DISCRETE MAXIMUM PRINCIPLES

Master's Thesis

by

Csirik Mihály

Applied mathematics (Msc)

Adviser:

Karátson János

docent

Department of Applied Analysis

EÖTVÖS LORÁND TUDOMÁNYEGYETEM
TERMÉSZETTUDOMÁNYI KAR

2013

This page is intentionally left blank.

Acknowledgments

First and foremost I would like to acknowledge the patience of Ancsi, exhibited during the writing of this thesis. I am thankful for *my* patience also. Second, the support of my and her family was highly necessary at certain times.

I would also like to thank my adviser, János, and other colleagues for the interesting discussions, not limited to the scope of this present work.

This work was typeset with LaTeXusing the memoir, hyperref, $\mathcal{A}_{\mathcal{M}}\mathcal{S}$-LaTeX, *&c, &c, &c* packages.

This page is intentionally left blank.

Contents

Acknowledgments **iii**

Contents **v**

Preface **vii**

1 Introduction **1**

1.1 Maximum principles for harmonic functions 1

1.2 Maximum principles for linear elliptic operators 2

Maximum principles for weak solutions 3

1.3 Maximum Principles for Nonlinear Elliptic Operators 6

Singular quasilinear equations 7

1.4 Variational theory for nonlinear elliptic PDEs 10

Euler–Lagrange equations 10

Potential operators 12

Nonpotential operators 14

1.5 Nemytskii operators 14

2 Discrete maximum principles **17**

2.1 The Ritz–Galerkin Method for Nonlinear Problems 17

2.2 Variational Properties of the Ritz–Galerkin Method 18

2.3 Lowest-order Finite Element Method 19

2.4 Discrete Maximum Principles 21

2.5 Geometric Constraints 21

2.6 Algebraic Maximum Principles 22

2.7 Discrete Green's functions 24

2.8 Variational Approach to Discrete Maximum Principles 27

Bibliography **31**

This page is intentionally left blank.

Preface

"Man muss immer generalisieren." – C. G. J. Jacobi

Elliptic partial differential equations describe wide ranging phenomena in physics. Engineering practice requires the solution of partial differential equations on many types of irregular domains that necessitates the use of a digital computer to obtain an approximation. The question of error estimation is a difficult one not only theoretically, but computationally – for an *a posteriori* scheme one has to solve a simpler problem of the same kind as the original. Therefore it is entirely natural to look for some criteria that is easily checked, and essential for a physically sound solution to possess. Fortunately, there *is* such a criterion.

Maximum principles where already studied in the nineteenth century, their significance was not overlooked. For example in complex analysis they constitute the first deep insight into the behavior of a holomorphic function. The serious investigation of maximum principles in elliptic equations began with the work of E. Hopf in the first quarter of the twentieth century. More recently the results were extended to nonlinear elliptic equations. We shall give a brief account of the developments relevant to us in the first chapter.

The second part of the present work is concerned with the validity of the maximum principle for a discrete, or approximate solution. Various, mainly linear algebraic techniques were developed over the years. These frameworks for studying discrete maximum principles are highly successful and popular, albeit they somewhat conceal the inner workings of the underlying discretization scheme, in our case the Ritz–Galerkin method. In particular, the Ritz–Galerkin method directly minimizes the energy functional of the problem on a finite dimensional subspace. There have been recent progress in exploiting the variational characteristics of the said discretization, and the results are rather instructive.

This page is intentionally left blank.

1 Introduction

1.1 Maximum principles for harmonic functions

The *maximum modulus principle* was well-known to mathematicians of the 19th century. In this section we summarize these elementary facts, for more, see [6].

The principle states that if f is holomorphic on a region Ω of $\mathbb{C}$, and the function $|f|$ attains its maximum in Ω, then f is necessarily constant. Results like this are called strong maximum principles. One of the many consequences to this fact is the following.

(1.1.1) Maximum Modulus Principle. Let $\Omega \subset \mathbb{C}$ be a bounded open set. If f is continuous on $\overline{\Omega}$ and holomorphic on Ω, then

$$\max\{|f(z)| \ : \ z \in \overline{\Omega}\} = \max\{|f(z)| \ : \ z \in \partial\Omega\}.$$

Recall that a function $u \ : \ \Omega \longrightarrow \mathbb{R}$ is called *harmonic* on a open set $\Omega \subset \mathbb{C}$, if u is twice continuously differentiable on $\overline{\Omega}$ and satisfies Laplace's equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \ \text{ (on } \Omega).$$

A holomorphic function f can always be written as $f = u + iv$, where u and v are harmonic functions which satisfy the Cauchy–Riemann partial differential equations. Therefore the study of harmonic functions may provide insight in holomorphic function theory. In fact, one can deduce the above Maximum Modulus Principle from the corresponding maximum principle for harmonic functions – a fact already known to Gauss, who proved it using the mean-value property of harmonic functions.

(1.1.2) Harmonic Maximum Principle. Suppose $\Omega \subset \mathbb{C}$ is a bounded region, and $u \ : \ \Omega \longrightarrow \mathbb{R}$ harmonic. Then

$$\max\{u(z) \ : \ z \in \overline{\Omega}\} = \max\{u(z) \ : \ z \in \partial\Omega\}.$$

A harmonic function on a bounded region Ω may be interpreted as steady-state distribution of heat on Ω. The Harmonic Maximum Principle then states the intuitive physical fact that in a thermal equilibrium, the hottest part of Ω must lie on its boundary. That is, by imposing sufficiently smooth Dirichlet boundary conditions on the sufficiently smooth boundary $\partial\Omega$, a particular (unique) element gets selected from the set of harmonic functions on Ω; and it can never be the case that heat somehow „builds up" inside Ω. This qualitative property is one of the most important notions in the study of partial differential equations.

These theoretical facts may be interpreted as a very natural, physical requirement on an approximate solution to Laplace's equation. Before investigating these discrete analogs of the maximum principle, let us see what other, more general maximum principles were discovered.

1.2 Maximum principles for linear elliptic operators

In this section, we present classical maximum principles for elliptic operators.

Let $\Omega \subset \mathbb{R}^n$ be a domain, i.e. a connected open set. Let $u \in C^2(\Omega) \cap C(\overline{\Omega})$ and consider the linear differential operator L given by the instruction

$$\forall x \in \Omega \ : \ (Lu)(x) := \sum_{1 \le i,j \le n} a_{ij}(x) \frac{\partial^2 u}{\partial x_i \partial x_j}(x) + \sum_{1 \le i \le n} b_i(x) \frac{\partial u}{\partial x_i}(x). \tag{1.1}$$

We assume that all the coefficient functions are bounded and continuous.

(1.2.1) Definition. We say that the coefficient functions $\{a_{ij}\}$ are *locally uniformly positive definite in* Ω if $a_{ij} = a_{ji}$ on Ω for every $1 \le i, j \le n$ and for every compact set $K \subset \Omega$

$$\exists \alpha > 0 \ \forall x \in K \ \forall \xi \in \mathbb{R}^n \ : \ \sum_{1 \le i,j \le n} a_{ij}(x) \xi_i \xi_j \ge \alpha \|\xi\|^2.$$

Note that the symmetry $a_{ij} = a_{ji}$ can always be achieved for nonsymmetric coefficient functions in a straightforward manner, therefore this requirement is not restrictive at all. If the coefficient functions $\{a_{ij}\}$ are locally uniformly positive definite, the corresponding differential operator L is called *locally uniformly elliptic*.

(1.2.2) Weak Maximum Principle. Let Ω be a bounded domain, $u \in C^2(\Omega) \cap C(\overline{\Omega})$ a so-called subsolution of $L + c$, i.e. $Lu + cu \ge 0$ (on Ω), where $c \le 0$ is a bounded function. Then

$$\max_{\overline{\Omega}} u \le \max_{\partial\Omega} u^+,$$

where $u^+ = u \vee 0$, and $\vee$ denotes the supremum.

The proof is fairly standard, see e.g. [16, Theorem 2.5]. By a transition from u to $-u$ one gets $\min_{\overline{\Omega}} u \ge \min_{\partial\Omega} u^-$, therefore if u is a solution to the homogeneous equation $Lu + cu = 0$ (with some prescribed boundary), we have the so-called **comparison principle**

$$\forall x \in \overline{\Omega} \ : \ \min_{\partial\Omega} u^- \le u(x) \le \max_{\partial\Omega} u^+.$$

For a motivation to Hopf's strong maximum principle, we first consider a much weaker, almost trivial version: if the strict inequality

$$Lu > 0 \ (\text{on } \Omega)$$

holds, then u cannot have a maximum in Ω. If, on the contrary $x^* \in \Omega$ is a maximum of u, then $\nabla u(x^*) = 0$ and the Hessian matrix at x^*,

$$[H(x^*)]_{ij} = \frac{\partial^2 u}{\partial x_i \partial x_j}(x^*)$$

is negative semidefinite, by elementary multivariate calculus. Note that at x^*, we have

$$(Lu)(x^*) = \sum_{1 \le i,j \le n} a_{ij}(x^*) \frac{\partial^2 u}{\partial x_i \partial x_j}(x^*).$$

We claim that $(Lu)(x^*) \le 0$. Let A denote the matrix formed by $a_{ij}(x^*)$. Then

$$(Lu)(x^*) = e^\top (A \circ H) e = -e^\top (A \circ (-H)) e$$

where $e = (1, \dots, 1)^\top$ and $\circ$ denotes the Hadamard product. By Schur's product theorem, $A \circ (-H)$ is positive semidefinite, therefore $(Lu)(x^*) \le 0$. But this contradicts our original hypothesis.

(1.2.3) Hopf's Maximum Principle. Let $\Omega \subset \mathbb{R}^n$ be a domain, and L an operator of the form (1.1). Suppose the coefficient functions $\{a_{ij}\}$ of L are locally uniformly positive definite in Ω, and the coefficient functions $\{b_i\}$ are locally bounded in Ω, c is locally bounded from below on Ω, and finally, that the function $u \in C^2(\Omega) \cap C(\overline{\Omega})$ satisfies

$$Lu + cu \ge 0 \text{ (on } \Omega).$$

Statements.

- If $c = 0$ and u attains its maximum M in Ω, then $u \equiv M$ on Ω.
- If $c \le 0$ and u attains its positive maximum/negative minimum M in Ω, then $u \equiv M$ on Ω.
- If attains its maximum/minimum in Ω, and its value is zero, then $u \equiv 0$ on Ω.

The full proof involves several original ideas, and although it has undergone some simplification over the years, it is still long and therefore will not be presented here. For details, see [25].

Maximum principles for weak solutions

There are versions of the maximum principle for the weak solutions of linear elliptic boundary value problems. To formulate this, a weak form of the elliptic problem has to be derived. The operator L of (1.1) is not in divergence form, therefore the transition is unclear. Let us introduce

a sufficiently general linear operator $\mathcal{L}$, where $\Omega \subset \mathbb{R}^n$ is a locally Lipschitz domain, by the instruction

$$\forall x \in \Omega \ : \ (\mathcal{L}u)(x) = -\operatorname{div}(A(x)\nabla u(x)) + (b(x), \nabla u(x)) + c(x)u(x),$$

and the classical problem of finding $u \in C^1(\overline{\Omega}) \cap C^2(\Omega)$, such that

$$\left.\begin{aligned} \mathcal{L}u &= f \\ u\big|_{\Gamma_0} &= g \\ \big(d(x)u + (A(x)\nabla u, \nu)\big)\big|_{\Gamma_1} &= h \end{aligned}\right\} \tag{1.2}$$

where

- $\Gamma_0 \subset \partial\Omega$ is the *Dirichlet boundary* and $\Gamma_1 \subset \partial\Omega$ is the *Neumann boundary*, both relatively open in $\partial\Omega$, $\Gamma_0 \cap \Gamma_1 = \emptyset$ and $\overline{\Gamma}_0 \cup \overline{\Gamma}_1 = \partial\Omega$.
- A is an uniformly positive, but not necessarily symmetric (real) matrix
- $c - \frac{1}{2}\operatorname{div} b \geq 0$ (on Ω)
- $d + \frac{1}{2}(b, \nu) \geq 0$ (on Γ_1)

Note the for the classical problem, restrictions on f, g, d and h are not so simple to ensure uniqueness and regularity. See [17] for details, and [28] where this model problem is used throughout the investigation of discrete maximum principles.

The weak form of (1.2) consists of the space

$$X := \{u \in H^1(\Omega) \ : \ u \equiv 0 \text{ (on } \Gamma_0)\},$$

and

$$\begin{aligned} (\mathcal{L}u, v)_X &= \int_\Omega \Big[(A\nabla u, \nabla v) + (b, \nabla u)v + cuv\Big]\, dx + \int_{\Gamma_1} duv\, d\sigma, \\ \Phi(v) &= \int_\Omega fv\, dx + \int_{\Gamma_1} hv\, d\sigma - (\mathcal{L}\tilde{g}, v)_X, \end{aligned}$$

so that the problem is to find $u_0 \in X$, such that

$$\forall v \in X \ : \ (\mathcal{L}u_0, v)_X = \Phi(v), \tag{1.3}$$

thereby the weak solution is $u = u_0 + \tilde{g}$. Here, the function $\tilde{g} \in H^1(\Omega)$, $\tilde{g} = g$ (on Γ_0) is arbitrary, as its particular choice does not affect u at all; it is called the *Dirichlet lift* and to simplify notations we use the same symbol g.

It is also not too hard to prove that $(\mathcal{L}u, v)_X$ is uniformly positive on X, therefore it defines an energy inner product. It is also bounded, as is Φ, therefore the Riesz representation theorem can be employed to represent Φ using an element in the Hilbert space X. This yields existence and uniqueness of the weak solution.

In the following definition we take into account the fact that u is not continuous anymore.

(1.2.4) Definition. We say that the solution u to (1.3) satisfies the *weak maximum principle* if the following implication holds:

$$f \leq 0 \ \text{(a.e. on } \Omega) \ \text{and} \ h \leq 0 \ \text{(a.e. on } \Gamma_1) \implies \operatorname{ess\,sup}_{\overline{\Omega}} u \leq \operatorname{ess\,sup}_{\Gamma_0} u^+.$$

(1.2.5) Definition. We say that the solution u to (1.3) satisfies the *weak nonnegativity principle* if the following implication holds:

$$f \geq 0 \ \text{(a.e. on } \Omega), \ \ g \geq 0 \ \text{(a.e. on } \Gamma_0) \ \text{and} \ h \geq 0 \ \text{(a.e. on } \Gamma_1) \implies u \geq 0 \ \text{(a.e. on } \Omega).$$

(1.2.6) Assumption. We have $c \geq 0$ (a.e. on Ω) and $d \geq 0$ (a.e. in Γ_1).

(1.2.7) Theorem. Suppose Assumption (1.2.6) holds.

(1) A weak solution u to (1.3) satisfies the weak nonnegativity principle if and only if u satisfies the weak maximum principle.

(2) The weak solution u to (1.3) satisfies the weak maximum principle.

Proof. **(1)** is trivial.

(2) The main ingredient is that $H^1(\Omega)$ forms a vector lattice. Therefore if $\gamma := \operatorname{ess\,sup}_{\Gamma_0} u^+$, then

$$v := (u - \gamma)^+ \in H^1(\Omega),$$

where u is a weak solution. By definition, $v \geq 0$ (a.e. on Ω), $v = 0$ (a.e. on Γ_0) and

$$u = v + \gamma \ \text{(a.e. on } E := \{v \neq 0\}).$$

The following chain of inequalities yields $v = 0$ (a.e. on Ω), thus $u \leq \gamma$ (a.e. on Ω). Let $f \leq 0$ and $h \leq 0$, then

$$\begin{aligned} 0 &\geq \int_\Omega f v \, dx + \int_{\Gamma_1} h v \, d\sigma - 0 = \Phi(v) \\ &= (\mathcal{L}u, v) = \int_E \Big[(A\nabla u, \nabla v) + (b, \nabla u) v + cuv \Big] dx + \int_{\Gamma_1} duv \, d\sigma \\ &= (\mathcal{L}v, v) + \gamma \int_E cv^2 \, dx + \gamma \int_{\Gamma_1} dv^2 \, d\sigma \geq (\mathcal{L}v, v) \geq \alpha \|v\|_X \geq 0 \end{aligned}$$

□

1.3 Maximum Principles for Nonlinear Elliptic Operators

In what follows, we present the recent results in the book [25] by P. Pucci and J. Serrin. The proofs of these results are omitted for they are highly technical and rather long.

Let $\Omega \subset \mathbb{R}^n$ be a bounded domain. Let us turn our attention to nonlinear differential inequalities of the form

$$\forall x \in \Omega \ : \ \big[\mathcal{A}(u)\big](x) := \operatorname{div} A(x, u, \nabla u) + B(x, u, \nabla u) \geq 0, \tag{1.4}$$

where $A \ : \ \Omega \times \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ and $B \ : \ \Omega \times \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}$. The nonlinear operator $\mathcal{A}$ is said to be in *divergence form*. It is reasonable to assume that $A(\cdot, u, \nabla u) \in [L^1_{\text{loc}}(\Omega)]^n$ and $B(\cdot, u, \nabla u) \in L^1_{\text{loc}}(\Omega)$. To exhibit the weak form of this ineqality, let $0 \leq v \in C^1_c(\Omega)$, for which $v = 0$ in a neighborhood of $\partial\Omega$, so that via integration by parts we get

$$(\mathcal{A}(u), v) := \int_\Omega \big(A(x, u, \nabla u), \nabla v\big) + B(x, u, \nabla u) v \, dx \geq 0. \tag{1.5}$$

Furthermore, we seek solutions $u \in W^{1,p}(\Omega)$ with the additional assumption that u is *p-regular*, that is

$$A(\cdot, u, \nabla u) \in [L^{p'}_{\text{loc}}(\Omega)]^n, \quad p' = \frac{p}{p-1}.$$

(1.3.1) Assumption. If $p > 1$, then there exists constants $a_1, a_2, b_1, b_2, a, b \geq 0$, such that

$$\forall (x, \eta, \xi) \in \Omega \times \mathbb{R}_+ \times \mathbb{R}^n \ : \ \big(A(x, \eta, \xi), \xi\big) \geq a_1 \|\xi\|^p - a_2 \eta^p - a^p, \quad \text{and}$$
$$B(x, \eta, \xi) \leq b_1 \|\xi\|^{p-1} + b_2 \eta^{p-1} + b^{p-1}.$$

For $p = 1$, some of the constants are no longer necessary, but the inequalities are the same.

(1.3.2) Theorem. Let $u \in W^{1,p}_{\text{loc}}(\Omega)$ be a weak solution of $\mathcal{A}(u) \geq 0$ (on Ω) for $p > 1$ and suppose that $\mathcal{A}$ satisfies Assumption (1.3.1) with $b_1 = b_2 = 0$. If there exists $M \geq 0$, such that for all $\delta > 0$, there is a neighborhood $U \supset \partial\Omega$ so that $u \leq M + \delta$ a.e. on U, then we have $u^+ \in L^\infty(\Omega)$ and

$$u \leq C(a + b + a_2^{1/p} M) + M \quad \text{(a.e. on } \Omega),$$

for a suitable constant C depending only on n, p, $|\Omega|$ and a_2.

(1.3.3) Theorem. Let $u \in W^{1,p}_{\mathrm{loc}}(\Omega)$ be a weak solution of $\mathcal{A}(u) \geq 0$ (on Ω) for $p > 1$ and suppose that $\mathcal{A}$ satisfies Assumption (1.3.1) with $a_2 = b_2 = 0$. If there exists $M \geq 0$, such that for all $\delta > 0$, there is a neighborhood $U \supset \partial\Omega$ so that $u \leq M + \delta$ a.e. on U, then we have $u^+ \in L^\infty(\Omega)$ and

$$u \leq C(a + b) + M \quad \text{(a.e. on } \Omega),$$

for a suitable constant C depending only on n, p, $|\Omega|$ and b_1.

(1.3.4) Example. Consider the p**-Poisson** boundary value problem $(1 < p < \infty)$[1]

$$\left.\begin{aligned} -\triangle_p u := -\operatorname{div}(\|\nabla u\|^{p-2}\nabla u) &= f \\ u\big|_{\partial\Omega} &= g \end{aligned}\right\}$$

therefore $A(x, \eta, \xi) = \|\xi\|^{p-2}\xi$. Obviously we have $(A(x, \eta, \xi), \xi) = \|\xi\|^p$, so Assumption (1.3.1) holds. We will later cite results that easily guarantee the existence of a weak solution $u - g \in W^{1,p}_0(\Omega)$, if $g \in W^{1,p}(\Omega)$. Thus Theorem (1.3.3) yields, for a sufficently nice domain Ω, that whenever $f \leq 0$, we have

$$u \leq \operatorname{ess\,sup}_{\partial\Omega} g \quad \text{a.e. on } \Omega,$$

in perfect analogy with the maximum princples for the weak solution presented in the previous section. Note however that there is further regularity, more precisely if $g \in C(\overline{\Omega})$, then $u \in C(\overline{\Omega})$ and $u\big|_{\partial\Omega} = g\big|_{\partial\Omega}$. See [23] for some details.

Singular quasilinear equations

Consider the quasilinear operator of divergence form (cf. [21])

$$\mathcal{Q}(u) := -\operatorname{div}\big(A(\|\nabla u\|)\nabla u\big) + q(u), \tag{1.6}$$

and the corresponding boundary value problem

$$\left.\begin{aligned} \mathcal{Q}(u) &= f \\ u\big|_{\partial\Omega} &= g \end{aligned}\right\} \tag{1.7}$$

It was shown in [21] that under suitable assumptions a continuous weak solution u of (1.7) satisfies the weak maximum principle.

[1] The case $p = 1$ corresponds to the mean curvature problem, and will not be treated here.

(1.3.5) Assumption. Let

1. $A \in C^1[0,+\infty)$, $\exists c_1, c_2 > 0 \ : \ c_1 \leq A \leq c_2$ and $A' > 0$.
2. $q \in C^1(\mathbb{R})$, and there exists constants $a_1 > 0$ and $a_2 > 0$, and an exponent $q \geq 2$ if $n = 2$ and $2 \leq q \leq 2n/(2-n)$ otherwise, such that

$$\frac{\partial q(\eta)}{\partial \eta} \leq a_1 + a_2|\eta|^q.$$

(1.3.6) Theorem. Suppose that the problem (1.7) satisfies Assumption (1.3.5) and $u \in C(\overline{\Omega}) \cap C^1(\Omega)$ is a weak solution. Then the weak maximum principle holds:

$$f \leq 0 \ \text{ (on } \Omega) \implies \max_{\overline{\Omega}} u \leq \max\{0, \max_{\partial\Omega} g\}.$$

This yields in particular, that

$$f \leq 0 \ \text{ (on } \Omega), \ g \leq 0 \ \text{ (on } \partial\Omega) \implies \max_{\overline{\Omega}} u \leq 0.$$

One should not overlook the physical significance of this fact, for example if $-u$ describes the concentration of some chemical.

As an interesting sidetrack, we consider nonpositive subsolutions of (1.6). The article [27] by Vázquez establish a necessary and sufficient condition, called the **Vázquez condition** on the function q for the strong maximum principle to hold for semilinear problem

$$u \geq 0, \ \ -\triangle u + q(u) = f$$

The generalization to the quasilinear case, for example the use of the p-Laplace operator admits similar characterization, see [25] or [24]. We now present these results in a nutshell. Consider the quasilinear operator inequality

$$u \leq 0, \ \ \mathcal{Q}(u) \geq 0$$

where

(Q1) $A \in C^1(0,+\infty)$,

(Q2) $q \in C[0,+\infty)$, $q(0) = 0$, and $\exists \delta > 0$ such that, q is non-decreasing on $(0, \delta)$

(Q3) $\Phi \in C(0,+\infty)$, $\Phi(s) := sA(s)$ is strictly increasing, and it admits a continuous extension at zero: $\Phi(0) := \lim_{0+} \Phi = 0$.

Now let

$$H \in C[0,+\infty), \quad H(s) := s\Phi(s) - \int_0^s \Phi(t)\,dt,$$

and

$$Q(u) := \int_0^u q(s)\,ds \tag{1.8}$$

(1.3.7) Vázquez condition. Either

1. $q = 0$ (on $[0,\mu)$), for some $\mu > 0$, or
2. $q > 0$ (on $(0,\delta)$) and
$$\int_0^\delta \frac{ds}{H^{-1}(Q(s))} = +\infty.$$

(1.3.8) Strong Maximum Principle. A nonpositive classical (distributionsal) solution of $\mathcal{Q}(u) \geq 0$ satisfies the strong maximum principle, i.e. the implication

$$\Big(\exists x \in \Omega \ : \ u(x) = 0\Big) \implies u \equiv 0 \ \text{ (on } \Omega)$$

holds, if and only if the Vázquez condition holds.

Note that this maximum principle implies the weak one for the linear case, for zero is a maximum value of a nonpositive function, therefore it is the constant zero function.

(1.3.9) Example. The nonlinear reaction-diffusion equation

$$u \geq 0, \quad -\operatorname{div}(\|\nabla u\|^{p-2}\nabla u) + u^\alpha = f, \tag{1.9}$$

$\alpha > 0$, $p > 1$, with some Dirichlet boundary. With the previous notations

$$A(s) = s^{p-2}, \quad q(u) = u^\alpha,$$

so $\Phi(s) = s^{p-1}$, and $H(s) = s^p(p-1)/p$, $Q(u) = u^{\alpha+1}/(\alpha+1)$. Putting it all together, $H^{-1}(t) = (tp/(p-1))^{1/p}$, and

$$\int_0^\delta \frac{ds}{\frac{p}{p-1}\frac{s^{\alpha+1}}{\alpha+1}} = \frac{(p-1)(\alpha+1)}{p} \int_0^\delta s^{-\alpha-1}\,ds = +\infty.$$

This means that the Vázquez condition is satisfied, so we may conclude that the strong maximum principle holds.

1.4 Variational theory for nonlinear elliptic PDEs

Euler–Lagrange equations

Consider the „energy" functional $j : W_0^{1,p}(\Omega) \longrightarrow \mathbb{R}$ given by

$$\forall u \in W^{1,p}(\Omega) \ : \ j(u) := \int_\Omega L(x, u, \nabla u)\, dx, \tag{1.10}$$

where the function L often called the *Lagrangian* by the physics community. In this section, we discuss the important case where the minimization problem

$$\inf\{ j(u) \ : \ u \in W_0^{1,p}(\Omega)\} \tag{1.11}$$

has a unique solution. The simplest case is the minimization of the Dirichlet integral, i.e. $L(x, \eta, \xi) := \|\xi\|^2$. See the standard work of Gelfand and Fomin [15], Jost and Li-Jost [19] or [7] for Courant's monograph on Dirichlet's principle. Note that for inhomogeneous boundary conditions one can takes $u - g \in W_0^{1,p}(\Omega)$ for minimization, where $j(g) < +\infty$.

(1.4.1) Assumption. The mapping $(\eta, \xi) \mapsto L(x, \eta, \xi)$ is $C^1(\mathbb{R} \times \mathbb{R}^n)$ for all $x \in \Omega$ and there exists constants $a_1, b_1, c_1 \geq 0$ and functions $a \in L^1(\Omega)$, $b \in L^{p/(p-1)}(\Omega)$ and $c \in L^{p/(p-1)}(\Omega)$ such that for every $(x, \eta, \xi) \in \Omega \times \mathbb{R} \times \mathbb{R}^n$

$$\begin{aligned} |L(x, \eta, \xi)| &\leq a_1(\|\xi\|^p + |\eta|^p) + a(x) \\ \left| \frac{\partial L(x, \eta, \xi)}{\partial \eta} \right| &\leq b_1(\|\xi\|^{p-1} + |\eta|^{p-1}) + b(x) \\ \left| \frac{\partial L(x, \eta, \xi)}{\partial \xi} \right| &\leq c_1(\|\xi\|^{p-1} + |\eta|^{p-1}) + c(x) \end{aligned}$$

This assumption, albeit restrictive, yields the following (semi-)classical result.

(1.4.2) Theorem. Suppose that $j : W^{1,p}(\Omega) \longrightarrow \mathbb{R}$ satisfies Assumption (1.4.1). Then j is Gateaux-differentiable at every $u \in W^{1,p}(\Omega)$ and in every direction $v \in W^{1,p}(\Omega)$ we have

$$(\partial_v j)(u) = \int_\Omega \left[\sum_{1 \leq i \leq n} \frac{\partial L(x, u, \nabla u)}{\partial \xi_i} \frac{\partial v}{\partial x_i} + \frac{\partial L(x, u, \nabla u)}{\partial \eta} v(x) \right] dx \tag{1.12}$$

The proof relies on Lebesgue's theorem, hence the assumptions on the various integrable bounds on the partials of the Lagrangian.

For a minimizer $u \in W_0^{1,p}(\Omega)$ of the homogeneous version of (1.10) we necessarily have $(\partial_v j)(u) = 0$ for every $v \in W_0^{1,p}(\Omega)$, called the *Euler-Lagrange equation*, which, after integration by parts and an application of the fundamental theorem of calculus of variations yields

$$-\operatorname{div} A(x, u, \nabla u) + B(x, u, \nabla u) = 0,$$

where

$$1 \leq i \leq n \ : \ A_i(x, \eta, \xi) := \frac{\partial L(x, \eta, \xi)}{\partial \xi_i}, \ \text{ and } \ B(x, \eta, \xi) := \frac{\partial L(x, \eta, \xi)}{\partial \eta}.$$

The following observation brings in the notion of convexity into the picture. Suppose that $\xi \mapsto L(x, \eta, \xi) \in C^2(\mathbb{R}^n)$, evaluating the second Gateaux-derivative $\partial_v^2 j$ on a minimizer u, we should get $\partial_v^2 j(u) \geq 0$. After a nontrivial calculation one obtains the *Legendre–Hadamard condition*

$$\forall (x, \eta, \xi) \in \Omega \times \mathbb{R} \times \mathbb{R}^n \ : \ \sum_{1 \leq i,k \leq n} \frac{\partial^2 L(x, \eta, \xi)}{\partial \xi_i \partial \xi_k} \xi_i \xi_k \geq 0.$$

From elementary multivariate analysis we obtain that $\xi \mapsto F(x, \eta, \xi)$ is convex iff the Legendre–Hadamard condition holds.

Furthermore, it turns out that convexity is enough for existence and strict convexity for uniqueness, but note that we do not necessarily have an Euler-Lagrange equation in this case.

(1.4.3) Assumption. Suppose that $L \in C(\overline{\Omega} \times \mathbb{R} \times \mathbb{R}^n)$,

$$\forall (x, \eta) \in \overline{\Omega} \times \mathbb{R} \ : \ \xi \mapsto L(x, \eta, \xi) \ \text{ is convex,}$$

and there exists $p > q \geq 1$ and $a_1 \in \mathbb{R}_+$ $a_2, a_3 \in \mathbb{R}$ such that

$$\forall (x, \eta, \xi) \in \overline{\Omega} \times \mathbb{R} \times \mathbb{R}^n \ : \ |L(x, \eta, \xi)| \leq a_1 \|\xi\|^p + a_2 |\eta|^q + a_3.$$

(1.4.4) Theorem. Let $\Omega \subset \mathbb{R}^n$ be a bounded Lipschitz domain and let the Lagrangian L satisfy Assumption (1.4.3). Then there exists a minimizer $u \in W_0^{1,p}(\Omega)$ of (1.11) that is unique if $(\eta, \xi) \mapsto L(x, \eta, \xi)$ is strictly convex for every $x \in \overline{\Omega}$.

The proof is difficult in this generality, we refer the reader to [8] for a proof that uses Assumption (1.4.1), and follow the references therein for the complete treatment.

Since the Gateaux derivative (1.12) exists in every direction it is reasonable to ask whether it is represented by an element $w \in W^{-1,p}(\Omega) = (W_0^{1,p}(\Omega))'$, i.e if $(\partial_v j)(u) = (w, u)$ for every $v \in W_0^{1,p}(\Omega)$. The following section presents the results of investigating this possibility in a general setting.

Potential operators

Potential operators have a highly developed theory, and offer an elegant framework for the treatment of a relatively large class of nonlinear boundary value problems. For an introduction, see the first chapter of Chabrowski's textbook [3], or Karátson's notes [20]. The book by Faragó and Karátson [13] contains a number of problems from physics that is treatable with this theory. To us, the most important aspect of this framework is that it naturally handles the Ritz–Galerin type discretizations to be discussed in the next chapter.

Let X be a reflexive Banach space and $\mathcal{A} : X \longrightarrow X'$ a (nonlinear) mapping. We say that $\mathcal{A}$ is a *potential operator* if there exists a (nonlinear) functional $\mathcal{J} : X \longrightarrow \mathbb{R}$ (called the *potential* of $\mathcal{A}$), such that $\mathcal{J}$ is Gateaux differentiable and

$$\forall u, v \in X \ : \ (\partial_v \mathcal{J})(u) = (\mathcal{A}(u), v),$$

where

$$(\partial_v \mathcal{J})(u) := \lim_{h \to 0} \frac{\mathcal{J}(u + hv) - \mathcal{J}(u)}{h},$$

and $v \mapsto \partial_v \mathcal{J} \in B(X, X')$. In this scenario, i.e. when $\mathcal{A}$ is a potential operator, the homogeneous problem $\mathcal{A}(u) = 0$ is reduced to finding the critical points of the real functional $\mathcal{J}$, a classical problem presented in the previous section if $\mathcal{J}$ is of the form (1.10). Note that for the inhomogeneous case $\mathcal{A}(u) = f$, where $f \in X'$, the functional to minimize is

$$j(u) := \mathcal{J}(u) - (f, u), \tag{1.13}$$

simply because $(\partial_v j)(u) = (\mathcal{A}(u), v) - (f, v)$ for every $u, v \in X$, that is $\mathcal{A}(u) = f$ for a minimizer $u \in X$. In what follows, we present results that, under suitable assumptions on the operator $\mathcal{A}$, yield the reverse implication, thereby establishing a *variational principle* – an equivalence between a solution of nonlinear operator equation $\mathcal{A}(u) = f$ and a minimization of the corresponding potential (1.13).

A nonlinear operator $\mathcal{A} : X \longrightarrow \mathcal{B}(Y, Z)$ is called *hemicontinuous* if

$$\forall u, v \in X \ \forall w \in Y \ : \ \mathbb{R} \ni t \mapsto \mathcal{A}(u + tv)w \in Z \ \text{ is continuous.}$$

As a generalization of the notion of a positive linear operator, a nonlinear operator $\mathcal{A} : X \longrightarrow X'$ is called *monotone*, if

$$\forall u, v \in X \ : \ (\mathcal{A}(u) - \mathcal{A}(v), u - v) \geq 0,$$

and *strictly monotone* if the inequality is a strict one, in other words if equality implies $u = v$. An important strengthening of the monotonicity is the concept of a *uniformly monotone* operator:

there exists a continuous function $\kappa : [0,+\infty) \longrightarrow [0,+\infty)$, with $\kappa(t) \longrightarrow +\infty$ whenever $t \longrightarrow +\infty$, $\kappa(0) = 0$ and

$$\forall u, v \in X : (\mathcal{A}(u) - \mathcal{A}(v), u - v) \geq \kappa(\|u - v\|)\|u - v\|.$$

The nonlinear analogue of a uniformly positive linear operator is a strengthening of uniform monotonicity with the choice $\kappa(t) := Ct$.

The operator $\mathcal{A}$ is called *coercive* if $(\mathcal{A}(u_k), u_k)/\|u_k\| \longrightarrow +\infty$ if $\|u_k\| \longrightarrow \infty$. It follows easily that a uniformly monotone operator is necessarily coercive.

We say that a functional $j : X \longrightarrow \mathbb{R}$ is *convex* if, for every $u, v \in X$, $[0,1] \ni \lambda \mapsto j((1-\lambda)u+\lambda v) \in \mathbb{R}$ is convex. The following theorem is the aforementioned variational principle for monotone potential operators.

(1.4.5) Theorem. Let $\mathcal{A} : X \longrightarrow X'$ be a monotone operator with potential $\mathcal{J}$. Then $u \in X$ is a solution of $\mathcal{A}(u) = f$ if and only if u is minimizer of $j(u) := \mathcal{J}(u) - (f, u)$.

Proof. We only need to show the "only if" part. First, we show that j is convex in two parts.

I. The monotonicity of $\mathcal{A}$ implies the convexity of its potential $\mathcal{J}$. To see this, let $v_1, v_2 \in X$ and

$$[0,1] \ni \lambda \mapsto \Phi_{v_1,v_2}(\lambda) := \mathcal{J}((1-\lambda)v_1 + \lambda v_2).$$

Then,

$$\Phi'_{v_1,v_2}(\lambda) = (\mathcal{A}((1-\lambda)v_1 + \lambda v), v_2 - v_1).$$

Therefore, using the monotonicity of $\mathcal{A}$ we have for $\lambda_2 > \lambda_1$

$$\Phi'_{u,v}(\lambda_2) - \Phi'_{u,v}(\lambda_1) = \big(\mathcal{A}((1-\lambda_2)v_1 + \lambda_2 v_2) - \mathcal{A}((1-\lambda_1)v_1 + \lambda_1 v_2), v_2 - v_1\big) \geq 0.$$

Thus Φ'_{v_1,v_2} is monotone, so by elementary calculus Φ_{v_1,v_2} is convex, therefore $\mathcal{J}$ is convex.

II. Since $u \mapsto (f, u)$ is linear, we deduce that $j(u) = \mathcal{J}(u) - (f, u)$ is convex.

III. For arbitrary $v_1, v_2 \in X$ it is easy to establish using the convexity and Gateaux differentiability of j that

$$j(v_1) - j(v_2) \geq (j'(v_2), v_1 - v_2). \tag{1.14}$$

IV. Finally, let $\mathcal{A}(u) = f$, or $(\partial_v j)(u) = 0$ for every $v \in X$. The inequality (1.14) now reduces to $j(v) \geq j(u)$ $(v \in X)$ which is precisely what we needed to show. □

There are of course various sets of conditions that guarantee *uniqueness* of the solution of such abstract minimization problems.

Nonpotential operators

Let X be a real Hilbert space, $\mathcal{A} : X \longrightarrow X$ a (nonlinear) operator.

(1.4.6) Assumption. Suppose $\mathcal{A}$ be a uniformly monotone operator:

$$\exists m > 0 \ \forall u, v \in X \ : \ \|\mathcal{A}(u) - \mathcal{A}(v)\| \geq m\|u - v\|^2.$$

Also, suppose that $\mathcal{A}$ is Lipschitz continuous:

$$\exists M > 0 \ \forall u, v \in X \ : \ \|\mathcal{A}(u) - \mathcal{A}(v)\| \leq M\|u - v\|.$$

(1.4.7) Theorem. Under Assumption (1.4.6), for every $f \in X$, there exists a unique solution $u \in X$ to the equation $\mathcal{A}(u) = f$.

1.5 Nemytskii operators

Unfortunately the theory sketched above fails to handle certain nonlinearities. An important nonlinear boundary value problem, that has a relatively accessible theory is the following.

$$\left.\begin{aligned} -\triangle u &= g(u) + f \\ u\big|_{\partial\Omega} &= 0 \end{aligned}\right\} \tag{1.15}$$

(1.5.1) Assumption. Suppose the following.

1. $q \in [1, 2^*)$, where $2^* = 2n/(n-2)$ is the critical Sobolev exponent.
2. $f \in L^{\frac{q}{q-1}}(\Omega)$
3. $\exists c > 0 \ : \ |g(s)| \leq c|s|^{q-1}$
4. $\forall s \in \mathbb{R} \ : \ sg(s) \leq 0$

Let $j : W_0^{1,2}(\Omega) \longrightarrow \mathbb{R}$ be

$$j(u) := \int_\Omega \frac{1}{2}\|\nabla u\|^2 \, dx - \int_\Omega \mathcal{G}(u) \, dx - \int f u \, dx,$$

where

$$\big[\mathcal{G}(u)\big](x) := \int_0^{u(x)} g(s) \, ds$$

is a Nemytskii operator and $\mathcal{G} : L^q(\Omega) \longrightarrow L^1(\Omega)$ continuous map as per the following remark.

(1.5.2) Remark. A function $g : \Omega \times \mathbb{R} \longrightarrow \mathbb{R}$ is called a *Carathéodory function*, if

$$\forall \eta \in \mathbb{R} \ : \ x \mapsto g(x, \eta) \ \text{ is Borel measurable, and}$$
$$\text{a.a. } x \in \Omega \ : \ \eta \mapsto g(x, \eta) \ \text{ is continuous.}$$

The following facts can be shown easily enough (or see [10]).

(a) Functions in $C(\Omega \times \mathbb{R})$ are Carathéodory,

(b) For every Borel measurable function $u : \Omega \longrightarrow \mathbb{R}$, the composition

$$\mathcal{G}(u) : \Omega \longrightarrow \mathbb{R}, \ \ \big[\mathcal{G}(u)\big](x) := g(x, u(x))$$

is measurable on Ω. The nonlinear map $\mathcal{G}$ is called the *Nemytskii operator* corresponding to the Carathéodory function g, and thus $\mathcal{G}$ maps measurable functions to measurable functions.

(c) Suppose that $1 \leq p, q < \infty$ and

$$\exists a > 0 \ \ \exists h \in L^q(\Omega) \ \ \forall \eta \in \mathbb{R} \ \ \text{a.a. } x \in \Omega \ : \ |g(x, \eta)| \leq h(x) + a|\eta|^{p/q}.$$

Then $\mathcal{G} : L^p(\Omega) \longrightarrow L^q(\Omega)$ and it is continuous. Moreover $\mathcal{G}$ maps bounded sets in $L^p(\Omega)$ to bounded sets in $L^q(\Omega)$.

□

It can be shown that functional j posess at least one minimizer (see [10, pp. 324–325]), therefore (1.16) has at least one weak solution in $W_0^{1,2}(\Omega)$. Uniqueness is guaranteed if we assume that g is nonincreasing and Assumption (1.5.1) (3) is satisfied with $q = 2^*$.

(1.5.3) Example. Consider the semilinear reaction-diffusion problem

$$\left.\begin{aligned} -\triangle u + |u|^{q-2}u &= f \\ u\big|_{\partial\Omega} &= 0 \end{aligned}\right\} \tag{1.16}$$

Now $g(\eta) = -|\eta|^{q-2}\eta$, and it is easy to see that it satisifes Assumption (1.5.1), and g is monotone decreasing; existence and uniqueness follow. Moreover

$$\big[\mathcal{G}(u)\big](x) = -\frac{1}{q}|u(x)|^q,$$

therefore the energy functional is

$$j(u) = \int_\Omega \frac{1}{2}\|\nabla u\|^2 \, dx + \int_\Omega \frac{1}{q}|u|^q \, dx - \int_\Omega f u \, dx. \tag{1.17}$$

For later purposes, we note that if $u \leq v$ (on Ω), then

$$\int_\Omega \frac{1}{q}|u|^q \, dx \leq \int_\Omega \frac{1}{q}|v|^q \, dx \tag{1.18}$$

2 Discrete maximum principles

2.1 The Ritz–Galerkin Method for Nonlinear Problems

The framework sketched in the previous chapter admits a popular finite-dimensional approximation scheme that involves the solution (of a sequence) of nonlinear algebraic equations. We now present this technique from the point of view of numerical analysis, but we note that these ideas also have a wide variety of theoretical applications.

Let $\mathcal{A} : X \longrightarrow X$, where X is a real Hilbert space. Consider the nonlinear problem of finding $u - g \in X$, such that

$$\mathcal{A}(u) = f, \tag{2.1}$$

for a given $f \in X'$ and $g \in X$. Suppose that $\mathcal{A}$ satisfies Assumption (1.4.6), then by Theorem (1.4.7), there exists a unique solution $u \in X$ to the problem. The *Ritz–Galerkin method* constructs a sequence of approximate solutions $\{u_n : n \in \mathbb{N}\} \subset X$ such that for every $n \in \mathbb{N}$, $u_n \in X_n$ for some finite-dimensional subspace $X_n \subset X$. The essential requirement for the convergence $\|u_n - u\| \longrightarrow 0$ as $n \longrightarrow \infty$ is that

$$\forall v \in X : \operatorname{dist}(v, X_n) \longrightarrow 0, \quad \text{whenever } n \longrightarrow \infty.$$

For the sake of completeness, we mention that convergence is due to the fact that the approximate solution $\{u_n\}$ enjoys a *quasi-optimality* property:

$$\|u - u_n\| \leq \frac{M}{m} \operatorname{dist}(u, X_n),$$

a relation easily deduced from the assumptions regarding the operator $\mathcal{A}$ and the *Galerkin-orthogonality* $\mathcal{A}(u_n) - \mathcal{A}(u) \in X_n^{\perp}$. We will not use these aspects of the Ritz–Galerkin method.

We will, on the other hand, refer to certain particularities of the discretization. To concretize the situation, let $J_n \subset \mathbb{N}$ denote a finite set indexing a basis of X_n, then

$$X_n = \operatorname{span}\{\varphi_{n,\alpha} \in X : \alpha \in J_n\},$$

so that $\dim X_n = |J_n|$. An approximate solution $u_n \in X_n$ ($n \in \mathbb{N}$) is defined by the relation

$$\forall v \in X_n : (\mathcal{A}(u_n), v) = (f, v),$$

which has a unique solution by Theorem (1.4.7). Choosing $v := \varphi_{n,\alpha}$ as test functions we hereby obtain

$$\forall \alpha \in J_n : (\mathcal{A}(u_n), \varphi_{n,\alpha}) = (f, \varphi_{n,\alpha}),$$

which is a nonlinear (algebraic) system of equations, since if

$$u_n = \sum_{\beta \in J_n} \xi_\beta \varphi_{n,\beta} \quad (\xi_\beta \in \mathbb{R}),$$

and for every $n \in \mathbb{N}$ and $\alpha \in J_n$,

$$A_{n,\alpha}(\xi) := \left(\mathcal{A}\Big(\sum_{\beta \in J_n} \xi_\beta \varphi_{n,\beta} \Big),\, \varphi_{n,\alpha} \right),$$
$$F_{n,\alpha} := (f, \varphi_{n,\alpha})$$

then the finite dimensional problem can be written compactly as

$$\text{given} \quad f \in X \quad \text{and} \quad X_n \subset X \quad \text{find} \quad \xi \in \mathbb{R}^{|J_n|} \;:\; A_n(\xi) = F_n. \tag{2.2}$$

Of course, this nonlinear equation is extremely difficult, if not impossible to solve in general. There are a number of cases where the Newton–Kantorovich method is applicable, see for example [13]. We are, however, exclusively concerned with the qualitative properties of the discretization itself, i.e. the choice of the approximating spaces $\{X_n\}$.

2.2 Variational Properties of the Ritz–Galerkin Method

Suppose $X = W_0^{1,p}(\Omega)$ and let us turn to the case when $\mathcal{A}$ has a potential $\mathcal{J}$ of the form (1.10) and it satisfies Assumptions (1.4.1) and (1.4.3). Also, assume that the Lagrangian is strictly convex to ensure uniqueness, see Theorem (1.4.4). Under these circumstances, Problem (2.1) is equivalent to finding the unique solution to the minimization problem

$$\min\{ j(v) \;:\; v - g \in X \} =: j(u), \tag{2.3}$$

where $g \in X$, and

$$j(v) = \int_\Omega L(x, v, \nabla v)\, dx - \int_\Omega f v. \tag{2.4}$$

The finite dimensional variant of this problem is

$$\min\{ j(v_n) \;:\; v_n - g \in X \} =: j(u_n),$$

which is uniquely solvable and $j(u_n) \geq j(u)$ holds. Therefore the Ritz–Galerkin method applied to problems that admit a variational formulation corresponds to the minimization of the energy functionals on finite dimensional subspaces. The relevance of this fact to the discrete maximum principles has only been realized recently, and we present these results due to Kreuzer et al. in a later section.

2.3 Lowest-order Finite Element Method

In this section, we discuss a popular subtype of the Ritz–Galerkin method, called the finite element method.

The sequence $\{X_n\}$ and bases $\{\varphi_{n,\alpha}\}$ are derived from a simplicial decomposition of the *polyhedral* domain Ω. (If Ω is not polyhedral, a fair amount of complications arise when approximating the boundary.) The simplicial decomposition $\mathcal{T}_n$ (also called the *grid*) is assumed to be free of hanging nodes, that is, a vertex of a simplex can only meet another simplex at a vertex. Also, a Ritz–Galerkin method, as we defined it, is *conforming*, i.e. $X_n \subset X$ for every $n \in \mathbb{N}$.

A traditional finite element scheme geometrically decomposes the polyhedral domain Ω into a finite number disjoint of simplices $\{T_{n,\alpha} \ : \ \alpha \in J\}$, where $n \in \mathbb{N}$. Let $\Omega_n \subset \mathbb{R}^d$ denote the finite set of nodes of the simplicial decomposition $\{T_{n,\alpha}\}$.

The *lowest order (linear) finite element* recipe is as follows: choose a basis

$$\{\varphi_{n,x} \ : \ x \in \Omega_n\} \subset X$$

of $\{T_{n,\alpha}\}$-piecewise linear functions, such that the δ-property

$$\forall x, x' \in \Omega_n \ : \ \varphi_x(x') = \delta(x - x') \tag{2.5}$$

holds[1], where $\delta(0) = 1$ and $\delta(z) = 0$ for $z \in \mathbb{R}^d \setminus \{0\}$. These conditions uniquely determine the basis $\{\varphi_x\}$, it contains "hat functions", the concrete formula is irrelevant to us, but easy enough to establish using barycentric coordinates. Note, in particular, that $0 \leq \varphi_x \leq 1$. For simplicity, we will omit the subscript n if it is clear from the context. Even in a more general higher-order finite element method the set

$$\operatorname{supp}(\varphi_x) \cap \operatorname{supp}(\varphi_{x'})$$

is empty or small whenever $x \neq x'$. In other words, if x and x' are far away, the supports of their corresponding so-called *nodal basis functions* are disjoint.

For later purposes, we record that the δ-property (2.5) implies that

$$\forall x \in \Omega_n \ : \ \sum_{x' \in \Omega_n} \varphi_x(x') = 1. \tag{2.6}$$

The treatment of Dirichlet and Neumann boundary conditions differ at this point: since $X = H_0^1(\Omega)$ consists of functions that vanish (in trace sense) on $\partial\Omega$, by density of $C_0(\Omega)$, only the interior nodes $\Omega_n^{\text{int}} := \Omega_h \cap \Omega$ matter – nodes $x \in \Omega_n$ for which $\operatorname{supp}(\varphi_x) \cap \partial\Omega = \emptyset$. Thus, for homogeneous Dirichlet boundary conditions the finite element space is

$$X_n = \operatorname{span}\{\varphi_x \ : \ x \in \Omega_n^{\text{int}}\} \subset C_0(\Omega).$$

[1]For simplicity, we will omit the subscript n if it is clear from the context.

Model problem I. We now turn to the simplest nontrivial case. Let $X := H_0^1(\Omega)$ (therefore $X \cong X' = H^{-1}(\Omega)$) and the Laplacian $\mathcal{A} : X \longrightarrow X$ is the linear isomorphism given by

$$(\mathcal{A}u, v) := \int_\Omega (K\nabla u, \nabla v), \quad \text{and} \quad (F, v) = \int_\Omega fv, \tag{2.7}$$

where $K \in \mathbb{R}^{d\times d}$ is constant symmetric positive definite matrix and $f \in L^2(\Omega)$. If the boundary is inhomogeneous, that is $u = g$ (on $\partial\Omega$), for some $g \in H^{1/2}(\Omega)$, the solution u has to satisfy $u - g \in X$. Since $\mathcal{A}$ is linear, this amounts to adding another term to the functional $F \in X'$, thereby reducing the inhomogeneous problem to the homogeneous one. Write the homogeneous approximate solution $u_{0,n} := g - u_n \in X_n$ in terms of the nodal basis $\{\varphi_x\}$,

$$u_{0,n} = \sum_{x\in\Omega_n^{\text{int}}} \xi_x^{\text{int}}\varphi_x, \quad (\xi_x \in \mathbb{R})$$

therefore problem (2.2) becomes a linear system of equations:

$$\text{given} \quad f \in X \quad \text{and} \quad X_n \subset X \quad \text{find} \quad \xi \in \mathbb{R}^{|J_n|} \ : \ A_n\xi^{\text{int}} = F_n, \tag{2.8}$$

where

$$A_n = ((\mathcal{A}\varphi_x, \varphi_{x'}) \ : \ x, x' \in \Omega_n^{\text{int}}) \tag{2.9}$$

$$\xi^{\text{int}} = (\xi_x \ : \ x \in \Omega_n^{\text{int}}) \tag{2.10}$$

$$F_n = ((F, \varphi_x) \ : \ x \in \Omega_n^{\text{int}}) \tag{2.11}$$

are real matrices.

The matrix A_n is historically known as the *stiffness matrix*, whereas F_n is called the *load vector*. From now on, let N denote the dimension of the space X_n, also called the *number of degrees of freedom*. These notions stem from elasticity, of course it was quickly realized that the method has a wide range of different applications, but the names stuck.

Using a suitable indexing[2] of Ω_n^{int} and $\Omega_n^\partial := \Omega_n \setminus \Omega_n^{\text{int}}$ it is sometimes convenient to write the linear system for the inhomogeneous problem in the form

$$\begin{pmatrix} A_n & A_n^\partial \\ 0 & I \end{pmatrix} \begin{pmatrix} \xi^{\text{int}} \\ \xi^\partial \end{pmatrix} = \begin{pmatrix} F_n \\ g_n \end{pmatrix}, \tag{2.12}$$

where

$$A_n^\partial = ((\mathcal{A}\varphi_x, \varphi_{x'}) \ : \ x \in \Omega_n^{\text{int}}, \ x' \in \Omega_n^\partial)$$

$$g_n = (g(x) \ : \ x \in \Omega_n^\partial).$$

[2]Note that the surjective map $x \mapsto \xi_x$ implicitly fix an ordering; its actual realization is only relevant from the point of view of the programmer.

Consequently, if we let $\xi = (\xi^{\text{int}}\ \xi^{\partial})$, the approximate solution u_n is can be written as

$$u_n = \sum_{x \in \Omega_n} \xi_x \varphi_x, \quad (\xi_x \in \mathbb{R}). \tag{2.13}$$

Formally, the inverse of the block matrix in (2.12) is

$$\begin{pmatrix} A_n^{-1} & -A_n^{-1} A_n^{\partial} \\ 0 & I \end{pmatrix}, \tag{2.14}$$

a fact that will turn out to be useful later.

2.4 Discrete Maximum Principles

The previous chapter enumerated a few problems for which the maximum principle holds, so we have reason to expect that approximate solutions to these problems also posess similar features. The validity of a so-called discrete maximum principle solely depends on the discretization itself, and not on the way the resulting algebraic system is solved.

Formulating the discrete version of the maximum principle is straightforward – for example for Model Problem I, we have Theorem (1.2.7).

(2.4.1) Definition. An approximate solution u is said to satisfy the *discrete maximum principle*, if $f \leq 0$ (a.e. on Ω) $\implies \max_{\overline{\Omega}} u = \max_{\partial\Omega} u^+$.

Similarly to Theorem (1.2.7) we introduce an equivalent notion.

(2.4.2) Definition. An approximate solution u is said to satisfy the *discrete nonnegativity principle*, if $f \geq 0$ a.e. on Ω and $g \geq 0 \implies u \geq 0$.

2.5 Geometric Constraints

As is often the case with both the theory and the discretization of PDEs, the one-dimensional case is exceptional, and as far as applications are concerned it is mostly uninteresting. For lowest-order finite element discretization of the Laplacian in one dimension, the discrete maximum principle always holds for every choice of simplicial (i.e. interval) decomposition of the domain Ω.

For dimensions $d \geq 2$, the discrete maximum principle does not hold without constraints on the mesh $\mathcal{T}_n$, for counterexamples, see [11]. Even worse, by choosing higher order basis functions, these constraints may become overly restrictive. In real-world applications however, one commonly uses higher order (typically $2 \leq p \leq 4$) basis functions – there isn't a final answer for these problems yet, even in one dimension [29].

The following lemma will turn out to be of paramount importance.

(2.5.1) Characterization of nonobtuseness. Suppose $T \subset \mathbb{R}^d$ is a simplex with vertices $P = \{x_1, \ldots, x_{d+1}\} \subset \mathbb{R}^d$. Then the angle between any two sides of T is less then or equal to $\pi/2$ if and only if

$$\forall x, x' \in P \ : \ x \neq x' \implies (\nabla\varphi_x\big|_T, \nabla\varphi_{x'}\big|_T) \leq 0$$

Proof. [3] For each of the $d+1$ sides $S_k \subset T$, let an inward pointing normal be denoted by $\nu_k \in \mathbb{R}^d$. Then the barycentric coordinates of a point $x \in T$ can be written as

$$1 \leq k \leq d+1 \ : \ \lambda_k(x) = (\nu_k, x) + \mu_k,$$

with some suitable constants μ_k. Therefore $\nabla\lambda_k(x) = \nu_k$. But the condition that for $k \neq \ell$, the angle between the sides S_k and S_ℓ is $\leq \pi/2$ is equivalent to the angle between ν_k and ν_ℓ being $\geq \pi/2$, in other words that (cf. [4])

$$0 \geq \cos(\nu_k, \nu_\ell) = \frac{(\nabla\lambda_k, \nabla\lambda_\ell)}{\|\nabla\lambda_k\|\|\nabla\lambda_\ell\|}.$$

The proof is finished once we observe that the barycentric coordinate functions λ_k are nothing but the linear nodal basis functions φ_{x_k}. □

For Model problem I., the stiffness matrix A_n contains entries of the form $(\mathcal{A}\varphi_x, \varphi_{x'})$, which in turn are defined in (2.7), thus the relevance of the preceding lemma is obvious. Since $(\mathcal{A}\varphi_x, \varphi_x) > 0$ always holds, if we can guarantee the geometric condition of the lemma on the whole mesh, also called the *non-obtuseness* criterion, a certain sign-structure of the stiffness matrix also follow. This fact motivates the linear algebraic investigations of the following section.

2.6 Algebraic Maximum Principles

In the seminal work [4], Ciarlet and Raviart establish the discrete maximum principe for the simplicial, lowest order finite element discretization of the d-dimensional inhomogeneous Helmholtz equation (or equivalently the eigenproblem for the Laplacian) with Dirichlet boundary. Here, we present this linear algebraic treatment of the problem.

As a motivation, suppose we are able to prove that the stiffness matrix A_n is monotone in the following sense.

(2.6.1) Definition. A matrix $A \in \mathbb{R}^{d\times d}$ is said to be *monotone* if for every $x \in \mathbb{R}^d$, $Ax \geq 0$ implies $x \geq 0$ (pointwise) or, equivalently, if A^{-1} exists and $A^{-1} \geq 0$.

[3] A geometric proof for the case $d = 2$ can found in [12, Appendix A.].

Then, from (2.8) we have $\xi^{\text{int}} = A_n^{-1} F_n \leq 0$, whenever $F_n \leq 0$. This means that $u_n\big|_\Omega \leq 0$, so the discrete maximum principle holds for the problem with the homogeneous Dirichlet boundary. Guaranteeing however that A_h is monotone isn't obvious. The following class enjoys the required monotonity, and is often used in the systematic study of finite difference and finite element matrices, see [26].

(2.6.2) Definition. A matrix $A \in \mathbb{R}^{d\times d}$ is said to be *irreducibly diagonally dominant* if

1. A is *irreducible*, that is, for every $k \neq \ell$ there exists distinct indices $s_1, \ldots, s_m$, such that $a_{k,s_1} a_{s_1,s_2} \ldots a_{s_m,\ell} \neq 0$,

2. A is *diagonally dominant*,

$$1 \leq k \leq d \ : \ \sum_{\ell \neq k} |a_{k\ell}| \leq |a_{kk}|,$$

3. there exists $1 \leq m \leq d$, such that

$$\sum_{\ell \neq m} |a_{m\ell}| < |a_{mm}|.$$

Diagonal dominance is usually not too hard to check, but irreducibility often seems impossible. Nevertheless, if we also have the sign-structure mentioned in the previous section (i.e., we have a non-obtuse mesh), the following theorem holds.

(2.6.3) Theorem. [26] Suppose A is an irreducibly diagonally dominant matrix. Furthermore, suppose that the diagonal of A is strictly positive, and the off-diagonal entries of A are nonpositive. Then $A^{-1} > 0$.

As noted by J. Karátson and S. Korotov [21], the quintessence of the discrete maximum principles lies in the following observation regarding the block matrix equation (2.12).

(2.6.4) Karátson–Korotov Theorem. Consider the matrix equation (2.12) and suppose the following:

1. diag $A_n > 0$, offdiag $A_n \leq 0$,
2. $A_n^\partial \leq 0$,
3. $A_n \mathbb{1}^{\text{int}} + A_n^\partial \mathbb{1}^\partial \geq 0$, where $\mathbb{1}^{\text{int}} = (1, \ldots, 1)^\mathsf{T}$ and $\mathbb{1}^\partial = (1, \ldots, 1)^\mathsf{T}$,
4. A_n is irreducibly diagonally dominant.

Statement.

$$\forall \xi = (\xi^{\text{int}}\ \xi^\partial)^\mathsf{T} \in \mathbb{R}^N \ : \ A_n \xi^{\text{int}} + A_n^\partial \xi^\partial \leq 0 \implies \max_x \xi_x \leq \max\{0, \max_x \xi_x^\partial\}$$
$$\Downarrow \text{ if also } A_n \mathbb{1}^{\text{int}} + A_n^\partial \mathbb{1}^\partial = 0$$
$$\max_x \xi_x = \max_x \xi_x^\partial$$

From this theorem one readily deduces the discrete maximum principle for the lowest-order finite element discretization of a number of problems. For example we have the following.

(2.6.5) Discrete Maximum Principle. Suppose we have a lowest order finite element discretization on a nonobtuse mesh of the Model problem I. Then the discrete maximum principle holds.

2.7 Discrete Green's functions

For the finite difference discretization of the Laplacian, Stoyan [18] uses the explicit form of the classical Green's function (which can be written in terms of the eigenstructure on an n-block) to obtain various *a priori* estimates and the discrete maximum principle. On a general triangle mesh, the concrete form of Green's functions are unknown (since the eigenstructure is unknown), therefore one must look for a discrete analogue to pursue this path. Note that although certain approximate forms of Green's functions appear in the applications of numerical analysis, we only view the concept as a theoretical tool or curiosity here; its use can be completely avoided.

A discrete Green's function, introduced by Ciarlet and Varga in 1970 [5] is a type of Aronszajn–Bergman reproducing kernel on a Hilbert space (see [1] and [2]). In [11] numerical experiments were conducted using the discrete Green's function. Vejchodskỳ also uses this approach in his survey [28]. The material presented in this section basically follows [28].

We consider Model problem I., one could also add a convective term and impose general boundary conditions, but one must be aware that Green's function is conceptually valid only for

linear problems. A (variational) *discrete Green's function* $G(\cdot, y) \in X_n$ concentrated on a point $y \in \overline{\Omega}$ is defined as

$$\forall v \in X_n \ : \ (\mathcal{A}G(\cdot, y), v) = v(y). \tag{2.15}$$

Existence and uniqeness questions are in this case automatically answered: this is a finite dimensional regular linear system of equations, due to properties inherited from the weak formulation.

In [28], a delicacy regarding the Dirichlet boundary is pointed out. Namely, we seek a solution u of the form $u = u_0 + g \in X_n$, where $u_0 \in X_n$ is the solution of homogeneous-boundary problem

$$\forall v \in X_n \ : \ (\mathcal{A}u_0, v) = (f, v) - (\mathcal{A}g, v),$$

but in general $g \notin X_n$. Therefore the boundary data g has to be approximated in X_n too, for example by an orthogonal projection $\pi g \in X_n$ – the so-called *elliptic projection* –, defined as

$$\forall v \in X_n \ : \ (\mathcal{A}v, g - \pi g) = 0. \tag{2.16}$$

By writing $v := u_0 + \pi g$ in (2.15) we get

$$\begin{aligned} u_0(y) + (\pi g)(y) &= (\mathcal{A}G(\cdot, y), u_0 + \pi g) \\ &= (\mathcal{A}G(\cdot, y), u_0 + g) + (\mathcal{A}G(\cdot, y), \pi g - g) \\ &= (G(\cdot, y), f). \end{aligned}$$

This is the *discrete Green's representation formula*

$$\forall y \in \overline{\Omega} \ : \ u(y) = \int_\Omega G(x, y) f(y)\, dy + g(y) - (\pi g)(y). \tag{2.17}$$

This pleasant-looking[4] formula is very much relevant to the discrete maximum principle.

(2.7.1) Proposition. The discrete nonnegativity principle holds for Model problem I. iff $G \geq 0$ on $\overline{\Omega} \times \overline{\Omega}$ and $g - \pi g \geq 0$, $g \geq 0$ on Ω.

To advance one step further to our goal, we now present an explicit formula for the discrete Green's function, see [28].

(2.7.2) Lemma. Suppose that X_n is the lowest-order finite element space, and A is the stiffness matrix, as before. Then, we have

$$\forall a, b \in \overline{\Omega} \ : \ G(a, b) = \sum_{x \in \Omega_n^{\text{int}}} \sum_{y \in \Omega_n^{\text{int}}} (A^{-1})_{x,y} \varphi_x(a) \varphi_y(b). \tag{2.18}$$

[4]Numerical experiments show that G is similarly badly behaved at the diagonal as its classical counterpart.

Proof. Let

$$G(\cdot, b) = \sum_y \gamma_y(b)\varphi_y, \tag{2.19}$$

so by (2.15) we have for all $x, b \in \Omega_n^{\text{int}}$

$$\varphi_x(b) = \sum_y \gamma_y(b)(\mathcal{A}\varphi_y, \varphi_x) = \sum_y A_{x,y}\gamma_y(b),$$

solving this for $\{\gamma_y(b) \ : \ y \in \Omega_n^{\text{int}}\}$ we get

$$\gamma_y(b) = \sum_x (A^{-1})_{x,y}\varphi_y(b).$$

Substituting this back into (2.19), we are finished. □

This formula gives a deeper explanation of the occurance of the inverse matrix A^{-1} in the previous section. Again referring to the work of Vejchodskỳ, numerical expriments can be done using this expression of the Green's function, but this is limited to simple problems for understandable reasons.

The term $g - \pi g$ also needs expanding in terms of the basis.

(2.7.3) Lemma. Let $g = \sum_{x\in\Omega_n} \zeta_x\varphi_x$, then

$$g - \pi g = \sum_{x\in\Omega_n^\partial} \zeta_x(\varphi_x - \pi\varphi_x),$$
$$\pi\varphi_x = \sum_{y,z\in\Omega_n^{\text{int}}} (A^{-1})_{y,z} A^\partial_{z,x}\varphi_y.$$

Proof. The first equation is a consequence of the fact that π leaves the basis functions corresponding to internal nodes fixed. As for the second equation, let $x \in \Omega_n^\partial$ and

$$\pi\varphi_x = \sum_{y\in\Omega_n^{\text{int}}} \sigma_{x,y}\varphi_y. \tag{2.20}$$

Then by the definition of the projection π, (2.16), we have for every $z \in \Omega_n^{\text{int}}$

$$(\mathcal{A}\varphi_x, \varphi_z) = \sum_{y\in\Omega_n^{\text{int}}} \sigma_{x,y}(\mathcal{A}\varphi_y, \varphi_z),$$

or,

$$A^\partial_{z,x} = \sum_{y\in\Omega_n^{\text{int}}} \sigma_{x,y}A_{z,y}.$$

Therefore

$$\sigma_{x,y} = \sum_{z\in\Omega_n^{\text{int}}} A^\partial_{z,x}(A^{-1})_{z,y},$$

which can be substituted back into (2.20). □

The following famous theorem is a conseqence of [14, Theorem 5.1], and is of utmost importance to us.

(2.7.4) Theorem. Let A be positive definite. If $\operatorname{offdiag}(A) \leq 0$, then $A^{-1} \geq 0$.

(2.7.5) Proposition. The discretization of Model problem I. (2.12) satisfies the discrete nonnegativity principle iff $A^{-1} \geq 0$ and $A^{-1}A^{\partial} \leq 0$.

Proof. From the previous two technical lemmas we have the following equivalences:

$$\begin{aligned} G \geq 0 &\iff A^{-1} \geq 0, \\ g - \pi g \geq 0 &\iff A^{-1}A^{\partial} \leq 0, \end{aligned}$$

from which the theorem follows. □

(2.7.6) Proposition. If $\operatorname{offdiag}(A) \leq 0$ and $A^{\partial} \leq 0$, then the discretization of Model problem I. (2.12) satisfies the discrete nonnegativity principle.

Proof. Since the is positive definiteness of A is inherited from the continuous problem, using Theorem (2.7.4), we get $A^{-1} \geq 0$. □

(2.7.7) Discrete Maximum Principle. Suppose we have a lowest order finite element discretization on a nonobtuse mesh of the Model problem I. Then the discrete maximum principle holds.

As noted by Vejchodskỳ, this approach using discrete Green's functions completely avoids the concept of irreducibility.

2.8 Variational Approach to Discrete Maximum Principles

The novel approach of Diening, Kreuzer and Schwarzacher [9] extends the validity of the discrete maximum principle to the lowest order finite element discretization of vector-valued monotone Nemyckii operators. In what follows, we present their results without discussing the vector-valued case, since this would not fit in our framework developed thus far.

Let $P^1(\mathcal{T})$ denote the set of continuous functions u, such that $u\big|_T$ is a linear d-variate polynomial for every simplex $T \in \mathcal{T}$ in a conformal simplicial decomposition of the bounded polyhedral domain Ω. (Note that we dropped the subscript n from $\mathcal{T}_n$ for convenience.) The set $P^1(\mathcal{T})$ forms a (normed) vector lattice[5], a structure that is necessary for the formulation of maximum principles – note that H^1 is also a vector lattice. Furthermore, let $P_0^1(\mathcal{T})$ denote the subspace of such

[5]Where the order is understood to be node-wise.

functions that vanish on $\partial\Omega$. Obviously the spaces $P^1(\mathcal{T})$ and $P_0^1(\mathcal{T})$ are spanned by the sets $\{\varphi_x \,:\, x \in \Omega_n\}$ and $\{\varphi_x \,:\, x \in \Omega_n^{\text{int}}\}$.

Let $u \in g + P_0^1(\mathcal{T})$, where $g \in P^1(\mathcal{T}_n)$. Then the discrete maximum principle says that whenever $f \leq 0$, then we *should* have

$$\max u(\Omega) \leq \gamma := \max\{0, \max u(\partial\Omega)\}. \tag{2.21}$$

Linear elements have the extremely useful property that they attain their maximum at nodes – a property that is not posessed by higher-order elements, even in one dimension. Therefore (2.21) is equivalent to

$$\max u(\Omega_n) \leq \gamma, \tag{2.22}$$

i.e. the maxima is taken over the nodes Ω_n determined by the conformal simplicial decomposition $\mathcal{T}$. The key observation in [9] for extending the discrete maximum principle to vector valued problems is that $u(\Omega_n)$ is contained in a *convex closed* set, in our case $C := (-\infty, \gamma]$. As is customary in the engineering practice, if one obtains an approximate solution that violates the maximum principle, one simply „cuts off" or „clamps" to the theoretically permissible range C. Geometrically, this corresponds to a projection π onto the set C, a notion that readily generalizes to the vector-valued case.

(2.8.1) Remark. Such a projection π preserves boundary values, that is, if $g \in P^1(\mathcal{T})$ and $g(\partial\Omega) \subset C$, then $\pi u \in g + P_0^1(\mathcal{T})$ for all $u \in g + P_0^1(\mathcal{T})$.

In order to have some control over this heuristic procedure, we need estimates guaranteeing that the solution is improved in some sense – preferably in the energy norm. This kind of investigation hasn't been done in depth until recently [22]. The following crucial lemma is due to [9], which we formulate only for the scalar case for the sake of simplicity; πu is just $u \wedge \gamma$ in our case.[6]

(2.8.2) Lemma. Let $\mathcal{T}$ be a non-obtuse simplicial decomposition. Then,

$$\forall \gamma \in \mathbb{R} \;\; \forall u \in P^1(\mathcal{T}) \;\; \forall T \in \mathcal{T} \;:\; \big(\nabla u, \nabla(u \wedge \gamma)\big) \geq \|\nabla(u \wedge \gamma)\|^2 \;\text{ on }\; T.$$

Proof. Since this is an element-wise estimate, it suffices to consider it on a fixed simplex $T \in \mathcal{T}$. Let $T = \operatorname{Hull} Y$, where $Y = \{y_0, \ldots, y_d\}$, then, from one of our previous observations we have

$$u\big|_T = \sum_{y \in Y} u(y)\, \varphi_y\big|_T,$$
$$\nabla u\big|_T = \sum_{y \in Y} u(y)\, \nabla\varphi_y\big|_T.$$

[6] Here the $\wedge$ denotes the infimum.

Let us rewrite the inner product on the left-hand side of the estimate as

$$\begin{aligned}\big(\nabla u\big|_T, \nabla(u\wedge\gamma)\big|_T\big) &= \left(\sum_{y\in Y} u(y)\,\nabla\varphi_y\big|_T, \sum_{z\in Y}(u(z)\wedge\gamma)\,\nabla\varphi_z\big|_T\right)\\ &= \sum_{y,z\in Y} u(y)(u(z)\wedge\gamma)\,\big(\nabla\varphi_y\big|_T, \nabla\varphi_z\big|_T\big)\\ &= \sum_{y\in Y}\Bigg[\sum_{z\in Y} u(y)(u(z)\wedge\gamma)\,\big(\nabla\varphi_y\big|_T, \nabla\varphi_z\big|_T\big) - \underbrace{\sum_{z\in Y} u(y)(u(y)\wedge\gamma)\,\big(\nabla\varphi_y\big|_T, \nabla\varphi_z\big|_T\big)}_{0,\text{ because of (2.6)}}\Bigg]\\ &= \sum_{y\in Y}\sum_{z\in Y} u(y)\big[u(z)\wedge\gamma - u(y)\wedge\gamma\big]\,\big(\nabla\varphi_y\big|_T, \nabla\varphi_z\big|_T\big).\end{aligned}$$

Due to the non-obtuseness condition, the inner products in this last expression are all nonpositive. Moreover, the other terms of the product can be estimated as

$$u(y)\big[u(z)\wedge\gamma - u(y)\wedge\gamma\big] \le (u(y)\wedge\gamma)\big[u(z)\wedge\gamma - u(y)\wedge\gamma\big],$$

because if $u(y) > \gamma$, then the above is equivalent to $(u(y)-\gamma)(u(z)\wedge\gamma-\gamma)\le 0$ a trivial relation; and equality holds iff $u(y)\le\gamma$. Summarizing, we have

$$\begin{aligned}\big(\nabla u\big|_T, \nabla(u\wedge\gamma)\big|_T\big) &\ge \sum_{y\in Y}\sum_{z\in Y}(u(y)\wedge\gamma)\big[u(z)\wedge\gamma - u(y)\wedge\gamma\big]\,\big(\nabla\varphi_y\big|_T, \nabla\varphi_z\big|_T\big)\\ &= \sum_{y\in Y}\sum_{z\in Y}(u(y)\wedge\gamma)(u(z)\wedge\gamma)\,\big(\nabla\varphi_y\big|_T, \nabla\varphi_z\big|_T\big) = \|\nabla(u\wedge\gamma)\|^2 \qquad \square\end{aligned}$$

(2.8.3) Corollary. An application of the Schwarz inequality yields

$$\forall\gamma\in\mathbb{R}\ \ \forall u\in P^1(\mathcal{T})\ \ \forall T\in\mathcal{T}\ :\ \|\nabla(u\wedge\gamma)\|\le\|\nabla u\|\ \text{ on }\ T.$$

Having obtained a low-level building block, let us examine a few energy functionals and try to deduce the discrete maximum principle. The key observation is the following. If we have $j(u\wedge\gamma)\le j(u)$, and u is the *unique* minimizer in the Ritz–Galerkin discretization space, then we necessarily have $u\wedge\gamma = u$. In other words $u\le\gamma$ (on Ω), from which the discrete maximum principle follows.

(2.8.4) Example. [9] Suppose that the Lagrangian is of the form $L(x,\eta,\xi) := L(\|\eta\|)$ and it is monotone increasing. If $f\le 0$, then obviously for the energy functional

$$j(u) = \int_\Omega L(\|\nabla u\|)\,dx - \int_\Omega f u\,dx,$$

we have

$$u - g\in P_0^1(\mathcal{T})\ :\ j(u\wedge\gamma)\le j(u),$$

where $\gamma = \max\{0, \max u(\partial\Omega)\}$ and j is defined in (2.4). This covers the *p-Laplace operator*, which has the Lagrangian

$$L(x, \xi, \eta) = \tfrac{1}{p}\|\eta\|^p.$$

We know that in the current *ansatz*, there exists a unique minimizer $\overline{u}$ of $j(\cdot)$ on $g + P_0^1(\mathcal{T})$. Since the boundary values are preserved after the cutoff, $\overline{u} \wedge \gamma \in g + P_0^1(\mathcal{T})$, so $j(\overline{u} \wedge \gamma) \leq j(\overline{u})$, therefore by uniqueness $\overline{u} \wedge \gamma = \overline{u}$. This precisely means that the discrete maximum principle holds.

(2.8.5) Example. Consider the nonlinear reaction-diffusion problem of Example (1.5.3). We can expect the maximum principle to hold (and in particular the nonpositivity principle), thus set $\gamma := 0$. Recall that this problem has the energy functional

$$u \in P_0^1(\Omega) \ : \ j(u) = \int_\Omega \frac{1}{2}\|\nabla u\|^2 \, dx + \int_\Omega \frac{1}{q}|u|^q \, dx - \int_\Omega f u \, dx.$$

Suppose that $f \leq 0$. From Corollary (2.8.3) and (1.18) we have $j(u \wedge \gamma) \leq j(u)$ for $u \in P_0^1(\mathcal{T})$, from which the discrete maximum principle follow as before.

Bibliography

[1] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc, 68 (1950), pp. 337–404.

[2] N. Aronszajn and K. T. Smith, *Characterization of positive reproducing kernels. applications to green's functions*, American Journal of Mathematics, (1957), pp. 611–622.

[3] J. Chabrowski, *Variational Methods for Potential Operator Equations: With Applications to Nonlinear Elliptic Equations*, De Gruyter Studies in Mathematics Series, De Gruyter, 1997.

[4] P. G. Ciarlet and P.-A. Raviart, *Maximum principle and unform convergence for the finite element method*, Computer Methods in Applied Mechanics and Engineering, (1973), pp. 17–31.

[5] P. G. Ciarlet and R. Varga, *Discrete variational green's function*, Numerische Mathematik, 16 (1970), pp. 115–128.

[6] J. Conway, *Functions of One Complex Variable I*, Functions of one complex variable / John B. Conway, Springer, 1978.

[7] R. Courant, *Dirichlet's principle, conformal mappings, and minimal surfaces*, Interscience Publishers, Inc., 1950.

[8] B. Dacorogna, *Introduction To The Calculus Of Variations*, Imperial College Press, 2004.

[9] L. Diening, C. Kreuzer, and S. Schwarzacher, *Convex Hull Property and Maximum Principle for Finite Element Minimisers of General Convex Functionals*, ArXiv e-prints, (2013).

[10] P. Drábek, *Lectures on Nonlinear Analysis*, Vydavatelsky Servis, 2004.

[11] A. Drăgănescu, T. Dupont, and L. Scott, *Failure of the discrete maximum principle for an elliptic finite element problem*, Mathematics of computation, 74 (2005), pp. 1–23.

[12] A. Drăgănescu, T. F. Dupont, and L. R. Scott, *Failure of the discrete maximum principle for an elliptic finite element problem*, Math. Comp, (2005), pp. 1–23.

[13] I. Faragó and J. Karátson, *Numerical Solution of Nonlinear Elliptic Problems Via Preconditioning Operators: Theory and Practice*, Advances in Computation : Theory and Practice, Volume 11, Nova Science Pub Incorporated, 2002.

[14] M. Fiedler, *Special matrices and their applications in numerical mathematics*, Martinus Nijhoff Publishers, 1986.

[15] S. Fomin and R. Silverman, *Calculus of Variations*, Dover Books on Mathematics Series, Dover Publ., 2000.

[16] L. Fraenkel, *An Introduction to Maximum Principles and Symmetry in Elliptic Problems*, Cambridge Tracts in Mathematics, Cambridge University Press, 2000.

[17] D. Gilbarg and N. Trudinger, *Elliptic partial differential equations of second order*, Classics in Mathematics Series, Springer London, Limited, 2001.

[18] S. Gisbert, *Numerikus módszerek 2-3.*, Typotex Elektronikus Kiadó Kft.

[19] J. Jost and X. Li-Jost, *Calculus of Variations*, Cambridge Studies in Advanced Mathematics, Cambridge University Press, 1998.

[20] J. Karátson, *Numerical functional analysis (in Hungarian)*, `http://www.cs.elte.hu/~karatson/nfa.pdf`.

[21] J. Karátson and S. Korotov, *Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions*, Numer. Math., 99 (2005), pp. 669–698.

[22] C. Kreuzer, *A note on why enforcing discrete maximum principles by a simple a posteriori cutoff is a good idea*, arXiv preprint arXiv:1208.3958, (2012).

[23] P. Lindqvist, *Notes on the p-Laplace equation*, `http://www.math.ntnu.no/~lqvist/p-laplace.pdf`, 2013.

[24] P. Pucci and J. Serrin, *The strong maximum principle revisited*, Journal of Differential Equations, 196 (2004), pp. 1–66.

[25] P. Pucci and J. Serrin, *The Maximum Principle*, Progress in Nonlinear Differential Equations and Their Applications, Birkhäuser, 2007.

[26] R. Varga, *Matrix Iterative Analysis*, Springer Series in Computational Mathematics, Springer, 2009.

[27] J. L. Vázquez, *A strong maximum principle for some quasilinear elliptic equations*, Applied Mathematics and Optimization, 12 (1984), pp. 191–202.

[28] T. Vejchodskỳ, *The discrete maximum principle for galerkin solutions of elliptic problems*, Central European Journal of Mathematics, 10 (2012), pp. 25–43.

[29] T. Vejchodskỳ and P. Šolín, *Discrete maximum principle for higher-order finite elements in 1d*, Mathematics of Computation, 76 (2007), pp. 1833–1846.