1 requires the solution of a discrete Hamilton–Jacobi–Bellman equation (7.17). This can be done by applying Algorithm 5.1 and justifies calling this method a *two-level* one. The convergence properties of Algorithm 7.1 are described below.

Proposition 7.27 (convergence of two-level Howard). Let $\mathcal{A}$ and $\mathcal{B}$ be finite sets. Assume that, for every $\boldsymbol{\beta} \in \mathcal{B}^N$, the map $\mathbf{F}^{\boldsymbol{\beta}}$ is monotone in the sense of Definition 3.5. Then the sequence of iterates, given by Algorithm 7.1 satisfies $\mathbf{x}_k \geq \mathbf{x}_{k+1}$ and converges, in a finite number of steps, to $\mathbf{x}$, the solution of (7.14).

Proof. Using equations (7.15)–(7.16), line 3 of Algorithm 7.1 and identity (7.17), we observe that the iterates $\{\boldsymbol{\beta}_k\}_{k\in\mathbb{N}}$ and $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ satisfy

$$\mathbf{F}^{\boldsymbol{\beta}_{k+1}}(\mathbf{x}_k) = \mathbf{F}^{\boldsymbol{\beta}(\mathbf{x}_k)}(\mathbf{x}_k) = \mathbf{F}(\mathbf{x}_k) \leq \mathbf{F}^{\boldsymbol{\beta}_k}(\mathbf{x}_k) = \mathbf{0} = \mathbf{F}^{\boldsymbol{\beta}_{k+1}}(\mathbf{x}_{k+1}),$$

so that, from the monotonicity of $\mathbf{F}^{\boldsymbol{\beta}_{k+1}}$, we conclude that $\mathbf{x}_k \geq \mathbf{x}_{k+1}$.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Algorithm 7.1: Two-level Howard's algorithm for non-convex problems.

Input : Sets $\mathcal{A}$ and $\mathcal{B}$.
Matrices $\{\mathbf{K}^{\alpha,\beta} : \alpha \in \mathcal{A}, \beta \in \mathcal{B}\} \subset \mathbb{R}^{N\times N}$.
right-hand sides $\{\mathbf{f}^{\alpha,\beta} : \alpha \in \mathcal{A}, \beta \in \mathcal{B}\} \subset \mathbb{R}^{N}$.
Output: Vector $\mathbf{x} \in \mathbb{R}^N$, solution of (7.14).

Initialization: Choose $\mathbf{x}_{-1} \in \mathbb{R}^N$;
for $k \geq 0$ **do**
  Set $\boldsymbol{\beta}_k = \boldsymbol{\beta}(\mathbf{x}_{k-1})$;
  Find: $\mathbf{x}_k \in \mathbb{R}^N$ that solves

$$\mathbf{F}^{\boldsymbol{\beta}_k}(\mathbf{x}_k) = \mathbf{0} \tag{7.17}$$

  ;
  if $\mathbf{F}(\mathbf{x}_k) = \mathbf{0}$ **then**
    return $\mathbf{x}_k$;
  end
end

As in Theorem 5.31, the fact that $\#\mathcal{B}$ is finite together with the monotonicity of iterates imply that Algorithm 7.1 converges in a finite number of steps. □

We conclude the discussion on Algorithm 7.1 by commenting that the case when $\mathcal{A}$ and $\mathcal{B}$ are compact spaces and when equation (7.17) is only solved approximately are also discussed by Bokanowski *et al.* (2009).

In addition to Algorithm 7.1, we present the Richardson-type iterative scheme of Kuo and Trudinger (1992, Section 4.2). Starting from $\mathbf{x}_0 \in \mathbb{R}^N$, the iterates are computed via

$$\mathbf{x}_{k+1} = \mathbf{G}(\mathbf{x}_k) := \mathbf{x}_k - \frac{1}{\Lambda_N}\mathbf{F}(\mathbf{x}_k), \tag{7.18}$$

where $\Lambda_N > 0$ is a sufficiently large constant that depends on N. It is shown there that, for $\Lambda_N \geq CN^{2/d}$, the mapping $\mathbf{G}$ is a contraction, that is, there is a constant $C > 0$ such that

$$|\mathbf{G}(\mathbf{v}) - \mathbf{G}(\mathbf{w})| \leq \left(1 - \frac{N^{-2/d}}{C}\right)|\mathbf{v} - \mathbf{w}|,$$

so that (7.18) is convergent. Some strategies that combine Algorithm 7.1 and (7.18) are suggested in González and Tidball (1992).

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

8. Outlook

Overall we know too much about linear PDE and too little about nonlinear PDE.

L. C. Evans

In this paper we have summarized some of the recent trends and advances in the discretizations and convergence analysis for strongly nonlinear PDEs, with an emphasis on convex and non-convex fully nonlinear equations. While these two classes of equations and discretization types have fundamentally different structure conditions, common themes permeate the analysis; for example, consistency, monotonicity and comparison principles, Alexandrov–Bakelman–Pucci estimates, wide stencils and smooth approximations. It is also evident from our discussion that, due to the pointwise definition of viscosity solutions and, correspondingly, the monotonicity criterion given in the Barles–Souganidis framework, there is a stark contrast between finite difference schemes and finite element methods for such problems; only within the last 10 years have *any* significant advances been made in the convergence analysis of finite element-type methods.

Despite the recent flurry of results for numerical fully nonlinear PDEs, there still remain fundamental open problems in the field. Probably the most pressing, at least on the finite element front, is an alternative framework that bypasses or relaxes the monotonicity requirement found in the Barles–Souganidis theory. Other completely open problems include, but are not limited to:

(i) derivation of *sharp* rates of convergence for non-convex fully nonlinear equations,

(ii) rates of convergence for finite difference schemes on *unstructured* grids,

(iii) rates of convergence of finite element methods for convex and non-convex PDEs in other norms, for example Sobolev norms such as H^1,

(iv) *a posteriori* error estimation and adaptive methods for fully nonlinear problems,

(v) rates of convergence for non-convex degenerate problems.

In light of these questions we feel that the quote from the preface of Evans (2010) given above nicely summarizes the current state of numerical PDEs. We hope that this overview will motivate current and future researchers to work on the numerical approximation of nonlinear problems.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

Acknowledgements

The authors are grateful that they were given the opportunity to write this review.

The first two authors were supported in part by the US National Science Foundation grants DMS-1417980 (Neilan) and DMS-1418784 (Salgado), and the Alfred Sloan Foundation (Neilan). This work was initiated at the conference 'Nonlinear PDEs, Numerical Analysis, and Applications' at the University of Pittsburgh, which was supported by NSF grant DMS-1541585, an IMA conference grant, and the Mathematical Research Center, University of Pittsburgh.

REFERENCES[1]

A. Alla, M. Falcone and D. Kalise (2015), 'An efficient policy iteration algorithm for dynamic programming equations', *SIAM J. Sci. Comput.* **37**, A181–A200.

G. Awanou (2015*a*), 'Pseudo transient continuation and time marching methods for Monge–Ampère type equations', *Adv. Comput. Math.* **41**, 907–935.

G. Awanou (2015*b*), 'Quadratic mixed finite element approximations of the Monge–Ampère equation in 2D', *Calcolo* **52**, 503–518.

G. Awanou (2015*c*), 'Standard finite elements for the numerical resolution of the elliptic Monge–Ampère equations: Classical solutions', *IMA J. Numer. Anal.* **35**, 1150–1166.

C. Baiocchi (1977), Estimations d'erreur dans L^∞ pour les inéquations à obstacle. In *Mathematical Aspects of Finite Element Methods: Proc. Conf., Consiglio Naz. delle Ricerche (C.N.R.), Rome, 1975*, Vol. 606 of *Lecture Notes in Mathematics*, Springer, pp. 27–34.

G. Barles and E. Jakobsen (2005), 'Error bounds for monotone approximation schemes for Hamilton–Jacobi–Bellman equations', *SIAM J. Numer. Anal.* **43**, 540–558.

G. Barles and E. Jakobsen (2002), 'On the convergence rate of approximation schemes for Hamilton–Jacobi–Bellman equations', *M2AN Math. Model. Numer. Anal.* **36**, 33–54.

G. Barles and P. Souganidis (1991), 'Convergence of approximation schemes for fully nonlinear second order equations', *Asymptotic Anal.* **4**, 271–283.

R. Bellman (2010), *Dynamic Programming*, Princeton Landmarks in Mathematics, Princeton University Press. Reprint of the 1957 edition.

J.-D. Benamou and Y. Brenier (1998), 'Weak existence for the semigeostrophic equations formulated as a coupled Monge–Ampère/transport problem', *SIAM J. Appl. Math.* **58**, 1450–1461.

J.-D. Benamou and Y. Brenier (2000), 'A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem', *Numer. Math.* **84**, 375–393.

[1] The URLs cited in this work were correct at the time of going to press, but the publisher and the authors make no undertaking that the citations remain live or are accurate or appropriate.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

J.-D. Benamou, F. Collino and J.-M. Mirebeau (2016), 'Monotone and consistent discretization of the Monge–Ampère operator', *Math. Comp.* **85**(302), 2743–2775.

J.-D. Benamou, B. Froese and A. Oberman (2014), 'Numerical solution of the optimal transportation problem using the Monge–Ampère equation', *J. Comput. Phys.* **260**, 107–126.

A. Bensoussan and J.-L. Lions (1984), *Impulse Control and Quasivariational Inequalities* (translated from the French by J. M. Cole), Gauthier-Villars.

K. Böhmer (2008), 'On finite element methods for fully nonlinear elliptic equations of second order', *SIAM J. Numer. Anal.* **46**, 1212–1249.

O. Bokanowski, S. Maroso and H. Zidani (2009), 'Some convergence results for Howard's algorithm', *SIAM J. Numer. Anal.* **47**, 3001–3026.

M. Boulbrachene and P. Cortey-Dumont (2009), 'Optimal L^∞-error estimate of a finite element method for Hamilton–Jacobi–Bellman equations', *Numer. Funct. Anal. Optim.* **30**, 421–435.

M. Boulbrachene and M. Haiour (2001), 'The finite element approximation of Hamilton–Jacobi–Bellman equations', *Comput. Math. Appl.* **41**, 993–1007.

S. Brenner and M. Neilan (2012), 'Finite element approximations of the three dimensional Monge–Ampère equation', *ESAIM Math. Model. Numer. Anal.* **46**, 979–1001.

S. Brenner and L. Scott (2008), *The Mathematical Theory of Finite Element Methods*, third edition, Vol. 15 of *Texts in Applied Mathematics*, Springer.

S. Brenner, T. Gudi, M. Neilan and L.-Y. Sung (2011), 'C^0 penalty methods for the fully nonlinear Monge–Ampère equation', *Math. Comp.* **80**(276), 1979–1995.

L. Caffarelli and X. Cabré (1995), *Fully Nonlinear Elliptic Equations*, Vol. 43 of *American Mathematical Society Colloquium Publications*, AMS.

L. Caffarelli and L. Silvestre (2010*a*), 'On the Evans–Krylov theorem', *Proc. Amer. Math. Soc.* **138**, 263–265.

L. Caffarelli and L. Silvestre (2010*b*), Smooth approximations of solutions to nonconvex fully nonlinear elliptic equations. In *Nonlinear Partial Differential Equations and Related Topics*, Vol. 229 of *American Mathematical Society Translations, Series 2*, AMS, pp. 67–85.

L. Caffarelli and P. Souganidis (2008), 'A rate of convergence for monotone finite difference approximations to fully nonlinear, uniformly elliptic PDEs', *Comm. Pure Appl. Math.* **61**, 1–17.

A. Calderón and A. Zygmund (1952), 'On the existence of certain singular integrals', *Acta Math.* **88**, 85–139.

J. Carnicer and W. Dahmen (1994), 'Characterization of local strict convexity preserving interpolation methods by C^1 functions', *J. Approx. Theory* **77**, 2–30.

P. Ciarlet (2002), *The Finite Element Method for Elliptic Problems*, Vol. 40 of *Classics in Applied Mathematics*, SIAM. Reprint of the 1978 original.

P. Cortey-Dumont (1987), 'Sur l'analyse numérique des équations de Hamilton–Jacobi–Bellman', *Math. Methods Appl. Sci.* **9**, 198–209.

M. Crandall and P.-L. Lions (1983), 'Viscosity solutions of Hamilton–Jacobi equations', *Trans. Amer. Math. Soc.* **277**, 1–42.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

M. Crandall, H. Ishii and P.-L. Lions (1992), 'User's guide to viscosity solutions of second order partial differential equations', *Bull. Amer. Math. Soc.* (N.S.) **27**, 1–67.

B. Dacorogna (2008), *Direct Methods in the Calculus of Variations*, second edition, Vol. 78 of *Applied Mathematical Sciences*, Springer.

E. Dean and R. Glowinski (2006*a*), '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**, 71–96.

E. Dean and R. Glowinski (2006*b*), 'Numerical methods for fully nonlinear elliptic equations of the Monge–Ampère type', *Comput. Methods Appl. Mech. Engrg* **195**, 1344–1386.

K. Debrabant and E. Jakobsen (2013), 'Semi-Lagrangian schemes for linear and fully non-linear diffusion equations', *Math. Comp.* **82**(283), 1433–1462.

H. Dong and N. Krylov (2007), 'The rate of convergence of finite-difference approximations for parabolic Bellman equations with Lipschitz coefficients in cylindrical domains', *Appl. Math. Optim.* **56**, 37–66.

I. Ekeland and R. Témam (1999), *Convex Analysis and Variational Problems*, English edition, Vol. 28 of *Classics in Applied Mathematics*, SIAM.

A. Ern and J.-L. Guermond (2004), *Theory and Practice of Finite Elements*, Vol. 159 of *Applied Mathematical Sciences*, Springer.

L. Evans (1980), 'On solving certain nonlinear partial differential equations by accretive operator methods', *Israel J. Math.* **36**, 225–247.

L. Evans (2010), *Partial Differential Equations*, second edition, Vol. 19 of *Graduate Studies in Mathematics*, AMS.

L. Evans and A. Friedman (1979), 'Optimal stochastic switching and the Dirichlet problem for the Bellman equation', *Trans. Amer. Math. Soc.* **253**, 365–389.

X. Feng and M. Jensen (2017), 'Convergent semi-Lagrangian methods for the Monge–Ampère equation on unstructured grids', *SIAM J. Numer. Anal.*, to appear.

X. Feng and T. Lewis (2014), 'Local discontinuous Galerkin methods for one-dimensional second order fully nonlinear elliptic and parabolic equations', *J. Sci. Comput.* **59**, 129–157.

X. Feng and M. Neilan (2009), 'Mixed finite element methods for the fully nonlinear Monge–Ampère equation based on the vanishing moment method', *SIAM J. Numer. Anal.* **47**, 1226–1250.

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

X. Feng, R. Glowinski and M. Neilan (2013), 'Recent developments in numerical methods for fully nonlinear second order partial differential equations', *SIAM Rev.* **55**, 205–267.

X. Feng, L. Hennings and M. Neilan (2017), 'C^0 discontinuous Galerkin finite element methods for second order linear elliptic partial differential equations in non-divergence form', *Math. Comp.*, to appear.

X. Feng, M. Neilan and S. Schnake (2016), Interior penalty discontinuous Galerkin methods for second order linear non-divergence form elliptic PDEs. arXiv:1605.04364

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

W. Fleming and H. Soner (2006), *Controlled Markov Processes and Viscosity Solutions*, second edition, Vol. 25 of *Stochastic Modelling and Applied Probability*, Springer.

W. Fleming and P. Souganidis (1989), 'On the existence of value functions of two-player, zero-sum stochastic differential games', *Indiana Univ. Math. J.* **38**, 293–314.

B. Froese (2016), Convergent approximation of surfaces of prescribed Gaussian curvature with weak Dirichlet conditions. arXiv:1601.06315

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

D. Gallistl (2017), 'Variational formulation and numerical analysis of linear elliptic equations in nondivergence form with Cordes coefficients', *SIAM J. Numer. Anal.*, to appear.

E. Georgoulis, P. Houston and J. Virtanen (2011), 'An *a posteriori* error indicator for discontinuous Galerkin approximations of fourth-order elliptic problems', *IMA J. Numer. Anal.* **31**, 281–298.

S. Gerschgorin (1930), 'Fehlerabschätzung für das Differenzenverfahren zur Lösung partieller Differentialgleichungen', *Z. Angew. Math. Mech.* **10**, 373–382.

D. Gilbarg and N. Trudinger (2001), *Elliptic Partial Differential Equations of Second Order*, Classics in Mathematics, Springer. Reprint of the 1998 edition.

R. González and M. Tidball (1992), Fast solution of general nonlinear fixed point problems. In *System Modelling and Optimization (Zurich, 1991)*, Vol. 180 of *Lecture Notes in Control and Information Sciences*, Springer, pp. 35–44.

P. Grisvard (2011), *Elliptic Problems in Nonsmooth Domains*, Vol. 69 of *Classics in Applied Mathematics*, SIAM. Reprint of the 1985 original.

C. Gutiérrez (2001), *The Monge–Ampère Equation*, Vol. 44 of *Progress in Nonlinear Differential Equations and their Applications*, Birkhäuser.

D. Han and J. Wan (2013), 'Multigrid methods for second order Hamilton–Jacobi–Bellman and Hamilton–Jacobi–Bellman–Isaacs equations', *SIAM J. Sci. Comput.* **35**, S323–S344.

Q. Han and F. Lin (2011), *Elliptic Partial Differential Equations*, second edition, Vol. 1 of *Courant Lecture Notes in Mathematics*, AMS.

M. Hintermüller, K. Ito and K. Kunisch (2002), 'The primal–dual active set strategy as a semismooth Newton method', *SIAM J. Optim.* **13**, 865–888.

H. Holden and N. Risebro (2015), *Front Tracking for Hyperbolic Conservation Laws*, second edition, Vol. 152 of *Applied Mathematical Sciences*, Springer.

R. Hoppe (1986), 'Multigrid methods for Hamilton–Jacobi–Bellman equations', *Numer. Math.* **49**, 239–254.

A. Iserles (2009), *A First Course in the Numerical Analysis of Differential Equations*, second edition, Cambridge Texts in Applied Mathematics, Cambridge University Press.

H. Ishii (1995), 'On the equivalence of two notions of weak solutions, viscosity solutions and distribution solutions', *Funkcial. Ekvac.* **38**, 101–120.

E. Jakobsen (2004), 'On error bounds for approximation schemes for non-convex degenerate elliptic equations', *BIT* **44**, 269–285.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

E. Jakobsen (2006), 'On error bounds for monotone approximation schemes for multi-dimensional Isaacs equations', *Asymptot. Anal.* **49**, 249–273.

M. Jensen and I. Smears (2013*a*), Finite element methods with artificial diffusion for Hamilton–Jacobi–Bellman equations. In *Numerical Mathematics and Advanced Applications 2011* (A. Cangiani *et al.*, eds), Springer, pp. 267–274.

M. Jensen and I. Smears (2013*b*), 'On the convergence of finite element methods for Hamilton–Jacobi–Bellman equations', *SIAM J. Numer. Anal.* **51**, 137–162.

M. Katsoulakis (1995), 'A representation formula and regularizing properties for viscosity solutions of second-order fully nonlinear degenerate parabolic equations', *Nonlinear Anal.* **24**, 147–158.

N. Katzourakis (2015), *An Introduction to Viscosity Solutions for Fully Nonlinear PDE with Applications to Calculus of Variations in* L^∞, Springer Briefs in Mathematics, Springer.

B. Kawohl and N. Kutev (2007), 'Comparison principle for viscosity solutions of fully nonlinear, degenerate elliptic equations', *Comm. Partial Diff. Equations* **32**, 1209–1224.

M. Kocan (1995), 'Approximation of viscosity solutions of elliptic partial differential equations on minimal grids', *Numer. Math.* **72**, 73–92.

S. Koike (2004), *A Beginner's Guide to the Theory of Viscosity Solutions*, Vol. 13 of *MSJ Memoirs*, Mathematical Society of Japan.

S. Koike and A. Świech (2013), 'Representation formulas for solutions of Isaacs integro-PDE', *Indiana Univ. Math. J.* **62**, 1473–1502.

I. Kossaczký, M. Ehrhardt and M. Günther (2016), 'On the non-existence of higher order monotone approximation schemes for HJB equations', *Appl. Math. Lett.* **52**, 53–57.

N. Krylov (1987), *Nonlinear Elliptic and Parabolic Equations of the Second Order* (translated from the Russian by P. L. Buzytsky), Vol. 7 of *Mathematics and its Applications (Soviet Series)*, Reidel.

N. Krylov (1996), *Lectures on Elliptic and Parabolic Equations in Hölder Spaces*, Vol. 12 of *Graduate Studies in Mathematics*, AMS.

N. Krylov (1997), 'On the rate of convergence of finite-difference approximations for Bellman's equations', *Algebra i Analiz* **9**, 245–256.

N. Krylov (2000), 'On the rate of convergence of finite-difference approximations for Bellman's equations with variable coefficients', *Probab. Theory Rel. Fields* **117**, 1–16.

N. Krylov (2005), 'The rate of convergence of finite-difference approximations for Bellman equations with Lipschitz coefficients', *Appl. Math. Optim.* **52**, 365–399.

N. Krylov (2008), *Lectures on Elliptic and Parabolic Equations in Sobolev Spaces*, Vol. 96 of *Graduate Studies in Mathematics*, AMS.

N. Krylov (2014), 'On the rate of convergence of difference approximations for uniformly nondegenerate elliptic Bellman's equations', *Appl. Math. Optim.* **69**, 431–458.

N. Krylov (2015), 'On the rate of convergence of finite-difference approximations for elliptic Isaacs equations in smooth domains', *Comm. Partial Diff. Equations* **40**, 1393–1407.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

H.-J. Kuo and N. Trudinger (1990), 'Linear elliptic difference inequalities with random coefficients', *Math. Comp.* **55**(191), 37–53.
H.-J. Kuo and N. Trudinger (1992), 'Discrete methods for fully nonlinear elliptic equations', *SIAM J. Numer. Anal.* **29**, 123–135.
H.-J. Kuo and N. Trudinger (1996), 'Positive difference operators on general meshes', *Duke Math. J.* **83**, 415–433.
H.-J. Kuo and N. Trudinger (2000), 'A note on the discrete Aleksandrov–Bakelman maximum principle', *Taiwanese J. Math.* **4**, 55–64.
M.-J. Lai and L. Schumaker (2007), 'Trivariate C^r polynomial macroelements', *Constr. Approx.* **26**, 11–28.
O. Lakkis and T. Pryer (2011), 'A finite element method for second order nonvariational elliptic problems', *SIAM J. Sci. Comput.* **33**, 786–801.
P.-L. Lions (1982), *Generalized Solutions of Hamilton–Jacobi Equations*, Vol. 69 of *Research Notes in Mathematics*, Pitman.
P.-L. Lions (1983), 'Optimal control of diffusion processes and Hamilton–Jacobi–Bellman equations II: Viscosity solutions and uniqueness', *Comm. Partial Diff. Equations* **8**, 1229–1276.
P.-L. Lions and B. Mercier (1980), 'Approximation numérique des équations de Hamilton–Jacobi–Bellman', *RAIRO Anal. Numér.* **14**, 369–393.
A. Maugeri, D. Palagachev and L. Softova (2000), *Elliptic and Parabolic Equations with Discontinuous Coefficients*, Vol. 109 of *Mathematical Research*, Wiley–VCH.
J.-L. Menaldi (1989), 'Some estimates for finite difference approximations', *SIAM J. Control Optim.* **27**, 579–607.
G. Mingione (2006), 'Regularity of minima: An invitation to the dark side of the calculus of variations', *Appl. Math.* **51**, 355–426.
J.-M. Mirebeau (2015), 'Discretization of the 3D Monge–Ampere operator, between wide stencils and power diagrams', *ESAIM Math. Model. Numer. Anal.* **49**, 1511–1523.
T. Motzkin and W. Wasow (1953), 'On the approximation of linear elliptic differential equations by difference equations with positive coefficients', *J. Math. Phys.* **31**, 253–259.
N. Nadirashvili and S. Vlăduţ (2007), 'Nonclassical solutions of fully nonlinear elliptic equations', *Geom. Funct. Anal.* **17**, 1283–1296.
N. Nadirashvili and S. Vlăduţ (2008), 'Singular viscosity solutions to fully nonlinear elliptic equations', *J. Math. Pures Appl.* (9) **89**, 107–113.
N. Nadirashvili, V. Tkachev and S. Vlăduţ (2012), 'A non-classical solution to a Hessian equation from Cartan isoparametric cubic', *Adv. Math.* **231**, 1589–1597.
M. Neilan (2010), 'A nonconforming Morley finite element method for the fully nonlinear Monge–Ampère equation', *Numer. Math.* **115**, 371–394.
M. Neilan (2013), 'Quadratic finite element approximations of the Monge–Ampère equation', *J. Sci. Comput.* **54**, 200–226.
M. Neilan (2017), 'Convergence analysis of a finite element method for second order non-variational elliptic problems', *J. Numer. Math.*, to appear.
L. Nirenberg (1953), 'On nonlinear elliptic partial differential equations and Hölder continuity', *Comm. Pure Appl. Math.* **6**, 103–156; addendum, 395.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

R. Nochetto and W. Zhang (2017*a*), 'Discrete ABP estimate and convergence rates for linear elliptic equations in non-divergence form', *Found. Comput. Math.*, to appear.

R. Nochetto and W. Zhang (2017*b*), Pointwise rates of convergence for the Oliker–Prussner's method for the Monge–Ampère equation. arXiv:1611.02786

R. Nochetto, D. Ntogakas and W. Zhang (2017), Two-scale method for the Monge–Ampère equation: Convergence rates. Preprint.

R. Nochetto, E. Otárola and A. J. Salgado (2015), 'Convergence rates for the classical, thin and fractional elliptic obstacle problems', *Philos. Trans. A* **373**(2050), 20140449.

A. Oberman (2006), 'Convergent difference schemes for degenerate elliptic and parabolic equations: Hamilton–Jacobi equations and free boundary problems', *SIAM J. Numer. Anal.* **44**, 879–895.

A. Oberman (2008), '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**, 221–238.

B. Øksendal (2003), *Stochastic Differential Equations: An Introduction with Applications*, sixth edition, Universitext, Springer.

V. Oliker and L. Prussner (1988), '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**, 271–293.

M. Safonov (1980), 'Harnack's inequality for elliptic equations and Hölder property of their solutions', *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)* **96**, 272–287, 312.

M. Safonov (1987), 'Unimprovability of estimates of Hölder constants for solutions of linear elliptic equations with measurable coefficients', *Mat. Sb.* (N.S.) **132**(174), 275–288.

A. J. Salgado and W. Zhang (2016), Finite element approximation of the Isaacs equation. arXiv:1512.09091v1

A. Schatz and L. Wahlbin (1982), 'On the quasi-optimality in L_∞ of the $\dot{H}^1$-projection into finite element spaces', *Math. Comp.* **38**(157), 1–22.

L. Silvestre (2015), Viscosity solutions of elliptic equations. http://math.uchicago.edu/~luis/preprints/viscosity-solutions.pdf

I. Smears and E. Süli (2013), 'Discontinuous Galerkin finite element approximation of nondivergence form elliptic equations with Cordès coefficients', *SIAM J. Numer. Anal.* **51**, 2088–2106.

I. Smears and E. Süli (2014), 'Discontinuous Galerkin finite element approximation of Hamilton–Jacobi–Bellman equations with Cordes coefficients', *SIAM J. Numer. Anal.* **52**, 993–1016.

I. Smears and E. Süli (2016), 'Discontinuous Galerkin finite element methods for time-dependent Hamilton–Jacobi–Bellman equations with Cordes coefficients', *Numer. Math.* **133**, 141–176.

S. Stojanovic (2005), 'Risk premium and fair option prices under stochastic volatility: The HARA solution', *CR Math. Acad. Sci. Paris* **340**, 551–556.

G. Strang and G. Fix (1973), *An Analysis of the Finite Element Method*, Prentice-Hall Series in Automatic Computation, Prentice-Hall.

L. Tolstoy (1998), *Anna Karenina*, Project Gutenberg, EBook #1399.

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press

N. Trudinger (1988), 'Comparison principles and pointwise estimates for viscosity solutions of nonlinear elliptic equations', *Rev. Mat. Iberoamericana* **4**, 453–468.

N. Trudinger and X.-J. Wang (2005), 'The affine Plateau problem', *J. Amer. Math. Soc.* **18**, 253–289.

N. Trudinger and X.-J. Wang (2008), 'Boundary regularity for the Monge–Ampère and affine maximal surface equations', *Ann. of Math.* (2) **167**, 993–1028.

O. Turanova (2015), 'Error estimates for approximations of nonhomogeneous nonlinear uniformly elliptic equations', *Calc. Var. Partial Diff. Equations* **54**, 2939–2983.

C. Villani (2003), *Topics in Optimal Transportation*, Vol. 58 of *Graduate Studies in Mathematics*, AMS.

C. Wang and J. Wang (2017), 'A primal–dual weak Galerkin finite element method for second order elliptic equations in non-divergence form', *Math. Comp.*, to appear.

X.-J. Wang (1996), 'Regularity for Monge–Ampère equation near the boundary', *Analysis* **16**, 101–107.

J. Witte and C. Reisinger (2012), 'Penalty methods for the solution of discrete HJB equations: Continuous control and obstacle problems', *SIAM J. Numer. Anal.* **50**, 595–625.

J. Xu and L. Zikatanov (1999), 'A monotone finite element scheme for convection-diffusion equations', *Math. Comp.* **68**(228), 1429–1446.

A. Ženíšek (1973), Hermite interpolation on simplexes in the finite element method. In *Proc. Equadiff III: Third Czechoslovak Conf. Differential Equations and their Appl., Brno, 1972*, Purkyně University, Brno, pp. 271–277. Vol. 1 of *Folia Fac. Sci. Natur. Univ. Purkynianae Brunensis, Ser. Monograph.*

https://doi.org/10.1017/S0962492917000071 Published online by Cambridge University Press