Numer. Math. (2005) 99: 485–508
Digital Object Identifier (DOI) 10.1007/s00211-004-0554-5

Numerische
Mathematik

Discrete minimum and maximum principles for finite element approximations of non-monotone elliptic equations

Ansgar Jüngel[1], Andreas Unterreiter[2]

[1] Fachbereich Mathematik und Informatik, Universität Mainz, Staudingerweg 9, 55099 Mainz, Germany; e-mail: juengel@mathematik.uni-mainz.de

[2] Institut für Mathematik, MA 6-3, TU Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany; e-mail: unterreiter@math.tu-berlin.de

Received January 8, 2004 /
Published online: November 26, 2004 – © Springer-Verlag 2004

Summary. Uniform lower and upper bounds for positive finite-element approximations to semilinear elliptic equations in several space dimensions subject to mixed Dirichlet-Neumann boundary conditions are derived. The main feature is that the non-linearity may be non-monotone and unbounded. The discrete minimum principle provides a positivity-preserving approximation if the discretization parameter is small enough and if some structure conditions on the non-linearity and the triangulation are assumed. The discrete maximum principle also holds for degenerate diffusion coefficients. The proofs are based on Stampacchia's truncation technique and on a variational formulation. Both methods are settled on careful estimates on the truncation operator.

Mathematics Subject Classification (2000): 65N30, 65N12

1 Introduction

In this paper uniform lower and upper bounds for finite-element discretizations of semi-linear elliptic boundary-value problems are derived. In short the following type of PDEs is considered:

$$(1) \qquad Lu = g(x,u) \quad \text{in } \Omega, \qquad u = u_D \quad \text{on } \Gamma_D, \qquad u_\nu = 0 \quad \text{on } \Gamma_N,$$

Correspondence to: A. Unterreiter

where L is the second-order differential operator

$$Lu = -\sum_{i,j=1}^{d} \partial_i(a_{ij}(x)\partial_j u) + \sum_{i=1}^{d} a_i(x)\partial_i u, \tag{2}$$

the function $g(x, u)$ may be non-monotone, and u_ν is the normal derivative of u associated to the operator L. The domain $\Omega \subset \mathbb{R}^d$ ($d \geq 1$) is bounded with boundary $\Gamma_D \cup \Gamma_N$. The *precise* assumptions are to be found in the next section.

The investigations are motivated by numerical approximations of the stationary quantum drift-diffusion model

$$\begin{aligned} \delta^2\Delta\sqrt{n} &= \sqrt{n}(\log(\sqrt{n}) + V - F), \\ \operatorname{div}(n\nabla F) &= 0, \\ -\Delta V &= n - C(x) \quad \text{in } \Omega, \end{aligned}$$

for the electron density n, the quantum quasi-Fermi potential F, and the electrostatic potential V. The parameter δ is the (scaled) Planck constant, and the prescribed function $C = C(x)$ is the concentration of fixed background charges [2,11]. The equations are supplemented by mixed Dirichlet-Neumann boundary conditions. The model describes the distribution of electrons in semiconductor devices whose performance relies on quantum-mechanical effects. Typically this model is used to simulate inversion layers in MOSFET devices [1] or to compute current-voltage characteristics of resonant tunneling diodes [12,18,20]. For $\delta = 0$ the model equations reduce to the classical drift-diffusion model [17].

In a Gummel-type iteration procedure [20] one has to solve for fixed F and V the equation

$$\delta^2\Delta u = g(x, u) := u(\log u + f(x)), \quad \text{in } \Omega, \tag{3}$$

where $u = \sqrt{n} \geq 0$ and $f(x) = V(x) - F(x)$. Here the function $g(x, u)$ is *not* monotone. It is important for a numerical scheme solving (3) to have the following two properties:

- The numerical approximation of the particle density $n(x)$ has to be positive.
- Uniform estimates on the numerical solution should be independent of the scaled Planck constant δ.

In this paper we prove that the linear finite-element approximation actually has these properties. More precisely, the numerical method is *positivity-preserving* and the upper bounds are *independent* of the parameter δ.

We remark that estimates independent of δ have been proved for the quantum drift-diffusion model (3)–(3) in [19] in a one-dimensional setting.

The peculiar non-linearity in (3) is the sum of a monotone and a bounded function. A combination of discrete maximum principles for bounded [6, Sec. 20] and monotone non-linearities [13] may be applied. However, the validity of discrete minimum and maximum principles does not rely on this specific structure.

There is a vast literature on discrete maximum principles, whereas much less references can be found for uniform positive lower bounds [13]. Discrete maximum principles for (linear) finite-element approximations have been first derived in [7] for linear elliptic equations. The method has been extended in [9] for a special system of non-linear equations. Other techniques are based on elliptic estimates [21] or matrix properties [10]. Stampacchia's method has been also applied to linear discrete variational inequalities [8]. It is well known that the validity of discrete maximum principles is closely related to geometric properties of the finite-element meshes, see, e.g., [5,7,14,15]. Discrete maximum principles for convection-diffusion equations have been derived in [4]. They also have been studied for finite-volume [3] and finite-difference schemes [16].

Let us check the paper's main results in advance. Let u_h be a (piecewise linear) finite-element approximation of (1) and let $g(x,u)$ be a Carathéodory function (a precise definition will be given later on) such that $g(x,u) \leq \overline{g}(x)$ for $x \in \Omega$ and $u \in \mathbb{R}$ and $g(x,u) \geq \underline{g}(x)$ for $x \in \Omega$ and $u \leq m_0$ for some $m_0 \in \mathbb{R}$, where $\overline{g}(x)$ and $\underline{g}(x)$ are L^p functions. Then, under some assumptions on the differential operator and the triangulation, there exist positive constants C_1, C_2, and α, independent of the maximal size h of the elements of the finite-element triangulation of Ω, such that

$$\min_{\Omega} u_h \geq \min\{m_0, \min_{\Gamma_D} u_{D,h}\} - C_1 \|\underline{g}\|_{L^p(\Omega)}^{1/2} - C_2 h^{\alpha}, \tag{4}$$

where $u_{D,h}$ is an approximation of u_D. Hence, for positive Dirichlet data, positive m_0, and sufficiently small $\underline{g}$ and h, the approximation u_h is strictly positive. Kerkhoven and Jerome [13] derived a similar result with $\alpha = 2$, however, only for *monotone* non-linearities. Our result applies to more general non-linearities and to space dimensions $d \leq 5$.

The proof is based on the Stampacchia truncation method. For a *continuous* weak solution u of (1), this technique provides the estimate

$$\inf_{\Omega} u \geq \min\{m_0, \inf_{\Gamma_D} u_D\} - C_0 \|\underline{g}\|_{L^p(\Omega)}, \tag{5}$$

where $C_0 > 0$ is some constant. Estimate (5) will follow if one uses the truncated function $(-u+m)^+ = \max\{0, -u+m\}$ as a test function in the weak formulation of (1) (see Section 3.1 for details). When setting $h = 0$ in (4), we do *not* recover the estimate (5). This is because the truncated function $(-u+m)^+$ cannot be used as test function in the *discretized* version of (1).

Instead of $(-u+m)^+$ we use a projection of $(-u+m)^+$ on the finite-element space as test function. This yields extra terms which are carefully estimated and (4) follows.

The second main result is a discrete maximum principle for equilibrium solutions which are minimizers of an energy functional associated with (1). We give sufficient conditions such that discrete equilibrium solutions are finite-element solutions. Furthermore, assuming (essentially) that there exists a constant $M_0 \in \mathbb{R}$ such that the primitive of $g(x, \cdot)$ is strictly increasing on (M_0, ∞), we prove that any equilibrium solution u_h of (1) satisfies

$$\sup_{\Omega} u_h \leq \max\{M_0, \inf_{\Gamma_D} u_{D,h}\}.$$

This result also holds for degenerate diffusion matrices. Thus, when applied to (3) the maximum principle holds for any value of the Planck constant $\delta \geq 0$. We can even allow for diffusion coefficients vanishing on (parts of) Ω. The proof of this result is based on estimates for the projected test function.

The paper is organized as follows. In the Section 2 we state our main hypotheses and we prove some auxiliary results. Section 3 is devoted to the discrete minimum principle. The discrete maximum principle is shown in Section 4.

2 Main assumptions and auxiliary results

We impose the following assumptions:

(A1) $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, is a bounded polyhedral domain, Γ_N is a measurable open subset of $\partial\Omega$, and $\Gamma_D = \partial\Omega \setminus \Gamma_N$.

(A2) $g : \Omega \times \mathbb{R} \to \mathbb{R}$ is a Carathéodory function, i.e., g is measurable and for all $x \in \Omega$, the function $g(x, \cdot) : \mathbb{R} \to \mathbb{R}$ is continuous.

(A3) $a_{ij} : \Omega \to \mathbb{R}$, $i, j = 1, \ldots, d$, are bounded, measurable functions. For each $x \in \Omega$ the matrix $(a_{ij}(x))_{i,j=1,\ldots,d}$ is symmetric and positive semi-definite. The functions $a_i : \Omega \to \mathbb{R}$, $i = 1, \ldots, d$, are bounded and measurable.

(A4) $u_D \in H^1(\Omega) \cap L^\infty(\Omega)$.

Let $C_0^\infty(\Omega \cup \Gamma_N)$ be the set of restrictions of functions $\phi \in C_0^\infty(\mathbb{R}^d)$ to Ω such that $\mathrm{supp}(\phi) \cap \partial\Omega \subset \Gamma_N$. Furthermore, let $H_0^1(\Omega \cup \Gamma_N)$ be the closure of $C_0^\infty(\Omega \cup \Gamma_N)$ in $H^1(\Omega)$ [23].

Each polyhedral domain has a Lipschitzian boundary. Thus the following Poincaré-Sobolev inequality holds if $\mathrm{meas}_{d-1}(\Gamma_D) > 0$:

$$\|u\|_{L^r(\Omega)} \leq C_s(r)\|\nabla u\|_{L^2(\Omega)} \quad \forall u \in H_0^1(\Omega \cup \Gamma_N), \tag{6}$$

where $r < \infty$ (if $d \leq 2$) and $r = 2d/(d-2)$ (if $d \geq 3$).

The finite-element approximation relies on a weak formulation of (1). The bilinear form associated with L is $a : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$, defined by

$$a(u, v) = \sum_{i,j=1}^{d} \int_\Omega a_{ij}(\partial_i u)(\partial_j v)\, dx + \sum_{i=1}^{d} \int_\Omega a_i(\partial_i u)v\, dx.$$

Assumption (A3) implies the existence of a constant $K > 0$ such that

$$a(u, v) \le K \|\nabla u\|_{L^2(\Omega)} \|v\|_{H^1(\Omega)} \quad \forall u, v \in H^1(\Omega). \tag{7}$$

We introduce the functional $F : H^1(\Omega) \times H^1_0(\Omega \cup \Gamma_N) \to \mathbb{R} \cup \{\infty\}$ by

$$F[u](v) = \begin{cases} \displaystyle\int_\Omega g(x, u)v\, dx & \text{if } g(., u)v \in L^1(\Omega) \\ \infty & \text{if } g(., u)v \notin L^1(\Omega). \end{cases}$$

Then the weak formulation of (1) reads

$$a(u, \phi) = F[u](\phi) \quad \forall \phi \in H^1_0(\Omega \cup \Gamma_N), \quad u - u_D \in H^1_0(\Omega \cup \Gamma_N). \tag{8}$$

For the finite-element discretization we assume:

(A5) T_h is an admissible, regular triangulation of $\overline{\Omega}$ in the sense of Ciarlet [6], made up of d-simplices $\tau \in T_h$.
(A6) The edges of each simplex $\tau \in T_h$ which are part of $\partial\Omega$ are entirely contained either in Γ_D or Γ_N.

The involved finite-element spaces are

$$X_h := \{v_h \in C(\overline{\Omega}) : v_h|_\tau \text{ is affine for all } \tau \in T_h\},$$
$$V_h := \{v_h \in X_h : v_h = 0 \text{ on } \Gamma_D\}.$$

Let x_i $(1 \le i \le N)$, x_i $(N + 1 \le i \le N + N_N)$, and x_i $(N + N_N + 1 \le i \le N_h := N + N_N + N_D)$ be the vertices of T_h that belong to Ω, to Γ_N, and to Γ_D, respectively. Furthermore, ϕ_i $(1 \le i \le N_h)$ are functions of X_h defined via

$$\phi_i(x_j) = \delta_{ij}, \qquad 1 \le i, j \le N_h,$$

i.e., the functions ϕ_i $(1 \le i \le N + N_N)$ and ϕ_i $(1 \le i \le N_h)$ are a basis of V_h or of X_h, respectively. The finite-element discretization of (8) is

$$a(u_h, v_h) = F[u_h](v_h) \quad \forall v_h \in V_h, \quad u_h - u_{D,h} \in V_h, \tag{9}$$

where $u_{D,h} \in X_h$ is an approximation of u_D. Finally we assume

(A7) The matrix $(a(\phi_i, \phi_j))_{ij}$ is an L_0 matrix, i.e. $a(\phi_i, \phi_j) \le 0$ for all $i \ne j$.
(A8) $u_{D,h} \in X_h$.

Assumption (A7) is a condition on the triangulation. As an example, for $L =$ Laplacian and $d = 2$, (A7) is satisfied if all angles of the triangles of T_h are not larger than $\pi/2$ [6, Thm. 20.2].

The starting point for the Stampacchia truncation method in the continuous equation is to use the truncated function $(u - M)^+ = \max\{0, u - M\} \in H_0^1(\Omega \cup \Gamma_N)$ with $M \in \mathbb{R}$ as test function in (8). However, $(u - M)^+$ is usually *not* in X_h. As a consequence, we cannot use $(u - M)^+$ as test function in (9). Instead we test (9) with the projected function

$$[u_h - M]^\pm := \sum_{i=1}^{N+N_N} (u_h(x_i) - M)^\pm \phi_i,$$

where $(u)^+ := \max\{0, u\}$ and $(u)^- := \min\{0, u\}$. We observe

Lemma 1. *Let (A1)–(A3), (A5)–(A7) hold and let L be uniformly elliptic, i.e., there is a constant $k > 0$ such that $a(u, u) \geq k\|u\|^2_{L^2(\Omega)}$ for all $u \in H_0^1(\Omega \cup \Gamma_N)$. Furthermore, let $v_h \in X_h$ and let $M \geq \sup_{\Gamma_D} v_h$. Then*

$$\|\nabla[v_h - M]^+\|_{L^2(\Omega)} \leq K_0 \|\nabla v_h\|_{L^2(\Omega)},$$

where $K_0 = (K/k)\sqrt{C_s(2)^2 + 1}$, and K and $C_s(2)$ are defined in (7), (6), *respectively.*

Remark 2. If $L =$ Laplacian, it is not difficult to check that $K_0 = 1$. In fact, the above estimate holds for *any* bilinear form $a(\cdot, \cdot)$ of a differential operator L provided the triangulation ensures that $a(\phi_i, \phi_j)$ is an L_0 matrix.

Proof of Lemma 1. Since $a(M, [v_h - M]^+) = 0$ we deduce via (A7)

$$\begin{aligned} & a(v_h, [v_h - M]^+) \\ & = a([v_h - M]^+, [v_h - M]^+) + a([v_h - M]^-, [v_h - M]^+) \\ & \geq k\|\nabla[v_h - M]^+\|^2_{L^2(\Omega)} \\ & \quad + \sum_{i \neq j} (v_h(x_i) - M)^-(v_h(x_j) - M)^+ a(\phi_i, \phi_j) \\ & \geq k\|\nabla[v_h - M]^+\|^2_{L^2(\Omega)}. \end{aligned} \tag{10}$$

Since $[v_h - M]^+ \in H_0^1(\Omega \cup \Gamma_N)$ via (6), (7),

$$\begin{aligned} k\|\nabla[v_h - M]^+\|^2_{L^2(\Omega)} & \leq a(v_h, [v_h - M]^+) \\ & \leq K\|\nabla v_h\|_{L^2(\Omega)} \|[v_h - M]^+\|_{H^1(\Omega)} \\ & \leq K\sqrt{C_s(2)^2 + 1}\|\nabla v_h\|_{L^2(\Omega)} \|\nabla[v_h - M]^+\|_{L^2(\Omega)}. \end{aligned}$$

□

Lemma 3. *Let (A1)–(A3), (A5)–(A7) hold, let $a_1 = \cdots = a_n = 0$ and let $v_h \in X_h$, $M \in \mathbb{R}$. Then*

$$a([v_h - M]^-, [v_h - M]^-) \leq a(v_h, v_h).$$

Proof. Via (A7),

$$a([v_h - M]^+, [v_h - M]^-) = \sum_{i \neq j} (v_h(x_i) - M)^+ (v_h(x_j) - M)^- a(\phi_i, \phi_j) \geq 0.$$

Since $(a(\phi_i, \phi_j)_{ij}$ is symmetric and positive semi-definite, we obtain

$$\begin{aligned} a(v_h, v_h) &= a([v_h - M]^-, [v_h - M]^-) + 2a([v_h - M]^+, [v_h - M]^-) \\ &\quad + a([v_h - M]^+, [v_h - M]^+) \\ &\geq a([v_h - M]^-, [v_h - M]^-) + a([v_h - M]^+, [v_h - M]^+) \\ &\geq a([v_h - M]^-, [v_h - M]^-). \end{aligned}$$

□

Furthermore, we need an estimate for $[v]^+ - (v)^+$ for $v \in X_h$. Since $[v]^+$ is the linear interpolation of $(v)^+$, we can use the interpolation results of [6]:

Lemma 4. *Let $1 \leq s \leq 2d/(d-2)$ ($s < \infty$ if $d \leq 2$) and let $v \in X_h$. Then*

$$\|[v]^+ - (v)^+\|_{L^s(\Omega)} \leq C_I h^{1+d/s-d/2} \|\nabla([v]^+ - (v)^+)\|_{L^2(\Omega)},$$

where $C_I > 0$ is a constant depending on s and d.

Actually the proof in [6] needs the assumption $H^1(\Omega) \hookrightarrow C^0(\overline{\Omega})$ which holds only for $d = 1$. However, one easily verifies along the argumentation in [6] that the estimate of Lemma 4 also holds for $H^1(\Omega) \cap C^0(\overline{\Omega})$, in particular for X_h.

The proof of the following technical lemma can be found in [6, p. 150]:

Lemma 5. *Let $r \in [1, \infty)$. Then there is a constant $\kappa_r > 0$ such that for each d-simplex τ in $\mathbb{R}^d$ (with vertices $x_1^{(\tau)}, \ldots, x_{d+1}^{(\tau)}$) and for each affine, non-negative function $v : \tau \to \mathbb{R}$,*

$$\|v\|_{L^r(\tau)}^r \geq \kappa_r \text{meas}(\tau) \sum_{i=1}^{d+1} v(x_i^{(\tau)})^r.$$

Remark 6. The constant κ_r can explicitly be calculated by transforming τ to a (reference) d-simplex σ in $\mathbb{R}^d$ with vertices $x_1^{(\sigma)}, \ldots, x_{d+1}^{(\sigma)}$. If $v : \sigma \to \mathbb{R}$ is affine and non-negative, then $v(x) = \sum_{i=1}^{d+1} v(x_i^{(\sigma)}) \phi_i^{(\sigma)}(x)$, where the barycentric coordinate functions $\phi_i^{(\sigma)} : \sigma \to \mathbb{R}$, $i = 1, \ldots, d+1$, are affine, non-negative with $\phi_i^{(\sigma)}(x_j^{(\sigma)}) = \delta_{ij}$, $i, j = 1, \ldots, d+1$, and it holds [6, p. 151]

$$\kappa_r = \text{meas}\,(\sigma)^{-1} \min_{i=1,\ldots,d+1} \|\phi_i^{(\sigma)})\|_{L^r(\sigma)}^r.$$

If $d = 2$, then we can choose σ to be the triangle with vertices $(0, 0)$, $(0, 1)$, $(1, 0)$ and we deduce $\kappa_r = 2/(r+1)(r+2)$.

3 A discrete minimum principle

We recall the minimum principle for the continuous case using the Stampacchia truncation technique. Then we prove the discrete minimum principle.

3.1 The continuous case

We assume:

(B1) $\mathrm{meas}_{d-1}(\Gamma_D) > 0$.
(B2) The operator L of (2) is uniformly elliptic, i.e., there exists $k > 0$ such that $a(u,u) \geq k\|\nabla u\|_{L^2(\Omega)}$ for all $u \in H_0^1(\Omega \cup \Gamma_N)$.
(B3) There exist $m_0 \in \mathbb{R}$, $p > \max\{1, d/2\}$ and $\underline{g} \in L^p(\Omega)$ such that

$$g(x,u) \geq \underline{g}(x) \quad \text{for } x \in \Omega \text{ and } u \leq m_0.$$

Proposition 7. *Let q, $r > 1$ be such that $1/p + 1/q + 1/r = 1$. If $d \leq 2$ we choose $r > q$, otherwise $r = 2d/(d-2)$. Furthermore, let $u \in H_0^1(\Omega \cup \Gamma_N) + u_D$ be a weak solution of* (8)*. Then*

$$\inf_{\Omega} u \geq \min\{m_0, \inf_{\Gamma_D} u_D\} - C_0\|\underline{g}\|_{L^p(\Omega)}, \tag{11}$$

where

$$C_0 = 2^{r/(r-q)}k^{-1}C_s(r)^2\mathrm{meas}\,(\Omega)^{(r-q)/rq} \tag{12}$$

and $C_s(r)$ is the Poincaré-Sobolev constant in (6)*.*

In particular, if $m_0 > 0$, if $\inf_{\Gamma_D} u_D > 0$ and if $\|\underline{g}\|_{L^p(\Omega)}$ is small enough then $u \geq c > 0$ in Ω for some constant c.

The proof is a variant of Stampacchia's maximum principle [22] and relies on the following lemma which is proved, for instance, in [23, p. 105].

Lemma 8. *Let $H : [\alpha, \beta) \to [0, \infty)$ be a non-increasing function with $\alpha < \beta \leq \infty$. Suppose there are positive constants κ, r, γ with $\gamma > 1$ and*

$$H(\mu) \leq \frac{\kappa^r}{(\mu - m)^r}H(m)^\gamma \quad \text{for } \alpha < m < \mu < \beta.$$

If $M^* = 2^{\gamma/(\gamma-1)}\kappa H(\alpha)^{(\gamma-1)/r}$ is such that $\alpha + M^* < \beta$, then

$$H(\alpha + M^*) = 0.$$

Proof of Proposition 7. We observe $a(v^-, v^+) = 0$ and therefore $a(v, v^+) = a(v^+, v^+)$ for any $v \in H_0^1(\Omega \cup \Gamma_N)$. Furthermore, the evaluation of $a(u, v)$ for $u, v \in H_0^1(\Omega \cup \Gamma_N)$ only involves the first argument's derivatives $\partial_1 u, \ldots, \partial_d u$. Thus, $a(-u + m, v) = a(-u, v) = -a(u, v)$ for any $m \in \mathbb{R}$ and any $u, v \in H_0^1(\Omega \cup \Gamma_N)$.

Let $m < \min\{m_0, \inf_{\Gamma_D} u_D\}$. Then $(-u + m)^+ \in H_0^1(\Omega \cup \Gamma_N)$ can be used as a test function in (8) and we obtain, by (B2), by the previously mentioned properties of a and by (B3),

$$\begin{aligned} k\|\nabla(-u+m)^+\|^2_{L^2(\Omega)} &\le a((-u+m)^+, (-u+m)^+) \\ &= -a(u, (-u+m)^+) \\ &= -\int_\Omega g(x,u)\,(-u+m)^+ dx \\ &\le -\int_\Omega \underline{g}(x)\,(-u+m)^+ dx \\ &\le \|\underline{g}\|_{L^p(\Omega)} \|(-u+m)^+\|_{L^r(\Omega)} (\text{meas}\,(u < m))^{1/q} \\ &\le \|\underline{g}\|_{L^p(\Omega)} C_s(r) \| \\ &\qquad \times \nabla(-u+m)^+\|_{L^2(\Omega)} (\text{meas}\,(u < m))^{1/q}, \end{aligned} \tag{13}$$

where p, q, r are specified above, and in the last inequality we used the Poincaré-Sobolev inequality (6). This estimate and the elementary inequality

$$\|(-u+m)^+\|_{L^r(\Omega)} \ge (m - \nu)(\text{meas}\,(u < \nu))^{1/r} \qquad \forall \nu < m$$

together imply via the Poincaré-Sobolev inequality (6)

$$\begin{aligned} (m-\nu)(\text{meas}\,(u<\nu))^{1/r} &\le C_s(r)\|\nabla(-u+m)^+\|_{L^2(\Omega)} \\ &\le C_s(r)^2 k^{-1} \|\underline{g}\|_{L^p(\Omega)} (\text{meas}\,(u < m))^{1/q} \end{aligned}$$

and for all $\nu < m$

$$\text{meas}\,(u < \nu) \le \frac{C_s(r)^{2r} k^{-r} \|\underline{g}\|^r_{L^p(\Omega)}}{(m-\nu)^r} (\text{meas}\,(u < m))^{r/q}.$$

The assumptions on q and r imply $\gamma := r/q > 1$, because due to assumption $p > \max\{1, d/2\}$. We set $\alpha = -\min\{m_0, \inf_{\Gamma_D} u_D\}$, $\beta = \infty$, and $H(y) = \text{meas}\,(u < -y)$ for $y \in [\alpha, \beta)$. Hence, we can apply Lemma 8 with $\kappa = C_s(r)^2\, k^{-1}\, \|\underline{g}\|_{L^p(\Omega)}$ to deduce

$$H\left(\alpha + 2^{\gamma/(\gamma-1)} \kappa H(\alpha)^{(\gamma-1)/r}\right) = 0.$$

In view of the estimate $H(\alpha) \le \text{meas}\,(\Omega)$ we conclude

$$u \ge \min\{m_0, \inf_{\Gamma_D} u_D\} - C_0 \|\underline{g}\|_{L^p(\Omega)} \quad \text{in } \Omega,$$

where C_0 is as above. □

3.2 The discrete case

We use the projected function $[-u_h + m]^+$ defined in Section 2 as a test function. Replacing in Stampacchia's argument the term $(-u_h + m)^+$ by $[-u_h + m]^+$ yields extra terms which can be estimated under additional assumptions on the non-linearity $g(x, u)$. For this, let $u_h \in X_h + u_{D,h}$ be a solution of (9) such that

$$\sup_{\Omega} u_h \leq M_0, \quad \|\nabla u_h\|_{L^2(\Omega)} \leq D_0. \tag{14}$$

Estimates for M_0 and D_0 are given in Section 3.3. We assume:

(B4) There exists $g_* \in L^p(\Omega)$ with $p > \max\{1, d/2\}$ such that

$$g(x, u) \geq g_*(x) \quad \text{for } x \in \Omega,\ m_0 \leq u \leq M_0,$$

where m_0 is as in (B3) and M_0 is defined in (14).

(B5) There exists $\overline{g} \in L^p(\Omega)$ with $p > \max\{1, d/2\}$ such that

$$g(x, u) \leq \overline{g}(x) \quad \text{for } x \in \Omega,\ u \in \mathbb{R}.$$

Theorem 9. *Let (A1)–(A8), (B1)–(B4) and* (14) *hold. Moreover, assume* $d \leq 5$ *and* $p > \max\{1, 2d/(6-d)\}$. *Then*

$$\min_{\Omega} u_h \geq \min\{m_0, \inf_{\Gamma_D} u_{D,h}\} - C_1\|\underline{g}\|^{1/2}_{L^p(\Omega)} - C_2 h^{\alpha}, \tag{15}$$

where the positive constants α, C_1 *and* C_2 *are defined as follows:*

$$\alpha = \frac{1}{2} + \frac{d}{2s} - \frac{d}{4},$$
$$C_{1,2} = 2^{r/(r-2q)}(\text{meas}\,(\Omega))^{(r-2q)/2rq} C_s(r)\kappa_r^{-1/r}(D_0/k)^{1/2} C^*_{1,2},$$

where

$$C_1^* = \sqrt{C_s(s)}, \quad C_2^* = \sqrt{C_I(K_0+1)(\|\underline{g}\|_{L^p(\Omega)} + \|g_*\|_{L^p(\Omega)})},$$

and $r > 2q$, $s > p/(p-1)$ *if* $d \leq 2$, $r = 2d/(d-2)$, $s \in (dp/(2p-d), 2d/(d-2)]$ *if* $3 \leq d \leq 5$, *and* $1/q = 1 - 1/p - 1/s \in (0, 1)$. *The constant* K_0 *is defined in Lemma 1.*

Remark 10. (1) For *monotone* non-linearities, Kerkhoven et al. [13] proved a similar discrete minimum principle with $\alpha = 2$. In our case the exponent α in Theorem 9 is always smaller than one.

(2) If we set $h = 0$ the bounds of Proposition 7 will not be recovered. This is not surprising since in the proof of Proposition 7 we can divide by $\|\nabla(-u + m)^+\|_{L^2(\Omega)}$. This is not possible in the proof of Theorem 9.

(3) The discrete solution u_h corresponding to the problem

$$-\Delta u = -u(\log u + f(x)) \text{ in } \Omega, \quad u = u_D > 0 \text{ on } \Gamma_D, \quad u_\nu = 0 \text{ on } \Gamma_N,$$

with $f_0 \leq f(x) \leq f_1$ for $x \in \Omega$ and for some $f_0, f_1 \in \mathbb{R}$ satisfies, by Theorem 9,

$$\min_\Omega u_h \geq \min\{\exp(-f_1), \inf_{\Gamma_D} u_{D,h}\} - C_2 h^\alpha > 0,$$

if $u_{D,h} > 0$ and if $h > 0$ is small enough.

Proof of Theorem 9. Let $m < m_1 = \min\{m_0, \inf_{\Gamma_D} u_{D,h}\}$ and use $[-u_h + m]^+ \in V_h$ as test function in (9) to obtain

$$a(u_h, [-u_h + m]^+) = F[u_h]([-u_h + m]^+). \tag{16}$$

The inequality (10) allows to estimate the left-hand side:

$$\begin{aligned} a(-u_h, [-u_h - (-m)]^+) &= -a(u_h, [-u_h + m]^+) \\ &\geq k \|\nabla[-u_h + m]^+\|^2_{L^2(\Omega)}. \end{aligned} \tag{17}$$

Concerning the right-hand side of (16) we introduce the set $E(m) = \{[-u_h + m]^+ > 0\}$ and employ (B3)–(B4) and the elementary inequality $[-u_h + m]^+ \geq (-u_h + m)^+$, yielding

$$\begin{aligned} &-F[u_h]([-u_h + m]^+) \\ &= -\int_{E(m)} g(x, u_h)(-u_h + m)^+ dx \\ &\quad - \int_{E(m)\cap\{u_h \leq m_0\}} g(x, u_h) \left([-u_h + m]^+ - (-u_h + m)^+\right) dx \\ &\quad - \int_{E(m)\cap\{u_h > m_0\}} g(x, u_h) \left([-u_h + m]^+ - (-u_h + m)^+\right) dx \\ &\leq -\int_{E(m)} \underline{g}(x)(-u_h + m)^+ dx \\ &\quad - \int_{E(m)\cap\{u_h \leq m_0\}} \underline{g}(x) \left([-u_h + m]^+ - (-u_h + m)^+\right) dx \\ &\quad - \int_{E(m)\cap\{u_h > m_0\}} g_*(x) \left([-u_h + m]^+ - (-u_h + m)^+\right) dx. \end{aligned}$$

The choice of the parameters p, s and q allows the use of the Hölder inequality:

$$\begin{aligned}
-F[u_h]([-u_h+m]^+) &\le \|\underline{g}\|_{L^p(\Omega)}\|(-u_h+m)^+\|_{L^s(\Omega)}(\text{meas}\,E(m))^{1/q}\\
&\quad+\Big(\|\underline{g}\|_{L^p(\Omega)}+\|g_*\|_{L^p(\Omega)}\Big)\,\|[-u_h+m]^+-(-u_h+m)^+\|_{L^s(\Omega)}\\
&\quad\times(\text{meas}\,E(m))^{1/q}\\
&\le \|\underline{g}\|_{L^p(\Omega)}C_s(s)\|\nabla(-u_h+m)^+\|_{L^2(\Omega)}(\text{meas}\,E(m))^{1/q}\\
&\quad+\Big(\|\underline{g}\|_{L^p(\Omega)}+\|g_*\|_{L^p(\Omega)}\Big)\,C_I h^{1+d/s-d/2}\\
&\quad\times\|\nabla([-u_h+m]^+-(-u_h+m)^+)\|_{L^2(\Omega)}(\text{meas}\,E(m))^{1/q}.
\end{aligned}$$

In the last inequality we used Lemma 4 which is possible since $s\le 2d/(d-2)$ (if $d\ge 3$) and $s<\infty$ if $d\le 2$. By Lemma 1,

$$\begin{aligned}
&\|\nabla([-u_h+m]^+-(-u_h+m)^+)\|_{L^2(\Omega)}\\
&\quad\le\|\nabla[-u_h+m]^+\|_{L^2(\Omega)}+\|\nabla(-u_h+m)^+\|_{L^2(\Omega)}\\
&\quad\le(K_0+1)\|\nabla u_h\|_{L^2(\Omega)},
\end{aligned}$$

and therefore, observing (14),

$$\begin{aligned}
-F[u_h]\left([-u_h+m]^+\right)\le D_0(\text{meas}\,E(m))^{1/q}\big[C_s(s)\|\underline{g}\|_{L^p(\Omega)}\\
+C_I(K_0+1)h^{1+d/s-d/2}(\|\underline{g}\|_{L^p(\Omega)}+\|g_*\|_{L^p(\Omega)})\big].
\end{aligned}$$

Putting together (17) and the above estimate we deduce from the Poincaré-Sobolev inequality (6) for $r=2d/(d-2)$ (if $d\ge 3$) or $2q<r<\infty$ (if $d\le 2$):

$$\begin{aligned}
\|[-u_h+m]^+\|^2_{L^r(\Omega)}\le C_s(r)^2(D_0/k)(\text{meas}\,E(m))^{1/q}\big[C_s(s)\|\underline{g}\|_{L^p(\Omega)}\\
+C_I(K_0+1)h^{1+d/s-d/2}(\|\underline{g}\|_{L^p(\Omega)}+\|g_*\|_{L^p(\Omega)})\big].
\end{aligned}$$

We estimate the left-hand side from below. For this, let $T_h(\nu)=\{\tau\in T_h:\tau$ has a vertex $x_j^{(\tau)}$ such that $\nu>u_h(x_j^{(\tau)}\}$. Then, by Lemma 5, for any $\nu<m$,

$$\begin{aligned}
\|[-u_h+m]^+\|^r_{L^r(\Omega)} &= \sum_{\tau\in T_h} \|[-u_h+m]^+\|^r_{L^r(\tau)}\\
&\geq \kappa_r \sum_{\tau\in T_h} \mathrm{meas}(\tau) \sum_{j=1}^{d+1} \left((-u_h(x_j^{(\tau)})+m)^+\right)^r\\
&= \kappa_r \sum_{\tau\in T_h} \mathrm{meas}(\tau) \sum_{j,\ m>u_h(x_j^{(\tau)})} (-u_h(x_j^{(\tau)})+m)^r\\
&\geq \kappa_r \sum_{\tau\in T_h(\nu)} \mathrm{meas}(\tau) \sum_{j,\ \nu>u_h(x_j^{(\tau)})} (-u_h(x_j^{(\tau)})+m)^r\\
&\geq \kappa_r \sum_{\tau\in T_h(\nu)} \mathrm{meas}(\tau) \sum_{j,\ \nu>u_h(x_j^{(\tau)})} (-\nu+m)^r\\
&\geq \kappa_r \sum_{\tau\in T_h(\nu)} \mathrm{meas}(\tau)(m-\nu)^r = \kappa_r(m-\nu)^r \mathrm{meas}(E(\nu)).
\end{aligned}$$

Hence, if we set

$$\begin{aligned}
\bar{K}(s,r) = C_s(r)\kappa_r^{-1/r}(D_0/k)^{1/2}\Big[&\sqrt{C_s(s)\|\underline{g}\|_{L^p(\Omega)}}\\
&+ h^{1/2+d/2s-d/4}\sqrt{C_I(K_0+1)(\|\underline{g}\|_{L^p(\Omega)}+\|g_*\|_{L^p(\Omega)})}\Big],
\end{aligned}$$

then for all $\nu < m$,

$$\mathrm{meas}\,(E(\nu)) \leq \frac{\bar{K}(s,r)^r}{(m-\nu)^r}(\mathrm{meas}\,(E(m))^{r/2q}.$$

We introduce the function $H : [\alpha,\infty) \to [0,\infty)$, where $\alpha = -m_1$, $\beta = \infty$, by $H(y) = \mathrm{meas}(E(-y))$. Then H is non-increasing and

$$H(y) \leq \frac{\bar{K}(r,s)^r}{(y-z)^r} H(z)^{r/2q} \quad \text{for } \alpha < z < y.$$

We claim that $r/2q > 1$. Indeed, if $d \geq 3$ then

$$\frac{r}{2q} = \frac{d}{d-2}\left(1-\frac{1}{p}-\frac{1}{s}\right) > \frac{d}{d-2}\left(1-\frac{1}{p}+\frac{1}{p}-\frac{2}{d}\right) = 1,$$

and if $d \leq 2$, then $r/2q > 1$ by assumption. We deduce from Stampacchia's Lemma 8 that $H(\alpha+M^*) = 0$, where $M^* = 2^{r/(r-2q)}H(\alpha)^{(r-2q)/2rq}\bar{K}(r,s)$, such that in view of $H(\alpha) \leq \mathrm{meas}\,(\Omega)$ the estimate

$$u_h \geq m_1 - 2^{r/(r-2q)}(\mathrm{meas}\,(\Omega))^{(r-2q)/2rq}\bar{K}(r,s)$$

follows. □

3.3 Estimates for u_h and ∇u_h

Theorem 9 involves upper estimates on u_h and $\|\nabla u_h\|_{L^2(\Omega)}$. In this section we give estimates for these quantities independent of h.

Proposition 11. *Let (A1)–(A8) and (B1)–(B2), (B5) hold and let u_h be a weak solution of* (9). *Then*

$$\sup_{\Omega} u_h \leq \sup_{\Gamma_D} u_{D,h} + C_3\|\overline{g}\|_{L^p(\Omega)},$$

where

$$C_3 = 2^{r/(r-q)}k^{-1}C_s(r)^2\kappa_r^{-1/r}(\text{meas }(\Omega))^{(r-q)/rq},$$

$C_s(r)$ is the Poincaré-Sobolev constant in (6)*, and $1 < q < r$ are defined by $1/p + 1/q + 1/r = 1$ and $r < \infty$ (if $d \leq 2$), $r = 2d/(d-2)$ (if $d > 2$).*

Proof. Let $M \geq \sup_{\Gamma_D} u_{D,h}$. Then $[u_h - M]^+$ is an admissible test function in (9):

$$a(u_h, [u_h - M]^+) = F[u_h]([u_h - M]^+).$$

The estimate (10) again yields

$$a(u_h, [u_h - M]^+) \geq k\|\nabla[u_h - M]^+\|^2_{L^2(\Omega)}. \tag{18}$$

For the estimate of $F[u_h]([u_h - M]^+)$ we introduce the set $E(M) = \{[u_h - M]^+ > 0\}$. Then, using (B5),

$$\begin{aligned} F[u_h]([u_h - M]^+) &= \int_{E(M)} g(x, u_h)[u_h - M]^+dx \\ &\leq \int_{E(M)} \overline{g}(x)[u_h - M]^+dx \\ &\leq \|\overline{g}\|_{L^p(\Omega)}\|[u_h - M]^+\|_{L^r(\Omega)}(\text{meas } E(M))^{1/q}, \end{aligned}$$

where p, q, r are specified above. We infer from the Poincaré-Sobolev embedding (6):

$$F[u_h]([u_h - M]^+) \leq C_s(r)\|\overline{g}\|_{L^p(\Omega)}\|\nabla[u_h - M]^+\|_{L^2(\Omega)}(\text{meas } E(M))^{1/q}.$$

Putting together the above estimates, we obtain

$$\|\nabla[u_h - M]^+\|_{L^2(\Omega)} \leq k^{-1}C_s(r)\|\overline{g}\|_{L^p(\Omega)}(\text{meas } E(M))^{1/q},$$

and, again with the Poincaré-Sobolev inequality,

$$\|[u_h - M]^+\|_{L^r(\Omega)} \leq k^{-1}C_s(r)^2\|\overline{g}\|_{L^p(\Omega)}(\text{meas } E(M))^{1/q}.$$

Proceeding as in the proof of Theorem 9 we deduce

$$\|[u_h - M]^+\|^r_{L^r(\Omega)} \geq \kappa_r(\mu - M)^r\text{meas }(E(\mu))$$

for all $\mu > M$. Therefore, setting $\alpha = \sup_{\Gamma_D} u_{D,h}$ and $\beta = \infty$, the function $H : [\alpha, \beta) \to [0, \infty)$, $H(\mu) = \text{meas}\,(E(\mu))$, is non-increasing and we infer for all μ, M with $\alpha < M < \mu < \beta$ the inequality

$$H(\mu) \leq \frac{C_s(r)^{2r} \|\overline{g}\|^r_{L^p(\Omega)}}{\kappa_r k^r (\mu - M)^r} H(M)^{r/q}.$$

By assumption, it holds $r/q > 1$. Hence we can apply Lemma 8 with $\kappa = C_s(r)^2 \|\overline{g}\|_{L^p(\Omega)} / \kappa_r^{1/r} k$ to deduce, taking into account $H(\alpha) \leq \text{meas}\,(\Omega)$,

$$u_h \leq \sup_{\Gamma_D} u_{D,h} + C_3 \|\bar{g}\|_{L^p(\Omega)} \quad \text{in } \Omega.$$

□

An estimate of $\|\nabla u_h\|_{L^2(\Omega)}$ clearly depends on the precise structure of the non-linearity $g(x, u)$. However, essentially under the assumptions (B3) and (B5), we can prove the following result.

Proposition 12. *Let (A1)–(A8) and (B1)–(B3), (B5) hold. Furthermore, we assume that* $G_1 = \max_{m \leq u \leq M} |g(\cdot, u)| \in L^p(\Omega)$*, where* $m = \min\{m_0, \inf_{\Gamma_D} u_{D,h}\}$ *and* $M = \sup_{\Gamma_D} u_{D,h}$*. Let* u_h *be a weak solution of* (9)*. Then*

$$\|\nabla u_h\|_{L^2(\Omega)} \leq (K C_s(2)/k + 1) \|\nabla u_{D,h}\|_{L^2(\Omega)} + C_4,$$

where k is the coercitivity constant defined in (B2), K is defined in (7)*,*

$$C_4 = C_s(p/(p-1)) (\|\underline{g}\|_{L^p(\Omega)} + \|\overline{g}\|_{L^p(\Omega)} + \|G_1\|_{L^p(\Omega)})/k,$$

and $C_s(p/(p-1))$ *is the Poincaré-Sobolev constant defined in* (6)*.*

Proof. With the test function $u_h - u_{D,h}$ we estimate

$$\begin{aligned}
& k \|\nabla(u_h - u_{D,h})\|^2_{L^2(\Omega)} \leq a(u_h - u_{D,h}, u_h - u_{D,h}) \\
&= \int_\Omega g(x, u_h)(u_h - u_{D,h}) dx - a(u_{D,h}, u_h - u_{D,h}) \\
&\leq \int_{\{u_h \leq m\}} \underline{g}(x)(u_h - u_{D,h}) dx + \int_{\{m < u_h < M\}} G_1(x) |u_h - u_{D,h}| dx \\
&\quad + \int_{\{u_h \geq M\}} \overline{g}(x)(u_h - u_{D,h}) dx \\
&\quad + K C_s(2) \|\nabla u_{D,h}\|_{L^2(\Omega)} \|\nabla(u_h - u_{D,h})\|_{L^2(\Omega)} \\
&\leq \left(\|\underline{g}\|_{L^p(\Omega)} + \|\overline{g}\|_{L^p(\Omega)} + \|G_1\|_{L^p(\Omega)} \right) \|u_h - u_{D,h}\|_{L^q(\Omega)} \\
&\quad + K C_s(2) \|\nabla u_{D,h}\|_{L^2(\Omega)} \|\nabla(u_h - u_{D,h})\|_{L^2(\Omega)},
\end{aligned}$$

where $q = p/(p-1)$. Since $p > \max\{1, d/2\} \geq 2d/(d+2)$ we have for $d \geq 3$, $q \leq 2d/(d-2)$. Thus, with the Poincaré-Sobolev inequality (6),

$$\begin{aligned}\|\nabla(u_h - u_{D,h})\|_{L^2(\Omega)} \leq k^{-1}C_s(q)(\|\underline{g}\|_{L^p(\Omega)} + \|\overline{g}\|_{L^p(\Omega)} + \|G_1\|_{L^p(\Omega)})\\ + k^{-1}KC_s(2)\|\nabla u_{D,h}\|_{L^2(\Omega)},\end{aligned}$$

from which the assertion follows. □

4 A discrete maximum principle

In this section maximum principles for equilibrium solutions are considered. In the first subsection equilibrium solutions are introduced. A corresponding maximum principle is formulated. The second subsection deals with a discrete version of this principle.

4.1 The continuous case

In the sequel let

$$G : \Omega \times \mathbb{R} \to \mathbb{R}, \quad G(x, s) = \int_0^s g(x, \sigma) d\sigma.$$

We assume

(C1) For all $i = 1, \ldots, d$: $a_i = 0$. We write $a_0(u, v)$ instead of $a(u, v)$.
(C2) For all $s \in \mathbb{R}$: $G(\cdot, s) \in L^1(\Omega)$.
(C3) There is a number $M_0 \in \mathbb{R}$ such that for all $x \in \Omega$ the function $G(x, \cdot)$ strictly decreases on (M_0, ∞).

Differently from the assumptions of the previous sections, the pure Neumann boundary case $\Omega = \Gamma_N$ and the case of degenerate diffusion matrices are included. Assumption (C3) is satisfied if, for instance, there is $M_0 \in \mathbb{R}$ such that for $x \in \Omega$ and $s > M_0$ it holds $g(x, s) < 0$.

We introduce the functional

$$E : \mathcal{C} \to \mathbb{R}, \quad E(v) = \frac{1}{2}a_0(v, v) - \int_\Omega G(x, v) dx, \tag{19}$$

where

$$\mathcal{C} = \{v \in u_D + H_0^1(\Omega \cup \Gamma_N) : G(\cdot, v) \in L^1(\Omega)\}.$$

Definition 13. *Let (A1)–(A8) and (C1)–(C3) hold. Then the function $u \in \mathcal{C}$ is an* equilibrium solution *iff u minimizes the functional E:*

$$E(u) = \inf_{v \in \mathcal{C}} E(v).$$

Here we do not discuss the existence or uniqueness of equilibrium solutions and whether or not equilibrium solutions are weak solutions of (8). Instead we are interested in the following maximum principle.

Proposition 14. *Let (A1)–(A8) and (C1)–(C3) hold and let u be an equilibrium solution of* (8). *Then*

$$\sup_{\Omega} u \leq \max\{\sup_{\Gamma_D} u_D, M_0\},$$

with the convention $\sup_{\Gamma_D} u_D = -\infty$ *whenever* Γ_D *has zero measure.*

Proof. Indirect. We assume $\sup_{\Omega} u > K := \max\{\sup_{\Gamma_D} u_D, M_0\}$. Then there exists $\varepsilon > 0$ such that $\Omega_\varepsilon := \{u > K + \varepsilon\}$ has non-zero measure. We introduce $u_\varepsilon(x) := u(x) - (u(x) - (K + \varepsilon))^+ = \min\{u(x), K + \varepsilon\}$ for $x \in \Omega$. Then $u_\varepsilon = u$ on $\Omega \backslash \Omega_\varepsilon$ and $u_\varepsilon = K + \varepsilon < u$ on Ω_ε.

We claim that $u_\varepsilon \in \mathcal{C}$. Clearly, $u_\varepsilon \in u_D + H_0^1(\Omega \cup \Gamma_D)$. Since $u \in \mathcal{C}$, $G(\cdot, u) \in L^1(\Omega)$. Moreover, by assumption (C2), $G(\cdot, K + \varepsilon) \in L^1(\Omega)$. Therefore $-G(\cdot, u_\varepsilon) = \min\{-G(\cdot, u), -G(\cdot, K+\varepsilon)\} \in L^1(\Omega)$. Thus $u_\varepsilon \in \mathcal{C}$ and $E(u) \leq E(u_\varepsilon)$.

Now we calculate

$$\begin{aligned} E(u_\varepsilon) - E(u) &= \frac{1}{2} \sum_{i,j=1}^{d} \int_{\Omega_\varepsilon} a_{ij}((\partial_i u_\varepsilon)(\partial_j u_\varepsilon) - (\partial_i u)(\partial_j u))dx \\ &\quad - \int_{\Omega_\varepsilon} (G(x, u_\varepsilon) - G(x, u))dx \\ &= -\frac{1}{2} \sum_{i,j=1}^{d} \int_{\Omega_\varepsilon} a_{ij}(\partial_i u)(\partial_j u)\, dx \\ &\quad - \int_{\Omega_\varepsilon} (G(x, K + \varepsilon) - G(x, u))dx \\ &< 0, \end{aligned}$$

since $(a_{ij}(x))$ is positive semi-definite and $G(x, \cdot)$ is strictly decreasing on $(K + \varepsilon, \infty)$. □

We illustrate Proposition 14 by two examples.

Example 15. Consider the equation (3) with (for the sake of simplicity) homogeneous Dirichlet boundary conditions:

$$-\delta^2 \Delta u = -u(\log u + f(x)) \text{ in } \Omega, \quad u = 0 \text{ on } \Gamma_D, \quad u_\nu = 0 \text{ on } \Gamma_N.$$

It can be seen that this problem has an equilibrium solution u minimizing the function

$$E_1(v) = \frac{\delta^2}{2} \int_\Omega |\nabla v|^2 dx + \frac{1}{4} \int_\Omega (v^+)^2 (2 \log(v^+) - 1 + 2f(x))dx$$

in the set

$$\mathcal{C}_1 = \{v \in H_0^1(\Omega) : (v^+)^2(\log(v^+) - 1 + 2f \in L^1(\Omega)\} = H_0^1(\Omega).$$

Since $g(x, s) := -(s)^+(\log(s)^+ + f(x)) < 0$ for $x \in \Omega$ and $s > \exp(-\inf_\Omega f)$, we deduce from Proposition 14:

$$\sup_\Omega u \leq \exp(-\inf_\Omega f).$$

This result can be also obtained from standard Stampacchia estimates for weak solutions (using $(u - M)^+$ for appropriate $M \in \mathbb{R}$ as a test function in the weak formulation).

Example 16. A rather extreme case concerns the choice $a_{ij} = 0$ for all i, j. Proposition 14 also applies in this situation. The crucial point, however, is the *existence* of equilibrium solutions. Let us consider

$$0 = \exp(u) - \frac{1}{1 + x^2}, \quad u \in H_0^1(0, 1), \tag{20}$$

with corresponding energy functional

$$E_2(v) = \frac{1}{2}\int_0^1 \left(\exp(v(x)) - 1 - \frac{v(x)}{1 + x^2}\right) dx,$$

to be minimized in $\mathcal{C}_2 = H_0^1(0, 1)$. E_2 has no minimizer in $\mathcal{C}_2$, because each minimizer has to satisfy (20), i.e. $u(x) = -\log(1 + x^2)$, which does not belong to $H_0^1(0, 1)$. Hence Proposition 14 does not yield any information in this situation.

However, if we consider the problem

$$0 = \exp(u) - \frac{1}{1 + x^2}, \quad u + x\ \log 2 \in H_0^1(0, 1)$$

then E_2 remains to be the corresponding energy functional but now to be minimized in

$$\mathcal{C}_2' = u_D + H_0^1(0, 1), \quad u_D(x) = -x \log 2.$$

In this case, $u(x) = -\log\left(1 + x^2\right)$ is the unique minimizer of E_2 in $\mathcal{C}_2'$ and we obtain $u \leq 0$ from Proposition 14.

4.2 The discrete case

In this subsection we are concerned with a discrete version of the maximum principle for equilibrium solutions. We assume in addition to (C1)–(C3) of the previous subsection:

(C4) There is a function $g_0 \in C^0(\mathbb{R})$ with primitive G_0 such that $g(x,s) \leq -g_0(s)$ for $x \in \Omega$ and $s \in \mathbb{R}$, $G_0(M_0) = \sup_{s<M_0} G_0(s)$, and G_0 is strictly increasing on (M_0, ∞), where M_0 is defined in (C3).

(C5) For each $A > 0$ there is a function $g_A \in L^1(\Omega)$ such that $|g(x,s)| \leq g_A(x)$ for $x \in \Omega$ and $-A \leq s \leq A$.

If $g_0(s)$ is positive for all $s > M_0$, then (C4) is satisfied. Assumption (C5) is needed in order to apply Lebesgue's dominated convergence theorem.

Now we introduce the discrete analogue of equilibrium solutions.

Definition 17. *Let (A1)–(A8) and (C1)–(C3) hold and let E be as in* (19). *We set*

$$\mathcal{C}_h := \{v_h \in u_{D,h} + V_h : G(x, v_h) \in L^1(\Omega)\}.$$

*Then u_h is an X_h-*equilibrium solution *of* (1) *iff*

$$u_h \in \mathcal{C}_h \quad \textit{and} \quad E(u_h) = \inf_{v_h \in \mathcal{C}_h} E(v_h).$$

Under the condition (C5) each X_h-equilibrium solution is a finite-element solution:

Proposition 18. *Let (A1)–(A8) and (C1)–(C3), (C5) hold. Then each X_h-equilibrium solution u_h of* (1) *is a solution of* (9).

Proof. Since $u_h \in \mathcal{C}_h$, we have $u_h - u_{D,h} \in V_h$. It remains to be proved that u_h satisfies the equation in (9). Since u_h is a minimizer of E, $E(u_h + \varepsilon v_h) - E(u_h) \geq 0$ for all $v_h \in \mathcal{C}_h$ and $\varepsilon > 0$ and therefore

$$\begin{aligned} 0 &\leq \liminf_{\varepsilon \to 0} \frac{1}{\varepsilon}(E(u_h + \varepsilon v_h) - E(u_h)) \\ &= a_0(u_h, v_h) - \limsup_{\varepsilon \to 0} \frac{1}{\varepsilon} \int_\Omega (G(x, u_h + \varepsilon v_h) - G(x, u_h))dx. \end{aligned}$$

Since

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon}(G(x, u_h(x) + \varepsilon v_h(x)) - G(x, u_h(x))) = g(x, u_h(x))v_h(x)$$

and due to (C5),

$$\begin{aligned} \left|\frac{1}{\varepsilon}(G(x, u_h(x) + \varepsilon v_h(x)) - G(x, u_h(x)))\right| &= |g(x, u_h(x) + \varepsilon\theta(\varepsilon, x)v_h(x))| \\ &\leq g_A(x) \in L^1(\Omega), \end{aligned}$$

for some $\theta(x,\varepsilon) \in (0,1)$. Thus Lebesgue's dominated convergence Theorem applies and we deduce

$$0 \leq a_0(v_h, v_h) - \int_\Omega g(x, u_h(x))v_h(x)dx.$$

This inequality will also hold if we replace v_h by $-v_h$. Thus u_h solves (9).□

Theorem 19. *Let (A1)–(A8) and (C1)–(C4) hold and let u_h be a X_h-equilibrium solution of* (1). *Then*

$$\max_\Omega u_h \leq \max\{M_0, \max_{\Gamma_D} u_{D,h}\}.$$

Proof. Indirect. We assume $\max_\Omega u_h > K := \max\{M_0, \max_{\Gamma_D} u_{D,h}\}$. Similar to the proof of Proposition 14 we consider for fixed $\varepsilon \in (0, \max_\Omega u_h - K)$ the projected function

$$u_h^\varepsilon := u_h - [u_h - (K+\varepsilon)]^+.$$

We shall prove $E(u_h^\varepsilon) < E(u_h)$.

For this, we introduce

$$T_h^* = \{\tau \in T_h : \max_{x\in\tau} u_h(x) > K + \varepsilon\}, \quad \Omega_\varepsilon = \bigcup_{\tau \in T_h^*} \tau.$$

Clearly, $T_h^* \neq \emptyset$. Since $u_h^\varepsilon = u_h$ on $\Omega\backslash\Omega_\varepsilon$, we obtain

$$\begin{aligned} E(u_h^\varepsilon) - E(u_h) &= \frac{1}{2}\big(a_0(u_h^\varepsilon, u_h^\varepsilon) - a_0(u_h, u_h)\big) \\ &\quad - \sum_{\tau \in T_h^*} \int_\tau \big(G(x, u_h^\varepsilon) - G(x, u_h)\big)\, dx. \end{aligned}$$

From assumption (C1) and Lemma 3 we deduce via $a_0(K+\varepsilon, \cdot) = a_0(\cdot, K+\varepsilon) = 0$,

$$\begin{aligned} &a_0(u_h^\varepsilon, u_h^\varepsilon) \\ &= a_0(u_h - (K+\varepsilon) - [u_h - (K+\varepsilon]^+, u_h - (K+\varepsilon) - [u_h - (K+\varepsilon]^+) \\ &= a_0([u_h - (K+\varepsilon)]^-, [u_h - (K+\varepsilon)]^-) \\ &\leq a_0(u_h, u_h). \end{aligned}$$

Therefore

$$E(u_h^\varepsilon) - E(u_h) \leq - \sum_{\tau \in T_h^*} \int_\tau \big(G(x, u_h^\varepsilon) - G(x, u_h)\big)\, dx,$$

and we shall prove for all $\tau \in T_h^*$,

$$\int_\tau G(x, u_h^\varepsilon(x))dx > \int_\tau G(x, u_h(x))dx. \tag{21}$$

We write

$$u_h|_\tau = \sum_{j=1}^{d+1} m_j^{(\tau)} \phi_j^{(\tau)},$$

where $u_h|_\tau$ is the restriction of u_h to τ, $x_1^{(\tau)}, \dots, x_{d+1}^{(\tau)}$ are the vertices of τ, and $\phi_1^{(\tau)}, \dots, \phi_{d+1}^{(\tau)}$ are the finite-element basis elements restricted to τ (see Section 2). Then

$$u_h^\varepsilon|_\tau = \sum_{j=1}^{d+1} [m_j^{(\tau)}] \phi_j^{(\tau)},$$

where $[m_j^{(\tau)}] = \min\{m_j^{(\tau)}, K + \varepsilon\}$. We set

$$J^+ = \{j \in \{1, \dots, d+1\} : m_j^{(\tau)} > K + \varepsilon\}.$$

For $\tau \in T_h^*$, J^+ is non-empty. Let $x \in \tau$. We calculate, using (C4) and $u_h^\varepsilon(x) \leq u_h(x)$,

$$\begin{aligned} G(x, u_h^\varepsilon(x)) - G(x, u_h(x)) &= -\int_{u_h^\varepsilon(x)}^{u_h(x)} g(x, s)ds \geq \int_{u_h^\varepsilon(x)}^{u_h(x)} g_0(s)ds \\ &= G_0(u_h(x)) - G_0(u_h^\varepsilon(x)). \end{aligned}$$

It remains to prove $\int_\tau G_0(u_h^\varepsilon)dx < \int_\tau G_0(u_h)dx$. For this, we introduce the auxiliary function

$$H_0 : \mathbb{R}^{d+1} \to \mathbb{R}, \quad H_0(c_1, \dots, c_{d+1}) = \int_\tau G_0\Big(\sum_{j=1}^{d+1} c_j \phi_j^{(\tau)}\Big)dx.$$

Setting $[m] := ([m_1^{(\tau)}], \dots, [m_{d+1}^{(\tau)}])$ and $m := (m_1^{(\tau)}, \dots, m_{d+1}^{(\tau)})$, we have to show $H_0([m]) < H_0(m)$.

Since G_0 is strictly increasing on (M_0, ∞) (by assumption (C4)), the inequality $H_0([m]) < H_0(m)$ is immediate if $\min\{[m_1^{(\tau)}], \dots, [m_{d+1}^{(\tau)}]\} = \max\{[m_1^{(\tau)}], \dots, [m_{d+1}^{(\tau)}]\} = K + \varepsilon$, since in this case, $u_h^\varepsilon < u_h$ holds in the interior of τ. We assume therefore that $\min\{[m_1^{(\tau)}], \dots, [m_{d+1}^{(\tau)}]\} < \max\{[m_1^{(\tau)}], \dots, [m_{d+1}^{(\tau)}]\} = K + \varepsilon$. In particular, $J^+ \neq \{1, \dots, d+1\}$. The function H_0 is continuously differentiable with

$$\frac{\partial H_0}{\partial c_\alpha}(c_1, \dots, c_{d+1}) = \int_\tau g_0\Big(\sum_{j=1}^{d+1} c_j \phi_j^{(\tau)}\Big)\phi_\alpha^{(\tau)} dx.$$

It holds $[m_j^{(\tau)}] = m_j^{(\tau)}$ for all $j \notin J^+$, hence

$$\begin{aligned} H_0(m) - H_0([m]) &= \int_0^1 \nabla H_0(sm + (1-s)[m]) \cdot (m - [m]) ds \\ &= \sum_{j \in J^+} \int_0^1 \frac{\partial H_0}{\partial c_j}(sm + (1-s)[m])(m_j^{(\tau)} - [m_j^{(\tau)}]) ds. \qquad (22) \end{aligned}$$

Set $c_j = sm_j^{(\tau)} + (1-s)[m_j^{(\tau)}]$. We assume without loss of generality that the vertices are numbered in such a way that $c_1 = \max\{c_1, \ldots, c_{d+1}\}$ and $c_{d+1} = \min\{c_1, \ldots, c_{d+1}\}$. Then $c_1 > c_{d+1}$ since J^+ is neither empty nor the whole set $\{1, \ldots, d+1\}$. With this numbering, $1 \in J^+$ and

$$c_1 = sm_1^{(\tau)} + (1-s)[m_1^{(\tau)}] = sm_1^{(\tau)} + (1-s)(K + \varepsilon) > K + \varepsilon > M_0.$$

As G_0 is strictly increasing on (M_0, ∞), this implies

$$G_0(c_1) > G_0(M_0) = \sup_{\sigma < M_0} G_0(\sigma) \geq \frac{1}{c_1 - c_{d+1}} \int_{c_{d+1}}^{c_1} G_0(\sigma) d\sigma. \qquad (23)$$

Introducing barycentric coordinates (see [6]) we can reformulate

$$\begin{aligned} &H_0(c_1, \ldots, c_{d+1}) \\ &\quad = \text{meas}(\tau) \int_{(0,1)^d} G_0(c_1\lambda_1 + \cdots + c_d\lambda_d + c_{d+1} \\ &\qquad\qquad \times (1 - \lambda_1 - \cdots - \lambda_d)) d(\lambda_1, \ldots, \lambda_d) \\ &\quad = \frac{\text{meas}(\tau)}{c_1 - c_{d+1}} \int_{c_{d+1}}^{c_1} G_0(\sigma) d\sigma, \end{aligned}$$

where $\sigma = (\sigma_1, \ldots, \sigma_d)$, $\sigma_2 = \lambda_2, \ldots, \sigma_d = \lambda_d$ and

$$\sigma_1 = c_1\lambda_1 + \cdots + c_d\lambda_d + c_{d+1}(1 - \lambda_1 - \cdots - \lambda_d).$$

Thus, by (23),

$$\begin{aligned} \frac{\partial H_0}{\partial c_1}(c_1, \ldots, c_{d+1}) = &\frac{\text{meas}(\tau)}{c_1 - c_{d+1}} \\ &\times \left(G_0(c_1) - \frac{1}{c_1 - c_{d+1}} \int_{c_{d+1}}^{c_1} G_o(\sigma) d\sigma \right) > 0. \end{aligned}$$

Finally, we obtain from (22):

$$H_0(m) - H_0([m]) > 0.$$

□

4.3 Discussion

We re-consider the finite element discretizations of Examples 15 and 16. It is assumed that (A1)–(A8) hold. The verification of (C1)–(C3) is left to the reader.

Example 15 (revisited). The proof of existence of a X_h-equilibrium solution u_h is as straight forward as in the "continuous" case. Choosing

$$g_\circ(s) = s^+ \left(\log(s^+) + \inf_\Omega f\right),$$

we easily check (C4) with $G_\circ(s) = \frac{1}{4}(s^+)^2 \left(2\log(s^+) - 1 + 2 \inf_\Omega f\right)$, and $M_0 = \sqrt{e} \exp\left(-\inf_\Omega f\right)$. We obtain an estimate on $\max u_h$ which is a bit worse (more precisely, a factor $\sqrt{e}$ larger) than in the "continuous" case.

Example 16 (revisited). Certainly each constant function $u_c = c \in (-\infty, 10]$ is a X_h-equilibrium solution of (20). The verification of (C4) is easy for $g(x,s) = g_0(s) = (s-10)^+$. Hence $u_h \leq 10$ for each X_h-equilibrium solutions of (20).

Acknowledgements. The authors acknowledge partial support of the Project "Hyperbolic and Kinetic Equations" of the European Union, grant HPRN-CT-2002-00282. The first author was supported by the Deutsche Forschungsgemeinschaft, grants JU 359/3 (Gerhard-Hess Award) and JU 359/5 (Priority Program "Multiscale Problems"). The second author acknowledges support from the DFG FZT86.

References

[1] Ancona, M.: Diffusion-drift models of strong inversion layers. COMPEL **6**, 11–18 (1987)

[2] Ben Abdallah, N., Unterreiter, A.: On the stationary quantum drift-diffusion model. Z. Angew. Math. Phys. **49**, 251–275 (1998)

[3] Bertolazzi, E.: Discrete conservation and discrete maximum principle for elliptic PDEs. Math. Models Meth. Appl. Sci. **8**, 685–711 (1998)

[4] Burman, E., Ern, A.: Nonlinear diffusion and discrete maximum principle for stabilized Galerkin approximations of the convection-diffusion-reaction equation. Comput. Meth. Appl. Mech. Engin. **191**, 3833–3855 (2002)

[5] Christie, I., Hall, C.: The maximum principle for bilinear elements. Internat. J. Numer. Meth. Engin. **20**, 549–553 (1984)

[6] Ciarlet, P., Lions, J.L.: Handbook of Numerical Analysis. Finite Element Methods (Part 1). North-Holland, Amsterdam, 1991

[7] Ciarlet, P., Raviart, P.: Maximum principle and uniform convergence for the finite element method. Comput. Meth. Appl. Mech. Engin. **2**, 17–31 (1973)

[8] Cortey-Dumont, P.: On finite element approximation in the L^∞-norm of variational inequalities. Numer. Math. **47**, 45–57 (1985)

[9] Ishihara, K.: On finite element schemes of the Dirichlet problem for a system of nonlinear elliptic equations. Numer. Funct. Anal. Optim. **3**, 105–136 (1981)

[10] Ishihara, K.: Strong and weak discrete maximum principles for matrices associated with elliptic problems. Lin. Algebra Appl. **88/89**, 431–448 (1987)

[11] Jüngel, A.: Quasi-hydrodynamic Semiconductor Equations. Birkhäuser, Basel, 2001

[12] Jüngel, A., Pinnau, R.: A positivity-preserving numerical scheme for a fourth-order parabolic equation. SIAM J. Numer. Anal. **39**, 385–406 (2001)

[13] Kerkhoven, T., Jerome, J.: L_∞ stability of finite element approximations to elliptic gradient equations. Numer. Math. **57**, 561–575 (1990)

[14] Korotov, S., Křížik, M.: Acute type refinements of tetrahedral partitions of polyhedral domains. SIAM J. Numer. Anal. **39**, 724–733 (2001)

[15] Korotov, S., Křížik, M., Neittaanmäki, P.: Weakened acute type condition for tetrahedral triangulations and the discrete maximum principle. Math. Comput. **70**, 107–119 (2001)

[16] Kuo, H.-J., Trudinger, N.: A note on the discrete Aleksandrov-Bakelman maximum principle. Taiwanese J. Math. **4**, 55–64 (2000)

[17] Markowich, P., Ringhofer, C., Schmeiser, C.: Semiconductor Equations. Springer, Vienna, 1990

[18] Micheletti, S., Sacco, R., Simioni, P.: Numerical simulation of resonant tunneling diodes with a quantum drift-diffusion model. Submitted for publication, 2003

[19] Pinnau, R.: Uniform convergence of an exponentially fitted scheme for the quantum drift-diffusion model. Submitted for publication, 2003

[20] Pinnau, R., Unterreiter, A.: The stationary current-voltage characteristics of the quantum drift-diffusion model. SIAM J. Numer. Anal. **37**, 211–245 (1999)

[21] Schatz, A.: A weak discrete maximum principle and stability of the finite element method in L_∞ on plane polygonal domains I. Math. Comput. **34**, 77–91 (1980)

[22] Stampacchia, G.: Equations elliptiques du second ordre à coefficients discontinus. Les Presses de l'Université de Montréal, Canada, 1966

[23] Troianiello, G.: Elliptic Differential Equations and Obstacle Problems. Plenum Press, New York, 1987