Chapter 2
The linear elliptic model problem

Abstract This chapter introduces the stationary diffusion problem along with suitable concepts of solutions. Besides the interest in the corresponding simplest boundary value problem with characteristic multiscale features: How can we seriously think that we understand some numerical scheme in a complicated real-life application if we do not understand it even for this model problem?

2.1 Weak solutions and stability

We want to reformulate the problem

$$\begin{aligned} -\operatorname{div}(A\nabla u) &= f \quad \text{in } D, \\ u &= u_0 \quad \text{on } \Gamma_0, \\ (A\nabla u)\cdot n &= g \quad \text{on } \Gamma_1, \end{aligned} \tag{2.1}$$

so that it also admits solutions which do not belong to $C^1(\overline{D}) \cap C^2(D)$. Throughout this section $D \subset \mathbb{R}^d$ ($d \in \mathbb{N}$) is a bounded Lipschitz domain, Γ_0 is a closed subset of ∂D with positive surface measure, $\Gamma_1 = \partial D \setminus \Gamma_0$, $f \in L^2(D)$, $g \in L^2(\Gamma_1)$, and $u_0 \in L^2(\Gamma_0)$ is such that there exists $\tilde{u}_0 \in H^1(D)$ with $\tilde{u}_0|_{\Gamma_0} = u_0$. The diffusion matrix $A \in L^\infty\left(D, \mathbb{R}^{d\times d}\right)$ is uniformly elliptic, that is

$$\alpha := \operatorname*{ess\,inf}_{x\in D} \inf_{0\neq v\in\mathbb{R}^d} \frac{(A(x)v)\cdot v}{v\cdot v} > 0 \quad \text{and} \tag{2.2.a}$$

$$\beta := \operatorname*{ess\,sup}_{x\in D} \sup_{0\neq v\in\mathbb{R}^d} \frac{(A(x)v)\cdot v}{v\cdot v} < \infty. \tag{2.2.b}$$

We shall pick up the notation from the introduction and collect the admissible coefficients in the class

$$\mathcal{M}(D,\alpha,\beta) := \{A \in L^\infty(D,\mathbb{R}^{d\times d}) \mid A \text{ satisfies (2.2)}\}$$

Throughout this chapter, we assume that $0 < \alpha \leq \beta < \infty$. We do not assume that A (or any other elements of $\mathcal{M}(D,\alpha,\beta)$) is necessarily symmetric. However, in some places of this manuscript, symmetry will be assumed for the sake of simplicity.

A function $u \in C^1(\overline{D}) \cap C^2(D)$ that satisfies (2.1) is called a *strong/classical solution* of (2.1). The subsequent example shows that such a strong solution does not always exist.

Example 2.1 *Let*

$$\begin{aligned} D &= \{r(\cos\varphi, \sin\varphi) \in \mathbb{R}^2 : 0 \leq r < 1, \varphi \in (0, \frac{3}{2}\pi)\}, \\ \Gamma_0 &= \{r(\cos\varphi, \sin\varphi) \in \mathbb{R}^2 : 0 < r < 1, \varphi \in \{0, \frac{3}{2}\pi\}\}, \\ \Gamma_1 &= \partial D \setminus \Gamma_0,\ u_0 = 0, \end{aligned}$$

$A = E_2$ the 2-dimensional identity matrix, and $g(r,\varphi) = \frac{2}{3}(-\sin\frac{\varphi}{3}, \cos\frac{\varphi}{3})$. Then $u(r,\varphi) = r^{2/3}\sin(\frac{2}{3}\varphi)$ satisfies (2.1) in polar coordinates, but $u \notin C^1(\overline{D})$.

In order to establish the notion of a weak solution of (2.1) we assume - for the time being - that A is smooth ($A \in C^1(\overline{D})$) and that there exists a strong solution u of (2.1) and apply Theorem A.18 with $F = A\nabla u$ and an arbitrary function $v \in C^1(D) \cap C(\overline{D})$ with $v|_{\Gamma_0} = 0$. Since $\operatorname{div} F = \operatorname{div} A\nabla u = -f$, $A\nabla u \cdot n = g$ on Γ_1 and $v = 0$ on Γ_0, we obtain

$$-\int_D f v\,dx + \int_D (A\nabla u)\cdot\nabla v\,dx = \int_{\partial D} v(A\nabla u)\cdot n\,ds = \int_{\Gamma_1} g v\,ds.$$

and, hence,

$$\int_D (A\nabla u)\cdot\nabla v\,dx = \int_D f v\,dx + \int_{\Gamma_1} g v\,ds. \tag{2.3}$$

Since functions $v \in C^1(\overline{D}) \cap C^2(D)$ with $v|_{\Gamma_0} = 0$ are dense in $H^1_{\Gamma_0}(D)$, the equality (2.3) holds in fact for all $v \in H^1_{\Gamma_0}(D)$. Moreover, (2.3) is well defined also for any $u \in H^1(D)$ and any $A \in L^\infty(D, \mathbb{R}^{d\times d})$, a fact that gives rise to the following notion of a weak solution.

Definition 2.1 (weak/generalized solution). A function $u \in H^1(D)$ with $u|_{\Gamma_0} = u_0$ (in the sense of traces) that satisfies

$$a(u,v) := \int_D (A\nabla u)\cdot\nabla v\,dx = \int_D f v\,dx + \int_{\Gamma_1} g v\,ds \tag{2.4}$$

for all $v \in H^1_{\Gamma_0}(D)$ is called a *weak/generalized solution* of (2.1). The variational problem (2.4) is called the weak formulation of the model problem (2.1).

Example 2.2 *The function u from Example 2.1 belongs to $H^1_{\Gamma_0}(D)$ and is a weak solution of* (2.1).

The derivation of the variational problem (2.4) shows that every strong solution of (2.1) is also a weak solution of (2.1) -- provided that $A \in C^1(D, \mathbb{R}^{d\times d})$. If a weak solution belongs to $C^1(\overline{D}) \cap C^2(D)$, then the converse statement is also true.

Theorem 2.1. *Let $u \in H^1(D)$ be a weak solution of* (2.1). *Suppose that $f \in C(\overline{D})$, $g \in C(\overline{\Gamma}_1)$, $u_0 \in C(\Gamma_0)$, $A \in C^1(D, \mathbb{R}^{d\times d})$ and that $u \in C^1(\overline{D}) \cap C^2(D)$. Then u is also a strong solution of* (2.1).

Proof. Equation (2.4) and Theorem A.18 show that

$$\int_D f v\, dx + \int_{\Gamma_1} g v\, ds \overset{(2.4)}{=} \int_D (A\nabla u)\cdot \nabla v\, dx$$
$$\overset{\text{Thm A.18}}{=} -\int_D (\operatorname{div} A\nabla u) v\, dx + \int_{\partial D} v (A\nabla u)\cdot n\, ds \quad (2.5)$$

for all $v \in C^1(\overline{D}) \cap C^2(D)$ with $v|_{\Gamma_0} = 0$. Since the boundary integrals vanish if $v \in \mathcal{D}(D)$, we have

$$\int_D (f + \operatorname{div} A\nabla u) v\, dx = 0 \quad \text{for all } v \in \mathcal{D}(D).$$

The fundamental lemma of the calculus of variations then yields $f + \operatorname{div} A\nabla u = 0$ in D. Since $u \in C(\overline{D})$ and $u_0 \in C(\Gamma_0)$ we have $u(x) = u_0(x)$ for all $x \in \Gamma_0$. It remains to show that $\frac{\partial u}{\partial n} = g$ on Γ_1. Since $f + \operatorname{div} A\nabla u = 0$, equation (2.5) reads

$$\int_{\Gamma_1} g v\, ds = \int_{\partial D} v (A\nabla u)\cdot n\, ds,$$

where we used $v|_{\Gamma_0} = 0$. The resulting identity

$$\int_{\Gamma_1} (g - (A\nabla u)\cdot n) v\, ds = 0 \quad \text{for all } v \in C^1(\overline{D}) \cap C(D)$$

and the fact that $g - (A\nabla u)\cdot n$ is piecewise continuous (because the outer normal n is piecewise continuous on Γ_1) imply $g = (A\nabla u)\cdot n$ on Γ_1.

2.1.1 Existence and uniqueness of a weak solution

The left hand side of the weak formulation (2.4) defines a bilinear form on $H^1_{\Gamma_0}(D) \times H^1_{\Gamma_0}(D)$. This section shows that this bilinear form fulfills the assumptions of the Lax-Milgram-Theorem A.3 and, hence, the existence and uniqueness of a weak solution. However, we shall first check that the solution space $H^1_{\Gamma_0}(D)$ is indeed a Hilbert space.

Theorem 2.2 (The Hilbert space $H^1_{\Gamma_0}(D)$). *The bilinear form*

$$\langle u,v\rangle_{H^1_{\Gamma_0}(D)} = \int_D \nabla u \cdot \nabla v\, dx \quad \textit{for all } u,v \in H^1_{\Gamma_0}(D)$$

is a scalar product on $H^1_{\Gamma_0}(D)$ *and* $H^1_{\Gamma_0}(D)$ *together with* $\langle u,v\rangle_{H^1_{\Gamma_0}(D)}$ *is a Hilbert space. Moreover, the semi-norm* $|\cdot|_{H^1}$ *defined by* $|u|^2_{H^1(D)} = \langle u,u\rangle_{H^1_{\Gamma_0}(D)}$ *for all* $u \in H^1(D)$ *is the induced norm on* $H^1_{\Gamma_0}(D)$.

Remark 2.1. Note that Theorem A.14 defines a different scalar product on the space $H^1_{\Gamma_0}(D)$.

Proof (Proof of Theorem 2.2). Friedrichs' inequality shows for $v \in H^1_{\Gamma_0}(D)$ that

$$\begin{aligned} \|v\|^2_{H^1(D)} &= \|v\|^2_{L^2(D)} + \|\nabla v\|^2_{L^2(D)} \\ &\le C_F^2 \|\nabla v\|^2_{L^2(D)} + \|\nabla v\|^2_{L^2(D)} \\ &= (1+C_F^2)\|\nabla v\|^2_{L^2(D)} \\ &= (1+C_F^2)|v|^2_{H^1(D)} \end{aligned}$$

and implies that $|\cdot|_{H^1(D)}$ is a norm on $H^1_{\Gamma_0}(D)$. Due to Schwarz' inequality, it holds that

$$\langle u,v\rangle_{H^1_{\Gamma_0}(D)} = \int_D \nabla u \cdot \nabla v\, dx \le \|\nabla u\|_{L^2(D)}\|\nabla v\|_{L^2(D)}.$$

This proves that $\langle\cdot,\cdot\rangle_{H^1_{\Gamma_0}(D)}$ is continuous. By the linearity of the trace operator, it holds

$$\gamma_1 v + \gamma_2 w \in H^1_{\Gamma_0}(D) \quad \text{for all } \gamma_1,\gamma_2 \in \mathbb{R} \text{ and } v,w \in H^1_{\Gamma_0}(D).$$

It remains to show that $H^1_{\Gamma_0}(D)$ is complete. Let (v_j) be a Cauchy sequence in $H^1_{\Gamma_0}(D)$, then the completeness of $H^1(D)$ shows that there exists $v \in H^1(D)$ such that $v_j \to v$ in $H^1(D)$. Since the trace operator $\gamma: H^1(D) \to L^2(\partial D)$ is bounded and linear and since $v_j|_{\Gamma_0} = 0$ for all j, it holds that

$$\|v\|_{L^2(\Gamma_0)} = \|v - v_j\|_{L^2(\Gamma_0)} \le \|v - v_j\|_{L^2(\partial D)} \le C_\gamma \|v - v_j\|_{H^1(\Gamma_0)} \to 0 \quad \text{for } j \to \infty.$$

Hence, v vanishes on Γ_0, i.e., $v \in H^1_{\Gamma_0}(D)$. This shows that $H^1_{\Gamma_0}(D)$ is a closed subspace of $H^1(D)$ and finishes the proof. □

With Theorem 2.2 the weak formulation fits into the classical Hilbert space theory at the end of Section A.4.

Theorem 2.3 (Existence and uniqueness of weak solutions -- homogeneous Dirichlet data). *Suppose* $u_0 = 0$ *on* Γ_0. *Then there exists a unique weak solution of* (2.1)*, i.e., there exists a unique* $u \in H^1_{\Gamma_0}(D)$ *such that*

$$a(u,v) = \int_D (A\nabla u)\cdot\nabla v\, dx = \int_D f v\, dx + \int_{\Gamma_1} g v\, ds \quad \textit{for all } v \in H^1_{\Gamma_0}(D). \qquad (2.6)$$

Proof. In order to apply the Lax-Milgram-Theorem A.3, we need to check whether the bilinear form

$$a : H^1_{\Gamma_0}(D) \times H^1_{\Gamma_0}(D), \quad (u,v) \mapsto \int_D (A\nabla u)\cdot \nabla v\, dx$$

is continuous and uniformly elliptic and the linear form

$$F : H^1_{\Gamma_0}(D) \to \mathbb{R},\ v \mapsto \int_D f v\, dx + \int_{\Gamma_1} g v\, ds$$

is bounded.

The boundedness of a follows from Hölder inequalities and (2.2.b),

$$a(u,v) \leq \|A\|_{L^\infty(D,\mathbb{R}^{d\times d})} |u|_{H^1(D)} |v|_{H^1(D)} = \beta |u|_{H^1(D)} |v|_{H^1(D)}. \tag{2.7}$$

The ellipticity follows from the uniform positive definiteness (2.2.a) of A,

$$a(u,u) \geq \alpha \langle u,u\rangle_{H^1_{\Gamma_0}(D)} = \alpha |u|^2_{H^1(D)}. \tag{2.8}$$

For any $v \in H^1_{\Gamma_0}(D)$, Cauchy-Schwarz inequalities in $L^2(D)$ and $L^2(\Gamma_1)$, Friedrichs' inequality, and the linearity of the trace operator imply

$$\begin{aligned}
|F(v)| &= \left|\int_D f v\, dx + \int_{\Gamma_1} g v\, ds\right| \\
&\leq \left|\int_D f v\, dx\right| + \left|\int_{\Gamma_1} g v\, ds\right| \\
&\leq \|f\|_{L^2(D)} \|v\|_{L^2(D)} + \|g\|_{L^2(\Gamma_1)} \|v\|_{L^2(\partial D)} \\
&\leq C_F \|f\|_{L^2(D)} \|\nabla v\|_{L^2(D)} + \|g\|_{L^2(\Gamma_1)} \|v\|_{L^2(\partial D)} \\
&\leq C_F \|f\|_{L^2(D)} \|\nabla v\|_{L^2(D)} + C_\gamma \|g\|_{L^2(\Gamma_1)} \|v\|_{H^1(D)}.
\end{aligned}$$

Since $\|v\|^2_{H^1(D)} \leq (1+C_F^2)\|\nabla v\|^2_{L^2(D)}$ we deduce that

$$|F(v)| \leq (C_F \|f\|_{L^2(D)} + C_\gamma \|g\|_{L^2(\Gamma_1)} (1+C_F^2)^{1/2}) \|\nabla v\|_{L^2(D)}. \tag{2.9}$$

Therefore, F is a bounded linear functional on $H^1_{\Gamma_0}(D)$ and the Lax-Milgram-Theorem (see Corollary A.3) implies the assertion. □

In case of inhomogeneous Dirichlet data $u_0 \neq 0$ we decompose the (unknown) solution into $u = \tilde{u} + \tilde{u}_0$, where $\tilde{u}_0 \in H^1(D)$ satisfies $\tilde{u}_0|_{\Gamma_0} = u_0$ and $\tilde{u} \in H^1_{\Gamma_0}(D)$, i.e., $\tilde{u}|_{\Gamma_0} = 0$ (note that $\tilde{u}_0$ is known and $\tilde{u}$ is unknown). In order to find a weak solution of (2.1) we then try to find $\tilde{u} \in H^1_{\Gamma_0}(D)$ such that

$$\int_D (A\nabla(\tilde{u}+\tilde{u}_0))\cdot \nabla v\, dx = \int_D f v\, dx + \int_{\Gamma_1} g v\, ds \quad \text{for all } v \in H^1_{\Gamma_0}(D),$$

or equivalently

$$\int_D (A\nabla\tilde{u})\cdot\nabla v\,dx = \int_D f v\,dx + \int_{\Gamma_1} g v\,ds - \int_D (A\nabla\tilde{u}_0)\cdot\nabla v\,dx \quad \text{for all } v\in H^1_{\Gamma_0}(D),$$

or equivalently

$$a(\tilde{u},v) = \tilde{F}(v) \quad \text{for all } v\in H^1_{\Gamma_0}(D),$$

where

$$\tilde{F}(v) := \int_D f v\,dx + \int_{\Gamma_1} g v\,ds - a(u_0,v). \tag{2.10}$$

Theorem 2.4 (Existence and uniqueness of weak solutions). *There exists a unique weak solution $u\in H^1(D)$ of the model problem* (2.1) *and*

$$|u|_{H^1} \le \frac{1}{\alpha}\|F\|_{(H^1_{\Gamma_0}(D))^*} \le \frac{1}{\alpha}\Big(\|f\|_{(H^1_{\Gamma_0}(D))^*} + \|g\|_{H^{-1/2}(\Gamma_1)} + \beta|\tilde{u}_0|_{H^1(D)}\Big). \tag{2.11}$$

Proof. The combination of the estimates (2.7) and (2.9) shows $\tilde{F}$ defined in (2.10) above is a bounded linear functional on $H^1_{\Gamma_0}(D)$. The statement then follows from the Lax-Milgram-Theorem A.3. □

2.2 Sobolev regularity for constant coefficient

The remaining part of this chapter discusses the regularity of weak solutions in Sobolev scales. To keep the focus on the impact of the coefficient, we restrict ourselves to the case of homogenous Dirichlet boundary conditions, that is, we assume $\Gamma_1 = \emptyset$ and $u_0 = 0$ throughout the remaining part of this chapter. Moreover, A is assumed to be symmetric or even scalar.

We shall recall some classical regularity results for the case of constant coefficients. Every weak solution u of the model problem (2.4) with constant coefficient $A = E_d$ the d-dimensional identity matrix allows some pointwise first and second derivatives almost everywhere (which coincide with its respective weak or distributional derivatives) written Du and D^2u and the following holds.

Theorem 2.5 (Local regularity). *Let $u\in H^1_0(D)$ be the weak solution of* (2.4) *with $A = E_d$. For any compactly included open set $G \subset\subset D$ there exists some constant $C(G,D) > 0$ such that*

$$Du|_G \in L^2(G;\mathbb{R}^d) \quad \textit{and} \quad D^2u|_G \in L^2(G;\mathbb{R}^{d\times d}_{sym})$$

and

$$\left\|D^2u\right\|_{L^2(G)} \le C(G,D)\,\|f\|_{L^2(D)}\,.$$

This result holds up to the boundary, if ∂D is smooth and $\Gamma_0 = \partial D$ and $u_0 = 0$ in (2.4).

Theorem 2.6 (Global regularity). *Let $u \in H_0^1(D)$ be the weak solution of* (2.4)*. For a C^2 domain D, there exists a constant $C(D) > 0$ such that for any right-hand side $f \in L^2(D)$ the weak solution $u \in H_0^1(D)$ belongs to $H^2(D)$ and*

$$\left\|D^2 u\right\|_{L^2(D)} \leq C(D) \|f\|_{L^2(D)} .$$

Theorem 2.7 (Global higher regularity). *Let $u \in H_0^1(D)$ be the weak solution of* (2.4) *and $2 \leq m \in \mathbb{N}$. For a C^m domain D, there exists a constant $C(D,m) > 0$ such that for any right-hand side $f \in H^{m-2}(D)$ the weak solution $u \in H_0^1(D)$ belongs to $H^m(D)$ and*

$$\|u\|_{H^m(D)} \leq C(D,m) \|f\|_{H^{m-2}(D)}.$$

For more general situations and for proofs we refer to, e.g., the textbook of Evans [2].

The assumption on the smoothness of the boundary of the domain D can be relaxed if D is convex. We consider only the case of homogenous Dirichlet conditions, i.e., $\Gamma_0 = \partial D$ and $u_0 = 0$ in (2.4).

Theorem 2.8 (H^2 regularity for convex domains). *The weak solution $u \in H_0^1(D)$ on any convex Lipschitz domain satisfies that $u \in H^2(D)$. Moreover, it holds that*

$$\left\|D^2 u\right\|_{L^2(D)} \leq \|f\|_{L^2(D)} .$$

A proof for convex polygons may be found in [4].

The results of the previous section can be generalized to the case of smooth coefficients, i.e. for smooth data (domain D and diffusion coefficient A) the condition $f \in H^{m-2}(D)$ implies $u \in H^m(D)$ and there is some constant C (depending on the data and m) such that

$$\|u\|_{H^m(D)} \leq C \|f\|_{H^{m-2}(D)} ,$$

see Theorem 2.7 and popular textbooks, e.g., [3] and [2].

If the domain D is polygonal and/or A is only piecewise smooth and discontinuous along polygonal interfaces, high order regularity can be preserved in weighted Sobolev spaces [6, 4, 7, 5, 1].

Straight-forward computations show that the constant C in the above regularity estimate depends on the L^∞ norms of derivatives of the coefficient A, e.g. $C = C(\|\nabla A\|_{L^\infty(D)})$ for $m = 2$. Considering highly variable A as in Section 1.3.1, this results in a constant $C \approx \frac{1}{\varepsilon}$ for some small parameter ε.

2.3 Sobolev regularity for scalar variable coefficients

Consider a model problem with coefficient $A(x) = \varkappa(x) E_d$ and $\varkappa \in L^\infty(D, [\alpha,\beta])$. The results of the previous section can be generalized to this setting if the data is smooth.

For a smooth $\varkappa$ and domain D, the condition $f \in H^{m-2}(D)$ still implies $u \in H^m(D)$ and the estimate

$$\|u\|_{H^m(D)} \leq C(D,A,m)\|f\|_{H^{m-2}(D)}$$

also holds (see [3] and [2]).

For the case $m = 2$, straight-forward computations show that $C(D,A)$ in the regularity estimate scales like the L^∞-norm of the gradient of A.

Corollary 2.1. *Let $A \in W^{1,\infty}(D) \cap \mathcal{M}(D,\alpha,\beta)$ be scalar and D be a convex Lipschitz domain. Then the weak solution $u \in H^1_0(D)$ of Problem 2.4 with $\Gamma_1 = \emptyset, g = 0, u_0 = 0$ and $f \in L^2(D)$ satisfies $u \in H^2(D)$ and*

$$\left\|D^2 u\right\|_{L^2(D)} \leq C(D,A)\|f\|_{L^2(D)}.$$

Proof. The product rule shows ($A : D \to \mathbb{R}$)

$$\operatorname{div} A\nabla u = A\Delta u + \nabla A \cdot \nabla u,$$

hence

$$\begin{aligned} \Delta u &= A^{-1}(A\Delta u) \\ &= A^{-1} \operatorname{div} A\nabla u - A^{-1}(\nabla A \cdot \nabla u). \end{aligned}$$

This implies

$$\begin{aligned} \|\Delta u\|_{L^2(D)} &\leq \alpha^{-1}\Big(\|\underbrace{\operatorname{div} A\nabla u}_{=-f}\| + \|\nabla A\|_{L^\infty(D)} \underbrace{\|\nabla u\|_{L^2(D)}}_{\leq \alpha^{-\frac{1}{2}} \frac{\operatorname{diam} D}{\pi}\|f\|_{L^2(D)}}\Big) \\ &\leq \underbrace{\alpha^{-1}\left(1 + \alpha^{-\frac{1}{2}} \frac{\operatorname{diam} D}{\pi}\|\nabla A\|_{L^\infty(D)}\right)}_{=C(D,A)}\|f\|_{L^2(D)}. \end{aligned}$$

□

References

1. Constantin Băcuţă, Anna L. Mazzucato, Victor Nistor, and Ludmil Zikatanov. Interface and mixed boundary value problems on n-dimensional polyhedral domains. *Doc. Math.*, 15:687--745, 2010.
2. Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.
3. D. Gilbarg and N.S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer-Verlag, 1983.
4. P. Grisvard. *Elliptic problems in nonsmooth domains*. Pitman (Advanced Publishing Program), Boston, MA, 1985.
5. Ben Qi Guo and Hae Soo Oh. The *h-p* version of the finite element method for problems with interfaces. *Internat. J. Numer. Methods Engrg.*, 37(10):1741--1762, 1994.

6. V. A. Kondratev. Boundary value problems for elliptic equations in domains with conical or angular points. *Trudy Moskov. Mat. Obšč.*, 16:209--292, 1967.
7. Serge Nicaise. Singularities in interface problems. In *Problems and methods in mathematical physics (Chemnitz, 1993)*, volume 134 of *Teubner-Texte Math.*, pages 130--137. Teubner, Stuttgart, 1994.

Chapter 3
Analytical approaches and methods

Abstract This chapter is devoted to a brief survey of G- and H-convergence as well as an constructive approaches in the mathematical theory of homogenization, i.e. the energy method.The chapter is meant to be a short introduction rather than a complete overview. The aim is to get an idea of this field of research. The presentation follows the survey [1] to a large extent.

3.1 G- and H-Convergence

Let $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain and α, β be positive constants such that $0 < \alpha \le \beta < \infty$. Recall the definition of the set $\mathcal{M}(D,\alpha,\beta)$ of admissible diffusion tensors on D with uniform coercivity constant α and $L^\infty(D)$-bound β,

$$\mathcal{M}(D,\alpha,\beta) = \left\{ \begin{matrix} A \in L^\infty(D;\mathbb{R}^{d\times d}) \text{ s.t. } \alpha|\xi|^2 \le (A(x)\xi)\cdot\xi \le \beta|\xi|^2 \\ \text{for a.e. } x \in D \text{ and all } \xi \in \mathbb{R}^d \end{matrix} \right\}. \tag{3.1}$$

In this chapter, we consider a sequence $(A_\varepsilon) \subset \mathcal{M}(D,\alpha,\beta)$ of admissible diffusion tensors indexed by a sequence of positive numbers ε tending to zero. Note that we do not assume symmetry of A_ε in general. We have seen in Chapter 2 that, for any $f \in L^2(D)$ and any A_ε, there exists a unique weak solution $u_\varepsilon \in H_0^1(D)$ of

$$\left.\begin{aligned} -\operatorname{div}(A_\varepsilon \nabla u_\varepsilon) &= f \quad \text{in } D, \\ u_\varepsilon &= 0 \quad \text{on } \partial D, \end{aligned}\right\} \tag{3.2}$$

and u_ε satisfies

$$\|\nabla u_\varepsilon\|_{L^2(D)} \le \alpha^{-1} C_F \|f\|_{L^2(D)};$$

see Lax-Milgram-Theorem A.3.

The mathematical theory of homogenization is concerned with the existence of some limiting coefficient $A_0 \in L^\infty(D;\mathbb{R}^{d\times d})$ such that, for any $f \in L^2(D)$, the sequence of solutions u_ε converges (in a suitable sense) to a limit $u_0 \in H_0^1(D)$ that solves

$$\left.\begin{aligned} -\operatorname{div}(A_0\nabla u_0) &= f \quad \text{in } D,\\ u_0 &= 0 \quad \text{on } \partial D. \end{aligned}\right\} \tag{3.3}$$

If such A_0 exists and if $A_0 \in \mathcal{M}(D,\alpha',\beta')$ for some $0 < \alpha' \leq \beta' < \infty$, it is called *homogenized / effective coefficient* and the corresponding well-posed diffusion problem (3.3) is denoted the *homogenized / effective problem*. Its unique solution u_0 is the *homogenized / effective solution* of (3.2).

Under the additional assumption of symmetry of A_ε, G-convergence as introduced by Spagnolo [8] is an appropriate notion of convergence in this context.

Definition 3.1 (G-convergence). The sequence of symmetric tensors $(A_\varepsilon) \subset \mathcal{M}(D,\alpha,\beta)$ is said to *G-converge* to a limit A_0 as ε tends to zero if, for any $f \in L^2(D)$, the sequence of solutions (u_ε) converges weakly in $H_0^1(D)$ to a limit $u_0 \in H_0^1(D)$, which is the unique solution of the homogenized problem (3.3).

The previous definition makes sense because the following compactness result of the set of admissible symmetric tensors

$$\mathcal{M}_{\text{sym}}(D,\alpha,\beta) := \{A \in \mathcal{M}(D,\alpha,\beta) \,|\, A \text{ is symmetric a.e. in } D\}.$$

Theorem 3.1. *For any sequence $(A_\varepsilon) \subset \mathcal{M}_{sym}(D,\alpha,\beta)$ there exists a subsequence (again indexed by ε) and a homogenized limit $A_0 \in \mathcal{M}_{sym}(D,\alpha,\beta)$ such that A_ε G-converges to A_0.*

Proof. See [8]. □

Remark 3.1 (One-dimensional periodic case.). In Chapter 1, Section 1.3.2, we have seen that in one dimension (where $\mathcal{M}_{\text{sym}} = \mathcal{M}$) sequences of periodic coefficients A_ε of the form

$$A_\varepsilon(x) = A_1(\tfrac{x}{\varepsilon})$$

with $A \in L^\infty(0,1)$ 1-periodic satisfy the definition of G-convergence and the G-limit

$$A_0 = \left(\int_0^1 A_1^{-1}(x)\,dx\right)^{-1}$$

equals the harmonic mean of A_1. Such explicit result holds only in one dimension, no explicit formulas are available in higher dimensions, not even in the periodic case.

We shall state some useful properties of G-convergence.

Proposition 3.1. *(a) If a sequence (A_ε) G-converges, then the G-limit is unique.*

(b) *If (A_ε) and (B_ε) are two sequences that G-converge to A_0 and B_0, respectively, and if $A_\varepsilon = B_\varepsilon$ in some subdomain $D' \subset D$ that is strictly included in D, then $A_0 = B_0$ in D'. This property is called locality of G-convergence.*
(c) *The G-limit of a sequence is independent of the source term f and the boundary condition on ∂D.*
(d) *If A_ε G-converges to A_0, then the sequence of energy densities $(A_\varepsilon \nabla u_\varepsilon) \cdot \nabla u_\varepsilon$ converges to $(A_0 \nabla u_0) \cdot \nabla u_0$ in the sense of distributions in D.*

The generalization of G-convergence so non-symmetric problems is H-convergence that was introduced by Murat and Tartar [5, 6, 7, 9].

Definition 3.2 (H-convergence). The sequence of admissible tensors $(A_\varepsilon) \subset \mathcal{M}(D,\alpha,\beta)$ is said to *H-converge* to a limit A_0 as ε tends to zero if, for any $f \in L^2(D)$, the sequence of solutions (u_ε) converges weakly in $H^1_0(D)$ to a limit $u_0 \in H^1_0(D)$, and the sequence of fluxes $A_\varepsilon \nabla u_\varepsilon$ converges weakly in $L^2(D)$ to $A_0 \nabla u_0$, where u_0 is the unique solution of the homogenized problem (3.3).

This definition makes sense because of the the following compactness result.

Theorem 3.2. *For any sequence $(A_\varepsilon) \subset \mathcal{M}(D,\alpha,\beta)$ there exists a subsequence (again indexed by ε) and a homogenized limit $A_0 \in \mathcal{M}(D,\alpha,\frac{\beta^2}{\alpha})$ such that A_ε H-converges to A_0.*

There is a constructive proof of a variant of the theorem based on the so-called energy method of the next section (cf. Theorem 3.3). A detailed proof may be found in [5] and [7].

Remark 3.2. (a) Like G-convergence, H-convergence satisfies the properties stated in Proposition 3.1, namely uniqueness of the H-limit, locality, independence of source and boundary condition and distributional convergence of $(A_\varepsilon \nabla u_\varepsilon) \cdot \nabla u_\varepsilon$ to $(A_0 \nabla u_0) \cdot \nabla u_0$.
(b) The set $\mathcal{M}(D,\alpha,\beta)$ is not stable with respect to H-convergence, because the $L^\infty(D)$ bound of the H-limit can be increased by a factor $\beta/\alpha \geq 1$. This is a specific effect of non-symmetry of a sequence. (In physical terms, it means that microscopic convective phenomena can yield macroscopic diffusive effects.)
(c) The requirement of convergence of fluxes in the definition of H-convergence is necessary to ensure uniqueness of the H-limit, as seen in the following example.

Example 3.1. Let B be a constant skew-symmetric matrix in $\mathbb{R}^{d\times d}$, i.e., its entries satisfy

$$B_{ij} = -B_{ji} \text{ for all } 1 \leq i,j \leq d.$$

Then any sufficiently smooth function u_0 satisfies

$$-\operatorname{div}(B\nabla u_0) = -\sum_{i,j=1}^{d} B_{ij}\partial_{ij}u_0 = \underbrace{-B}_{\text{skew-sym.}} : \underbrace{D^2 u_0}_{\text{sym.}} = 0$$

Therefore, if u_0 is a solution of the homogenized problem (3.3), then u_0 also solves

$$\begin{aligned} \operatorname{div}\Big((A_0 + B)\nabla u_0\Big) &= f \qquad \text{in } D, \\ u_0 &= 0 \quad \text{on } \partial D. \end{aligned}$$

If the definition of H-convergence would coincide with the definition of G-convergence, this simple calculation would contradict the desired property of uniqueness of the H-limit. The role of the convergence of fluxes is exactly to ensure such uniqueness.

3.2 The energy method

We will now prove the compactness result of Theorem 3.2 using the so-called energy method of Tartar and Murat which is also called the method of oscillating test functions. Beyond its theoretical interest in this context it is also relevant in practical applications because it provides a recipe for the homogenization of any second-order elliptic system. For the sake of simplicity, we present the method for a periodic model problem only. For the non-periodic case we refer to [5, 7, 3, 4].

Let $Y := (0,1)^d$ denote the rescaled unit cube and let $\varepsilon > 0$ be a very small parameter. Given positive constants α,β as in the previous sections, we consider periodic admissible coefficients A_ε of the following form. Let

$$A \in C(D; L^\infty_\#(Y;\mathbb{R}^{d\times d}))$$

and set $A_\varepsilon(x) := A(x, \frac{x}{\varepsilon})$. $A(x,y)$ is smooth with respect to the so-called *slow variable* $x \in D$ and Y-perodic with respect to the so-called *fast variable* $y \in Y$. The function A satisfies the usual coercivity and boundedness conditions, i.e., for any $\xi \in \mathbb{R}^d$ and almost all $(x,y) \in D \times Y$,

$$\alpha|\xi|^2 \le \Big(A(x,y)\xi\Big)\cdot\xi \le \beta|\xi|^2. \tag{3.4}$$

This implies that, for any $\varepsilon > 0$, $A_\varepsilon \in \mathcal{M}(D,\alpha,\beta)$.

The goal of this section is to prove the following homogenization result for a sequence of such periodic coefficients A_ε indexed by a positive zero sequence ε and corresponding solutions $u_\varepsilon \in H^1_0(D)$ of the model diffusion problem (3.2).

Theorem 3.3. *The sequence of solutions* (u_ε) *of* (3.2) *converges weakly in* $H^1_0(D)$ *and strongly in* $L^2(D)$ *to a limit* $u_0 \in H^1_0(D)$ *which is the unique solution of the homogenized problem* (3.3). *The homogenized diffusion coefficient* A_0 *is defined by its entries*

$$\Big[A_0(x)\Big]_{ij} := \int_Y \Big(A(x,y)(e_j + \nabla_y w_j(x,y))\Big) \cdot \Big(e_i + \nabla_y w_i(x,y)\Big) dy, \tag{3.5}$$

where $w_i(x,y)$ are defined, at each point $x \in D$ and each coordinate direction $i \in \{1,\dots,d\}$, as the unique solutions in $H^1_\#(Y)/\mathbb{R}$ of the so-called cell problems

$$\left.\begin{aligned} -\operatorname{div}_y\Big(A(x,y)(e_i + \nabla_y w_i(x,y))\Big) = \quad 0 \quad &\textit{in } Y, \\ y \mapsto w_i(x,y)\ &Y\textit{-periodic}, \end{aligned}\right\} \tag{3.6}$$

with the canonical basis $(e_i)_{i=1}^d$ of $\mathbb{R}^d$.

Remark 3.3. The cell problems (3.6) are well posed. Using Poncaré's inequality Theorem A.22, it is easily checked that the bilinear form

$$(v,w) \mapsto \int_Y \Big(A(x,y)\nabla_y v(y)\Big) \cdot \nabla w(y)\, dy$$

associated with the variational formulation is coercive on periodic functions $H^1_\#(Y)/\mathbb{R}$ (factorized by constants) with coercivity constant $(1 - C_P^2)/\alpha$. Moreover the functional

$$v \mapsto \int_Y -\Big(A(x,y)e_i\Big) \cdot \nabla v\, dy$$

is linear and bounded on $H^1_\#(Y)/\mathbb{R}$ by β so that the Lax-Milgram-Theorem A.3 is applicable.

Remark 3.4. Theorem 3.3 is consistent with our findings in Section 1.3.2, where $D = (0,1)$, $A(x,y) = A_1(y)$ independent of the slow variable. In one dimension the cell problem reads

$$-\frac{\partial}{\partial y}\Big(A(x,y)\frac{\partial}{\partial y} w_1(x,y)\Big) = \frac{\partial}{\partial y} A(x,y)$$

and straight forward computations show that

$$w_1(x,y) = w(y) = -y + \left(\int_0^1 A_1^{-1}(z)\, dz\right)^{-1} \int_0^y A^{-1}(z)\, dz + c$$

and

$$w'(y) = -1 + \left(\int_0^1 A_1^{-1}(z)\, dz\right)^{-1} A^{-1}(y).$$

According to (3.5), the homogenized coefficient, hence, reads

$$A_0 = \left(\int_0^1 A_1^{-1}(z)\, dz\right)^{-1},$$

which is in agreement with Section 1.3.2.

The proof of Theorem 3.3 relies on the following result on periodic functions.

Lemma 3.1. *Let $w \in C(D; L^2_\#(Y))$ and set $w_\varepsilon(x) := w(x, \frac{x}{\varepsilon})$. Then the sequence (w_ε) converges weakly in $L^2(D)$, as ε tends to zero, to*

$$w_0 : x \mapsto \int_Y w(x,y)\,dy \quad \in L^2(D).$$

Proof. A detailed proof may be found, e.g., in [2]. To understand the key idea of the proof, we shall first consider the case where w does not depend on x, w_ε is a ε-periodic function in D, and w_0 is a constant (w_0 is the mean value of w and all w_ε). Let $\mathcal{Q}_\varepsilon$ be a regular subdivision of D into cubes of size ε (shifted and scaled copies of Y). Let

$$\Pi_\varepsilon : L^2(D) \to P_0(\mathcal{Q}_\varepsilon)$$

be the $L^2(D)$ orthogonal projection onto $\mathcal{Q}_\varepsilon$-piecewise constants. Obviously, we have, for any $v \in L^2(D)$ and any $Q \in \mathcal{Q}_\varepsilon$,

$$(\Pi_\varepsilon v)|_Q = \fint_Q v(z)\,dz.$$

In particular, we have that

$$w_0 = \Pi_\varepsilon w = \Pi_\varepsilon w_\varepsilon$$

for any ε. This implies that

$$\begin{aligned}\int_D (w_\varepsilon - w_0) v\,dx &= \int_D (w_\varepsilon - \Pi_\varepsilon w_\varepsilon)(v - \Pi_\varepsilon v)\,dx,\\ &= \int_D w_\varepsilon (v - \Pi_\varepsilon v)\,dx\end{aligned}$$

using the orthogonality of Π_ε. Since $\bigcup_{\varepsilon>0} P_0(\mathcal{Q}_\varepsilon)$ is dense in $L^2(D)$ and $\|w_\varepsilon\|_{L^2(D)} = \|w\|_{L^2(D)}$ for all $\varepsilon > 0$, we finally get

$$0 \le \lim_{\varepsilon\to 0} \left| \int_D (w_\varepsilon - w_0) v\,dx \right| \le \|w\|_{L^2(D)} \lim_{\varepsilon\to 0} \|v - \Pi_\varepsilon v\|_{L^2(D)} = 0.$$

Since $v \in L^2(D)$ was arbitrary, this is the assertion for the case where w is independent of x.

For the general case, observe that we have just proved

$$\begin{aligned}\lim_{\varepsilon\to 0} \left| \int_D (w_\varepsilon - \Pi_\varepsilon w_\varepsilon) v\,dx \right| &= \lim_{\varepsilon\to 0} \left| \int_D w_\varepsilon (v - \Pi_\varepsilon v)\,dx \right|\\ &\le \|w\|_{C(D;L^2_\#(Y))} \lim_{\varepsilon\to 0} \|v - \Pi_\varepsilon v\|_{L^2(D)}\\ &= 0.\end{aligned}$$

Noting that $\Pi_\varepsilon w_\varepsilon \to w_0$ in $L^2(D)$, this is the assertion. □

Proof (Theorem 3.3). Let $\varphi \in \mathcal{D}(D)$ be a smooth function with compact support in D. The oscillating test function φ_ε is defined by

$$\varphi_\varepsilon(x) := \varphi(x) + \varepsilon \sum_{i=1}^{d} \frac{\partial \varphi}{\partial x_i}(x) w_i^*(x, \tfrac{x}{\varepsilon}), \tag{3.7}$$

where $w_i^*(x,\cdot)$ is defined for any $x \in D$ as the unique solution in $H^1_\#(D)/\mathbb{R}$ of the dual cell problem

$$\left.\begin{aligned} -\operatorname{div}_y\Big(A^T(x,y)(e_i + \nabla_y w_i^*(x,y))\Big) &= \quad 0 \quad \text{in } Y, \\ y \mapsto w_i^*(x,y) \text{ Y-periodic.} & \end{aligned}\right\} \tag{3.8}$$

The difference between (3.6) and (3.8) is that $A(x,y)$ has been replaced by its transpose $A^T(x,y)$. If A were symmetrc than $w_i^*(x,y)$ coincides with the solution $w_i(x,y)$ of the (primal) cell problem (3.6). For the sake of simplicity we assume that A is continuously differentiable with respect to x with partial derivatives in $C(D; L^\infty_\#(Y;\mathbb{R}^{d\times d}))$. Under this assumption, the function $x \mapsto w_i^*(x, \frac{x}{\varepsilon}) \in H^1(D)$ and, hence, $\varphi_\varepsilon \in H^1(D)$. Since φ_ε inherits the compact support of φ, we even get $\varphi_\varepsilon \in H^1_0(D)$. Using φ_ε as a test function in the variational formulation of (3.2), we get

$$\int_D \Big(A(x, \tfrac{x}{\varepsilon})\nabla u_\varepsilon(x)\Big)\cdot \nabla\varphi_\varepsilon(x)\,dx = \int_D f(x)\varphi_\varepsilon(x)\,dx. \tag{3.9}$$

It is easy to see that the right-hand side tends to

$$\int_D f(x)\varphi(x)\,dx$$

as ε tends to zero. Since

$$\nabla\varphi_\varepsilon(x) = \sum_{i=1}^{d} \frac{\partial\varphi}{\partial x_i}(x)\Big(e_i + \nabla_y w_i^*(x,\tfrac{x}{\varepsilon})\Big) + \varepsilon\sum_{i=1}^{d}\left(\frac{\partial\nabla\varphi}{\partial x_i}(x) w_i^*(x,\tfrac{x}{\varepsilon}) + \frac{\partial\varphi}{\partial x_i}(x)\nabla_x w_i^*(x,\tfrac{x}{\varepsilon})\right),$$

the left-hand side of (3.9) can be written as

$$\begin{aligned} &\int_D A(x,\tfrac{x}{\varepsilon})\nabla u_\varepsilon(x)\cdot\sum_{i=1}^{d}\frac{\partial\varphi}{\partial x_i}(x)\Big(e_i + \nabla_y w_i^*(x,\tfrac{x}{\varepsilon})\Big)dx \\ &\quad + \varepsilon\int_D A(x,\tfrac{x}{\varepsilon})\nabla u_\varepsilon(x)\cdot\sum_{i=1}^{d}\left(\frac{\partial\nabla\varphi}{\partial x_i}(x)w_i^*(x,\tfrac{x}{\varepsilon}) + \frac{\partial\varphi}{\partial x_i}(x)\nabla_x w_i^*(x,\tfrac{x}{\varepsilon})\right)dx. \end{aligned} \tag{3.10}$$

The second term is bounded by some multiple of

$$\varepsilon \|A\|_{C^1(D;L^\infty(Y;\mathbb{R}^{d\times d}))} \|f\|_{H^{-1}(D)} \|\varphi\|_{C^2(D)} \sum_{i=1}^{d} \|w_i^*\|_{H^1(D)}$$

and, hence, tends to zero in the limit $\varepsilon \to 0$. Integration by parts in the first term of (3.10) yields

$$\begin{aligned}
&\int_D A(x, \tfrac{x}{\varepsilon}) \nabla u_\varepsilon(x) \cdot \sum_{i=1}^{d} \frac{\partial \varphi}{\partial x_i}(x) \Big(e_i + \nabla_y w_i^*(x, \tfrac{x}{\varepsilon}) \Big) dx \\
&= - \int_D u_\varepsilon(x) \underbrace{\operatorname{div}\Big(A^T(x, \tfrac{x}{\varepsilon}) \sum_{i=1}^{d} \frac{\partial \varphi}{\partial x_i}(x) \Big(e_i + \nabla_y w_i^*(x, \tfrac{x}{\varepsilon}) \Big) \Big)}_{=: g(x, \frac{x}{\varepsilon})} dx.
\end{aligned}$$

Computing the divergence in g with $y = y(x) = \frac{x}{\varepsilon}$ yields

$$\begin{aligned}
g(x, y) = {} & \operatorname{div}_x \Big(A^T(x, y) \sum_{i=1}^{d} \frac{\partial \varphi}{\partial x_i}(x) \Big(e_i + \nabla_y w_i^*(x, y) \Big) \Big) \\
& + \frac{1}{\varepsilon} \sum_{i=1}^{d} \frac{\partial \varphi}{\partial x_i}(x) \underbrace{\operatorname{div}_y \Big(A^T(x, y) \Big(e_i + \nabla_y w_i^*(x, y) \Big) \Big)}_{= 0 \text{ by } (3.8)}.
\end{aligned}$$

Therefore, $g \in C^0(D; L^\infty_\#(Y))$ and $x \mapsto g(x, \frac{x}{\varepsilon})$ converges weakly in $L^2(D)$ to its average

$$x \mapsto \int_Y g(x, y)\, dy$$

by Lemma 3.1. Recall further that (u_ε) is bounded in $H^1(D)$ and the Rellich theorem (Theorem A.21) ensures that there is a subsequence (again indexed by ε) and a limit $u_0 \in L^2(D)$ such that $u_\varepsilon \to u_0$ strongly in $L^2(D)$. This allows us to pass to the limit in (3.10) and

$$\begin{aligned}
\lim_{\varepsilon \to 0} \int_D & (A(x, \tfrac{x}{\varepsilon}) \nabla u_\varepsilon(x)) \cdot \nabla \varphi_\varepsilon(x)\, dx \\
&= - \int_D u_0(x) \operatorname{div}_x \underbrace{\Big(\int_Y A^T(x, y) \sum_{i=1}^{d} \frac{\partial \varphi}{\partial x_i}(x) \Big(e_i + \nabla_y w_i^*(x, y) \Big) dy \Big)}_{=: A_0^T(x) \nabla \varphi(x)} dx \\
&= \int_D \nabla u_0(x) \cdot \Big(A_0^T(x) \nabla \varphi(x) \Big) dx \\
&= \int_D \Big(A_0(x) \nabla u_0(x) \Big) \cdot \nabla \varphi(x)\, dx \\
&\overset{(3.9)}{=} \int_D f(x) \varphi(x)\, dx.
\end{aligned}$$

As $\varphi \in \mathcal{D}(D)$ was arbitrary, and by density of $\mathcal{D}(D)$ in $H_0^1(D)$, the result holds for all $\varphi \in H_0^1(D)$.

Hence u_0 satisfies the homogenized problem (3.3). Since A_0 satisfies the same coercivity bound as A, the Lax-Milgram-Theorem A.3 states that the solution $u_0 \in H_0^1(D)$ of (3.3) is unique. This implies that any subsequence (u_ε) converges to the same limit u_0 and, hence, the entire sequence (u_ε) converges strongly in $L^2(D)$ to the homogenized solution u_0. The weak convergence in $H_0^1(D)$ follows from the following duality argument.

Let $F \in H^{-1}(D)$. Then, for any $\varepsilon > 0$, there exists $z_\varepsilon \in H_0^1$ such that

$$F(u_\varepsilon) = \int_D \Big(A_\varepsilon(x) \nabla u_\varepsilon(x) \Big) \cdot \nabla z_\varepsilon(x)\, dx = \int_D f(x) z_\varepsilon(x)\, dx.$$

The previous proof shows that $z_\varepsilon \to z_0$ strongly in $L^2(D)$, where $z_0 \in H_0^1(D)$ solves

$$\int_D \Big(A_0(x) \nabla v(x) \Big) \cdot \nabla z_0(x)\, dx = F(v) \qquad \text{for all } v \in H_0^1(D).$$

This yields, for $\varepsilon \to 0$, that

$$F(u_\varepsilon) = \int_D f(x) z_\varepsilon(x)\, dx \to \int_D f(x) z_0(x)\, dx = F(u_0). \quad \square$$

Remark 3.5. (a) Theorem 3.3 implies that the sequence of diffusion coefficients (A_ε) H-converges to the homogenized coefficient A_0 defined in (3.5). To see this, let $f \in L^2(D)$ and $\varphi \in \mathcal{D}(D)$. Then

$$\int_D A_\varepsilon \nabla u_\varepsilon \cdot \nabla \varphi\, dx = \int_D A_\varepsilon \nabla u_\varepsilon \cdot \nabla \varphi_\varepsilon\, dx + \underbrace{\int_D A_\varepsilon \nabla u_\varepsilon \cdot \nabla(\varphi - \varphi_\varepsilon)\, dx}_{= \int_D f(\varphi - \varphi_\varepsilon)\, dx}, \qquad (3.11)$$

with φ_ε defined in (3.7).

The previous proof shows that

$$\int_D A_\varepsilon \nabla u_\varepsilon \cdot \nabla \varphi_\varepsilon\, dx \to \int_D A_0 \nabla u_0 \cdot \nabla \varphi\, dx$$

while the second term on the right-hand side of (3.11) tends to zero as $\varepsilon \to 0$. Since $\varphi \in \mathcal{D}(D)$ was arbitrary, this is the weak convergence of fluxes which was the missing piece for H-convergence.

(b) Recall that the homogenized coefficient A_0 satisfies the same coercivity bound as A_ε but the $L^\infty(D)$-bound may be amplified by a factor β/α.

(c) Note that even for scalar A_ε the homogenized coefficient might be truly matrix valued.

References

1. Grégoire Allaire. Mathematical approaches and methods. In Ulrich Hornung, editor, *Homogenization and porous media*, volume 6 of *Interdisciplinary Applied Mathematics*, pages 225--250. Springer-Verlag, New York, 1997.
2. J. M. Ball and F. Murat. $W^{1,p}$-quasiconvexity and variational problems for multiple integrals. *J. Funct. Anal.*, 58(3):225--253, 1984.
3. Doïna Cioranescu and François Murat. Un terme étrange venu d'ailleurs. II. In *Nonlinear partial differential equations and their applications. Collège de France Seminar, Vol. III (Paris, 1980/1981)*, volume 70 of *Res. Notes in Math.*, pages 154--178, 425--426. Pitman, Boston, Mass.-London, 1982.
4. Doina Cioranescu and François Murat. A strange term coming from nowhere [MR0652509 (84e:35039a); MR0670272 (84e:35039b)]. In *Topics in the mathematical modelling of composite materials*, volume 31 of *Progr. Nonlinear Differential Equations Appl.*, pages 45--93. Birkhäuser Boston, Boston, MA, 1997.
5. F. Murat and L. Tartar. H-convergence. *Séminaire d'Analyse Fonctionnelle et Numérique de l'Université d'Alger*, 1978.
6. François Murat and Luc Tartar. Calculus of variations and homogenization [MR0844873 (87i:73059)]. In *Topics in the mathematical modelling of composite materials*, volume 31 of *Progr. Nonlinear Differential Equations Appl.*, pages 139--173. Birkhäuser Boston, Boston, MA, 1997.
7. François Murat and Luc Tartar. H-convergence. In *Topics in the mathematical modelling of composite materials*, volume 31 of *Progr. Nonlinear Differential Equations Appl.*, pages 21--43. Birkhäuser Boston, Boston, MA, 1997.
8. S. Spagnolo. Sulla convergenza di soluzioni di equazioni paraboliche ed ellittiche. *Ann. Scuola Norm. Sup. Pisa (3) 22 (1968), 571-597; errata, ibid. (3)*, 22:673, 1968.
9. L. Tartar. Quelques remarques sur les systèmes semi-linéaires. In *Étude numérique des grands systèmes (Proc. Sympos., Novosibirsk, 1976)*, volume 7 of *Méthodes Math. Inform.*, pages 198--212. Dunod, Paris, 1978.

Chapter 4
Finite element approximation of the homogenized model problem

Abstract This chapter aims to study the numerical solution of the homogenized equations derived in the previous chapter. We will focus on linear finite elements and investigate the error committed by this approach when compared with either the homogenized or the original model problem.

4.1 Elliptic model problems

Consider the abstract variational problem posed in some closed subspace $V \subset H^1(D)$ of $H^1(D)$: Find $u \in V$ such that

$$a(u,v) = F(v) \text{ for all } v \in V. \tag{4.1}$$

Here, the bilinear form $a : V \times V \to \mathbb{R}$ and the linear functional $F : V \to \mathbb{R}$ are such that the Lax-Milgram-Theorem A.3 is applicable, i.e., a is V-elliptic and bounded and F is bounded. In this case there is a unique solution $u \in V$ of (4.1) and

$$\|u\|_{H^1(D)} \leq \gamma^{-1} \|F\|_{V'}, \tag{4.2}$$

where γ is the ellipticity constant and V' is the dual space of V.
This abstract model problem covers all variational problems that we have seen in Chapter 3, the oscillatory problem (3.2), the homogenized problem (3.3), as well as the cell problem (3.6). In all these cases, the bilinear form a is defined by

$$a(u,v) := \int_D \Big(A(x)\nabla u(x)\Big) \cdot \nabla v(x)\, dx$$

for $u, v \in V$ with some admissible $A \in \mathcal{M}(D,\alpha,\beta)$ and $0 < \alpha \leq \beta$. Here A refers to A_ε, A_0 or A depending on the context. The other thing that varies is the boundary condition encoded in the definition of V so that

$$V = \begin{cases} H_0^1(D) & \text{in (3.2) and (3.3)}, \\ H_\#^1(Y) & \text{in (3.6)}. \end{cases}$$

In all cases, γ is proportional to α.

4.2 Finite element meshes

The finite element method is based on a description of the physical domain D (or the unit cell Y) in terms of regular simplicial meshes. Hence, in this lecture, we will make the simplifying assumption that D is an open bounded connected polyhedral domain in $\mathbb{R}^d$ for $d \in \{1,2,3\}$.

Definition 4.1 (Regular simplicial mesh). A finite subdivision

$$\mathcal{T} := \{T_j \mid 1 \le j \le N_{\mathcal{T}}\}$$

of $D \subset \mathbb{R}^d$ into closed non-empty simplices (denoted *elements*), i.e.

$$\overline{D} = \bigcup_{j=1}^{N_{\mathcal{T}}} T_j,$$

is said to be *regular* if any two elements $T_1, T_2 \in \mathcal{T}$ are either disjoint or share exactly on vertex or one edge ($d \ge 2$) or one face ($d = 3$).

The set of vertices of a regular mesh is denoted by $\mathcal{N}(\mathcal{T})$. By $\mathcal{F}(\mathcal{T})$ we will refer to the set of $(d-1)$-dimensional intersections of elements, i.e., the set of edges in $2d$ and the set of faces in $3d$.

In this lecture, we will often quantify properties of associated finite element spaces by a single mesh parameter - the *mesh size*, typically denoted by H or h. The mesh size is the maximal diameter of elements from the mesh. We will index the mesh by the mesh global size and write $\mathcal{T}_h$. Doing this, we will not be able to exploit any sort of local mesh refinement in the error analysis and we will restrict ourselves to *quasi-uniform* meshes. This is a reasonable choice for many homogenization problems. Recall that quasi-uniform means that there are constants $c(\mathcal{T}_h)$, $C(\mathcal{T}_h) > 0$ such that, for any two elements $T, K \in \mathcal{T}_h$,

$$c(\mathcal{T}_h)|T| \le |K| \le C(\mathcal{T}_h)|T|. \tag{4.3}$$

We also assume meshes to be *non-degenerate* (or *shape regular*) in the sense that there is a universal constant $\varrho(\mathcal{T}_h) > 0$ such that, for any $T \in \mathcal{T}_h$,

$$\left(|T|^{1/d} \le\right) \operatorname{diam}(T) \le \varrho(\mathcal{T}_h)|T|^{1/d}. \tag{4.4}$$

Of course, given a single regular mesh, one can always find constants $c(\mathcal{T}_H)$, $C(\mathcal{T}_h)$ and $\varrho(\mathcal{T}_H)$ such that (4.3) and (4.4) hold true. The point is that we want these constants to be independent of the mesh size. Whenever we deal with families of meshes $\{\mathcal{T}_h\}_{h>0}$ indexed by the mesh size h, we implicitly assume that the constants $c(\mathcal{T}_h)$, $C(\mathcal{T}_h)$ and $\varrho(\mathcal{T}_h)$ are independent of h, so that the mesh properties (4.3) and (4.4) remain unchanged under mesh refinement $h \to 0$.

4.3 Finite element spaces

Given a regular mesh $\mathcal{T}_h$ with mesh size $h > 0$, we define the finite element space

$$S^1(\mathcal{T}_h) := \{v_h \in C^0(\bar{D}) | \forall T \in \mathcal{T}_h : v_h|_T \text{ is affine}\}$$

of continuous $\mathcal{T}_h$-piecewise affine functions over D. A basis of $S^1(\mathcal{T}_h)$ is given by the so-called hat functions Λ_z associated with the vertices $z \in \mathcal{N}(\mathcal{T}_h)$ of the mesh. For any $z \in \mathcal{N}(\mathcal{T}_h)$, $\Lambda_z \in S^1(\mathcal{T}_h)$ is the unique function that takes values

$$\Lambda_z(y) = \begin{cases} 1, & z = y \\ 0, & z \neq y \end{cases}$$

in the vertices $y \in \mathcal{N}(\mathcal{T}_h)$. Any finite element function $v_h \in S^1(\mathcal{T}_h)$ is hence determined by its values in the vertices of the underlying mesh, i.e.,

$$v_h(x) = \sum_{z \in \mathcal{N}(\mathcal{T}_h)} v_h(z) \Lambda_z(x)$$

for all $x \in \bar{D}$. We will denote by V_h the intersection of $S^1(\mathcal{T}_h)$ with the solution space V of the abstract model problem (4.1) that encodes essential boundary conditions. In the case of homogeneous Dirichlet boundary conditions this means that

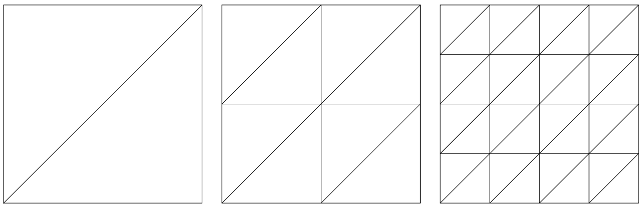
$$V_h = \{v_h \in S^1(\mathcal{T}_h) \,|\, \forall z \in \mathcal{N}(\mathcal{T}_h) \cap \partial D : v_h(z) = 0\}.$$

A meaningful definition of $V_h = S^1(\mathcal{T}_h) \cap H^1_\#(D)$ in the case of periodic boundary conditions for the cell problems requires periodicity of the underlying mesh in the sense that the traces of the mesh on opposite faces of the unit cube $Y = [0,1]^d$ (which is the domain D in this case) match. See Figure 4.1 for an illustration.

4.4 Approximation properties

Before we apply finite elements in a Galerkin approach for the discretization of the variational problem, we shall recall its stability and approximation properties. For

(a)



(b)

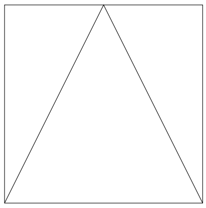


Fig. 4.1 (*a*) A sequence of prototypical periodic meshes of the unit square. (*b*) A mesh that is not periodic.

this purpose we introduce some Clément-type quasi-interpolation operator based on volumetric averaging.

Definition 4.2 (Quasi-interpolation operator). Given some regular mesh $\mathcal{T}_h$, a *quasi-interpolation operator* $I_h : L^2(D) \to S^1(\mathcal{T}_h)$ is defined, for any $v \in L^2(D)$, by its values

$$(I_h v)(z) := \left|\{K \in \mathcal{T}_h \,|\, z \in K\}\right|^{-1} \sum_{T \in \mathcal{T}_h : z \in T} \Pi_T^1(v)(z)$$

in the vertices $z \in \mathcal{N}(\mathcal{T}_h)$ of $\mathcal{T}_h$. Here the operators Π_T^1 from $L^2(T)$ into the set of affine functions are $L^2(T)$-orthogonal projections onto affine functions on the element T, i.e.,

$$\int_T \Pi_T^1(v) w \, dx = \int_T v w \, dx$$

for all affine functions w.

Remark 4.1. The operator I_h in Definition 4.2 is a projection, i.e., for any finite element function $v_h \in S^1(\mathcal{T}_h)$, we have that

$$I_h v_h = v_h.$$

Lemma 4.1 (Stability and approximation of the quasi-interpolation). *The operator I_h of Definition 4.2 has the following properties. There is a constant $C_I > 0$ independent of the mesh size h such that for any $T \in \mathcal{T}_h$ and*

(a) any $v \in L^2(D)$

$$\|I_h v\|_{L^2(T)} \le C_I \|v\|_{L^2(\mathrm{N}(T))}$$

(b) any $v \in H^1(D)$

$$\|\nabla I_h v\|_{L^2(T)} + h^{-1}\|v - I_h v\|_{L^2(T)} \le C_I \|\nabla v\|_{L^2(\mathsf{N}(T))}$$

(c) any $v \in H^2(D)$

$$\|\nabla(v - I_h v)\|_{L^2(T)} + h^{-1}\|v - I_h v\|_{L^2(T)} \le C_I h \left\|D^2 v\right\|_{L^2(\mathsf{N}(T))},$$

where $\mathsf{N}(T)$ *denotes the element patch of* T*, i.e.,*

$$\mathsf{N}(T) := \bigcup \Big\{ K \in \mathcal{T}_h \mid K \cap T \neq \emptyset \Big\}.$$

Remark 4.2. Note that, on a periodic mesh $\mathcal{T}_h$, the quasi-interpolation operator I_h preserves periodicity, i.e., if $v \in H^1_\#(D)$, then $I_h v \in H^1_\#(D) \cap S^1(\mathcal{T}_h)$. In particular, for any constant function v we have that $I_h v = v$.

Proof (Lemma 4.1). For the proof it is useful to rewrite the operator I_h as the concatenation of the $L^2(D)$-orthogonal projection Π^1_h onto the space of (discontinuous) $\mathcal{T}_h$-piecewise affine functions $P_1(\mathcal{T}_h)$ and the averaging operator $E_h : P_1(\mathcal{T}_h) \to S^1(\mathcal{T}_h)$ that maps discontinuous finite element functions onto continuous ones by nodal averaging, i.e., for any $z \in \mathcal{N}(\mathcal{T}_h)$,

$$E_h v_h(z) := \{K \in \mathcal{T}_h \mid z \in K\}^{-1} \sum_{\substack{T \in \mathcal{T}_h: \\ z \in T}} \Big(v_h|_T\Big)(z).$$

Since $\Pi^1_h(v)|_T = \Pi^1_T(v)$ is the $L^2(T)$ best approximation of v on an element T, we have that (a), (b) and (c) are valid for Π^1_h instead of I_h (even with $\mathsf{N}(T)$ replaced by T on the right-hand sides). This follows from the inverse inequality

$$\|\nabla v_h\|_{L^2(T)} \le C \operatorname{diam}(T)^{-1} \|v_h\|_{L^2(T)}$$

and standard approximation results such as the Poincaré inequality or, more generally, the Bramble-Hilbert lemma [2] (see also [3, Ch. 4])

$$\inf_{v_h \in P_1(T)} \|v - v_h\|_{H^k(T)} \le C_m \operatorname{diam}(T)^{m-k} \|v\|_{H^m(T)}$$

for some uniform $C_m > 0$ independent of $\operatorname{diam}(T)$, with $m \in \mathbb{N}$, $k \in \{0, \dots, m-1\}$. The second ingredient are stability and approximation properties of the averaging operator E_h, i.e., for any $T \in \mathcal{T}_h$ and any $v_h \in P_1(\mathcal{T}_h)$,

$$\|E_h v_h\|_{L^2(T)} \le C \|v_h\|_{L^2(\mathsf{N}(T))}$$

and

$$\|(1 - E_h)v_h\|_{L^2(T)} + \operatorname{diam}(T)\|\nabla(1 - E_h)v_h\|_{L^2(T)} \le C h^{-1/2} \sum_{\substack{F \in \mathcal{F}(\mathcal{T}_h): \\ F \subset T}} \|[v_h]_F\|_{L^2(F)},$$

where $[v_h]_F$ denotes the jump of v_h across the face F. For a proof of these bounds we refer to [7].
The assertion follows by the combination of the previous results for Π_h^1 and E_h, the observation that for any $v \in H^1(D)$, $[v]_F = 0$ as well as the trace inequality for H^1 functions. For details see [6] which contains a full proof for an even more general result. □

The operator in Definition 4.1 does not preserve homogeneous Dirichlet boundary conditions as it is based on volumetric averaging. This means that, in general, $v \in H_0^1(D)$ does not imply that $I_h v \in H_0^1(D)$. In connection with Dirichlet problems I_h needs to be modified and one possibility how to do that is given below.

Definition 4.3 (Quasi-interpolation with homogeneous Dirichlet data). Given some regular mesh, a quasi-uniform interpolation operator $I_h^0 : H_0^1(D) \to S^1(\mathcal{T}_h) \cap H_0^1(D)$ that preserves homogeneous Dirichlet boundary conditions is given by

$$I_h^0 v := \underbrace{\sum_{z \in \mathcal{N}(\mathcal{T}_h) \cap D}}_{\text{interior vertices}} (I_h v)(z) \Lambda_z,$$

where I_h is the operator of Definition 4.2.

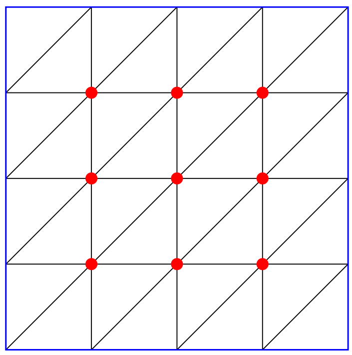


Fig. 4.2 The modified version of the interpolation operator fixes zero values on the boundary (and, hence, all over the boundary) and takes values of $I_h v$ at the interior vertices.

The stability and approximation properties of Lemma 4.1 remain valid under this modification of the quasi-interpolation operator.

Lemma 4.2. *The results of Lemma 4.1 remain valid for I_h replaced by I_h^0 provided that v is taken from $H_0^1(D)$ in (a) and (b) and from $H_0^1(D) \cap H^2(D)$ in (c).*

Proof. The proof for elements T that do not touch the boundary ∂D was already given in Lemma 4.1. For the other elements the proof follows from similar arguments and Friedrichs' inequality. See [6] for details. □

4.5 Galerkin approximation and error estimates

The goal is to approximate the unknown solution $u \in V$ of the abstract model problem (4.1) by a discrete function $u_h \in V_h \subset V$ in some finite element space V_h (cf. Section 4.3) based on a regular mesh $\mathcal{T}_h$ of D. Since u is only implicitely given through the variational problem, the direct approximation by quasi-interpolation is not available.
The most popular approach is Galerkin's method. The idea is to simply replace the infinite-dimensional space V in (4.1) by the finite-dimensional subspace V_h, i.e., the Galerkin method determines an approximation $u_h \in V_h$ of $u \in V$ by solving the discrete variational problem: Find $u_h \in V_h$ such that

$$a(u_h, v_h) = F(v_h) \tag{4.5}$$

for all $v_h \in V_h$. Since V_h is a closed subspace of V, the Lax-Milgram-Theorem A.3 ensures well-posedness of (4.5), i.e., there is a unique solution $u_h \in V_h$ of (4.5) and

$$\|u_h\|_{H^1(D)} \leq \gamma^{-1}\|F\|_{V'};$$

cf. the corresponding bound on u in (4.2).
To quantify the error of this approach, let us re-interpret the Galerkin method as an operator

$$\mathcal{G}_h : V \to V_h$$

that maps any $u \in V$ onto its Galerkin approximation $u_h \in V_h$ that solves (4.5). This operator is obviously linear and

$$a(u, v_h) = F(v_h) = a(\mathcal{G}_h u, v_h) \tag{4.6}$$

for all $v_h \in V_h$. If a is a scalar product, then the property 4.6 is called *Galerkin orthogonality*, because it says that the error $u - \mathcal{G}_h u$ is a-orthogonal on the finite element space V_h, i.e.,

$$a(u - \mathcal{G}_h u, v_h) = 0.$$

In other words, $\mathcal{G}_h$ is the a-orthogonal projection from V onto the subspace V_h. In the general non-symmetric case (4.6) still shows that $\mathcal{G}_h$ is a projection. This suffices to prove quasi-optimality in the sense that, for any $v_h \in V_h$,

$$\begin{aligned} \|u - u_h\|_V &= \|(1 - \mathcal{G}_h)u\|_V \\ &= \|(1 - \mathcal{G}_h)(u - v_h)\|_V \\ &\leq \|1 - \mathcal{G}_h\|_{\mathcal{L}(V,V)}\|u - v_h\|_V \\ &= \|\mathcal{G}_h\|_{\mathcal{L}(V,V)}\|u - v_h\|_V. \end{aligned} \tag{4.7}$$

Here we have used that the operator norms of a projection and its complementary projection in a Hilbert space coincide, i.e.,

$$\|1 - \mathcal{G}_h\|_{\mathcal{L}(V,V)} = \|\mathcal{G}_h\|_{\mathcal{L}(V,V)};$$

see, e.g., [9]. It remains to check that $\mathcal{G}_h$ is bounded. Since a is V-elliptic and bounded, we have that

$$\alpha\|\mathcal{G}_h u\|_V^2 \leq a(\mathcal{G}_h u, \mathcal{G}_h u) \overset{(4.6)}{=} a(u, \mathcal{G}_h u) \leq \beta\|u\|_V\|\mathcal{G}_h u\|_V.$$

Hence, the operator norm of $\mathcal{G}_h$ satisfies

$$\|\mathcal{G}_h\|_{\mathcal{L}(V,V)} = \sup_{0\neq v\in V} \frac{\|\mathcal{G}_h v\|_V}{\|v\|_V} \leq \frac{\beta}{\alpha},$$

where the generic C is related to either the Friedrichs or the Poincaré inequality. This and (4.7) lead to the error estimate for Galerkin's method known as *Céa's lemma*,

$$\|u-u_h\|_V \leq \tfrac{\beta}{\alpha} \min_{v_h\in V_h} \|u-v_h\|_V.$$

If a is symmetric, we even have the better bound

$$\|u-u_h\|_V \leq \sqrt{\tfrac{\beta}{\alpha}} \min_{v_h\in V_h} \|u-v_h\|_V.$$

Inserting $v_h = I_h u$, where I_h is the quasi-interpolation operator of Definition 4.2 (in the case of homogeneous Dirichlet boundary conditions take I_h^0 instead), yields an upper bound of the error. Assuming $u \in H^2(D)$, Lemma 4.1 (resp. Lemma 4.2) yields

$$\|u-u_h\|_V \leq C\tfrac{\beta}{\alpha} C_I' h \left\|D^2 u\right\|_{L^2(D)}.$$

Note that C_I' is a modified version of the constant C_I from Lemma 4.1--4.2.
We have seen in Section 2.3 that in convex domains with scalar coefficient A, $D^2 u \in L^2(D;\mathbb{R}^{d\times d})$ exists provided that $A \in W^{1,\infty}(D)$ and

$$\left\|D^2 u\right\|_{L^2(D)} \leq C\Big(\|A\|_{W^{1,\infty}(D)}, \alpha\Big)\|F\|_{L^2(D)}.$$

For less regular coefficients $A \in W^{s,\infty}(D)$ for some $s \in [0,1]$, we only get bounds for the $H^{1+s}(D)$ norm of u and the convergence rate is reduced correspondingly,

$$\|u-u_h\|_V \leq C\Big(\|A\|_{W^{s,\infty}(D)}, \alpha, \beta\Big) h^s \|F\|_{H^{-s}(D)}. \tag{4.8}$$

We shall now discuss (4.8) for each of the three model problems separately.

4.5.1 *Error of the direct finite element approximation of the oscillatory model problem* (3.2)

For the model problem (3.2) with an admissible coefficient $A_\varepsilon \in \mathcal{M}(D,\alpha,\beta)$, (4.8) yields an error estimate for the Galerkin approximation $u_{\varepsilon,h} \in V_h = S^1(\mathcal{T}_h) \cap H^1_0(D)$ of the solution $u_\varepsilon \in V = H^1_0(D)$ on a regular mesh of width h.
Provided that A_ε is a bit more regular than just L^∞, say $A_\varepsilon \in W^{s,\infty}(D;\mathbb{R}^{d\times d})$, the rate of convergence with respect to the mesh size is s. However, this rate is only observed once the characteristic length scale ε is resolved by the mesh, i.e. $h\varepsilon^{-1} \lesssim 1$, see Chapter 1. In the error estimate this is encoded in the $\|\bullet\|_{W^{s,\infty}(D;\mathbb{R}^{d\times d})}$ norm of the coefficient A_ε which scales like ε^{-s} for typical oscillatory coefficients such as the one given in (1.10).
Altogether, we can expect that

$$\|u_\varepsilon - u_{\varepsilon,h}\|_V \leq C(\alpha,\beta)\left(\tfrac{h}{\varepsilon}\right)^s \|f\|_{L^2(D)}. \tag{4.9}$$

If $A_\varepsilon \in W^{1,\infty}(D;\mathbb{R}^{d\times d})$, then $s = 1$. The estimate (4.9) is sharp in the sense that there are $f \in L^2(D)$ and $c > 0$ independent of h such that

$$\|u_\varepsilon - u_{\varepsilon,h}\|_V \geq c\tfrac{h}{\varepsilon}\|f\|_{L^2(D)}.$$

In the cases of less regularity, the estimate (4.9) can be pessimistic with respect to the rate s because the actual performance also depends on how the underlying mesh is elongated with singularities of A_ε but these subtle details are certainly beyond the scope of this lecture.

4.5.2 *Error of the Galerkin finite element approximation of the homogenized model problem* (3.3)

Here we are facing an admissible coefficient $A_0 \in \mathcal{M}(D,\alpha,\beta^2/\alpha)$. Since variations on the scale of ε have been averaged already, there will be no crucial ε-dependence in the error.
Let $\mathcal{T}_H$ be a regular mesh of D. As before, we use the capital letter H to indicate that the mesh can be significantly larger than the microscopic parameter ε ($H \gg \varepsilon \gg h$). We call H the *scale of observation* or *interest*. Let $u_0 \in H^1_0(D)$ denote the homogenized solution and

$$u_{0,H} \in S^1(\mathcal{T}_H) \cap H^1_0(D)$$

its Galerkin finite element approximation. If $A_0 \in W^{s,\infty}(D;\mathbb{R}^{d\times d})$, then we have the error estimate

$$\|u_0 - u_{0,H}\|_V \le C(\|A_0\|_{W^{s,\infty}(D;\mathbb{R}^{d\times d})}, \alpha, \beta) H^s \|f\|_{L^2(D)}.$$

4.5.3 Error of the finite element approximation of the cell problem

For the cell problem (3.6), we fix $x \in D$ and a coordinate direction $i \in \{1,...,d\}$ and we seek $w_i(x,\cdot) \in H^1_\#(Y)/\mathbb{R}$ such that

$$\int_Y \Big(A(x,y)\nabla_y w_i(x,y)\Big) \cdot \nabla v(y)\, dy = -\int_Y \Big(A(x,y)e_i\Big) \cdot \nabla v(y)\, dy \qquad (4.10)$$

for all $v \in H^1_\#(Y)/\mathbb{R}$. The Galerkin finite element method applied to (4.10) seeks

$$w_{i,\delta}(x,\cdot) \in V_\delta(Y) := S^1(\mathcal{T}_\delta) \cap H^1_\#(Y)/\mathbb{R}$$

such that

$$\int_Y \Big(A(x,y)\nabla_y w_{i,\delta}(x,y)\Big) \cdot \nabla v_\delta\, dy = -\int_Y \Big(A(x,y)e_i\Big) \cdot \nabla v_\delta\, dy \qquad (4.11)$$

for all $v_\delta \in V_\delta$. Here $\mathcal{T}_\delta$ is a regular periodic mesh of $Y = [0,1]^d$. The mesh parameter δ is linked to the fine scale discretization parameter h by rescaling, i.e., $\delta \approx h/\varepsilon$. Before we can apply the error estimate (4.8) we need to compute the L^2-norm of the right-hand side. This requires smoothness of A, that is $A(x,\cdot) \in W^{1,\infty}(Y;\mathbb{R}^{d\times d})$. Observe that

$$\begin{aligned}\left| -\int_Y \Big(A(x,y)e_i\Big) \cdot \nabla v\, dy \right| &= \left| \int_Y \operatorname{div}_y \Big(A(x,y)e_i\Big) v\, dy \right| \\ &\le \|\operatorname{div}_y(A(x,\cdot)e_i)\|_{L^\infty(Y)} \|v\|_{L^2(Y)}.\end{aligned}$$

This means that under the smoothness assumption on the coefficient, the L^2-norm of the functional

$$F_x : v \mapsto \int_Y -\Big(A(x,y)e_i\Big) \cdot \nabla v\, dy$$

can be estimated by

$$\|F_x\|_{L^2(Y)} \le C_d \|A(x,\cdot)\|_{W^{1,\infty}(Y;\mathbb{R}^{d\times d})}.$$

Employing this and Remark (3.3) in (4.8) yields

$$\|w_i(x,\cdot) - w_{i,\delta}(x,\cdot)\|_{H^1(Y)} \le C(\|A(x,\cdot)\|^2_{W^{1,\infty}(Y;\mathbb{R}^{d\times d})}, \alpha, \beta)\, \delta.$$

If $A(x,\cdot) \in W^{s,\infty}(Y;\mathbb{R}^{d\times d})$ is less smooth, we get

$$\|w_i(x,\cdot) - w_{i,\delta}(x,\cdot)\|_{H^1(Y)} \le C(\|A(x,\cdot)\|^2_{W^{s,\infty}(Y;\mathbb{R}^{d\times d})}, \alpha, \beta)\,\delta^s. \tag{4.12}$$

4.6 The heterogeneous multiscale method

In the previous section, we have discussed the finite element approximation of both the homogenized problem and the cell problems that define the corresponding homogenized coefficient. We shall now put these pieces together to derive a computational method for the approximation of the unknown homogenized solution $u_0 \in H^1_0(D)$ that solves (3.3).

Let $A \in C(\bar{D}; L^\infty_\#(Y;\mathbb{R}^{d\times d}))$, where $Y = [0,1]^d$, and set $A_\varepsilon(x) := A(x, \frac{x}{\varepsilon})$ for $x \in D$ as in Section 3.2. Of course, A shall be such that $A_\varepsilon \in \mathcal{M}(D,\alpha,\beta)$ for some $0 < \alpha \le \beta < \infty$. Theorem 3.3 characterizes the homogenized coefficient $A_0 \in \mathcal{M}(D,\alpha,\beta^2/\alpha)$ by

$$[A_0(x)]_{ji} := \int_Y \Big(A(x,y)(e_i + \nabla_y w_i(x,y))\Big) \cdot \Big(e_j + \nabla_y w_j(x,y)\Big)\,dy, \tag{4.13}$$

where $w_i(x,\cdot) \in H^1_\#(Y)/\mathbb{R}$ are defined, at each point $x \in D$, as the unique solutions of (4.10). Note that we can as well drop the last term $\nabla_y w_j(x,y)$ in (4.13) without changing the formula. To see this consider the variational formulation of (3.6) with the test function $w_j(x,\cdot)$.
Replacing the correctors $w_i(x,\cdot)$ by their Galerkin finite element approximations $w_{i,\delta}(x,\cdot) \in S^1(\mathcal{T}_\delta) \cap H^1_\#(Y)/\mathbb{R}$ based on some regular periodic mesh $\mathcal{T}_\delta$ of Y leads to an approximative homogenized coefficient $A_{0,\delta} \in \mathcal{M}(D,\alpha,\beta^2/\alpha)$ defined by

$$[A_{0,\delta}(x)]_{ji} = \int_Y \Big(A(x,y)(e_i + \nabla_y w_{i,\delta}(x,y))\Big) \cdot e_j\,dy \tag{4.14}$$

where $w_{i,\delta}(x,\cdot)$ are solutions of (4.11) for any $x \in D$. The obvious idea for a numerical method is to use $A_{0,\delta}$ instead of A_0. This leads to a perturbed homogenized problem that seeks $u_{0,\delta} \in H^1_0(D)$ such that

$$\int_D \Big(A_{0,\delta} \nabla u_{0,\delta}\Big) \cdot \nabla v\,dx = \int_D f v\,dx \tag{4.15}$$

for all $v \in H^1_0(D)$. We then discretize (4.15) on the coarse regular target mesh $\mathcal{T}_H$ of D by the Galerkin finite element method. This defines an approximation

$$\tilde{u}_{0,H,\delta} \in V_H := S^1(\mathcal{T}_H) \cap H^1_0(D)$$

that satisfies

$$\int_D \Big(A_{0,\delta} \nabla \tilde{u}_{0,H,\delta}\Big) \cdot \nabla v_H\,dx = \int_D f v_H\,dx \tag{4.16}$$

for all $v_H \in V_H$. Assuming that the integral in (4.14) can be computed exactly for any $x \in D$ or approximated to high accuracy at negligible cost (which requires some knowledge of $A(x,\cdot)$ or smoothness), we are left with the issue of evaluating the bilinear form in (4.16) for the assembly of the finite element stiffness matrix. In the heterogeneous multiscale method (HMM) introduced by E and Engquist [4, 5, 1], this issue is resolved by numerical quadrature. Since we are only dealing with linear finite elements, a one-point quadrature rule is sufficient, i.e.,

$$\begin{aligned} a_{0,\delta}(u_H, v_H) &:= \int_D \Big(A_{0,\delta}(x)\nabla u_H\Big)\cdot \nabla v_H\, dx \\ &\approx \sum_{T\in\mathcal{T}_H} |T|\Big(A_{0,\delta}(c_T)\nabla u_H\Big)\cdot \nabla v_H \\ &=: a_{0,H,\delta}(u_H, v_H). \end{aligned} \tag{4.17}$$

This means that $A_{0,\delta}$ is replaced by the $\mathcal{T}_H$-piecewise constant approximation $A_{0,H,\delta} \in P_0(\mathcal{T}_H;\mathbb{R}^{d\times d})\cap \mathcal{M}(D,\alpha,\beta^2/\alpha)$ defined by

$$A_{0,H,\delta}|_T := A_{0,\delta}(c_T) \tag{4.18}$$

for any $T\in\mathcal{T}_H$, where $c_T := \fint_T x\, dx$ denotes the centroid of T.
The finite element heterogeneous multiscale method, hence, seeks $u_{0,H,\delta}\in V_H = S^1(\mathcal{T}_H)\cap H^1_0(D)$ such that

$$a_{0,H,\delta}(u_{0,H,\delta}, v_H) = \int_D (A_{0,H,\delta}\nabla u_{0,H,\delta})\cdot \nabla v_H\, dx = \int_D f v_H\, dx \tag{4.19}$$

for all $v_H\in V_H$. We have seen in the previous section that (4.19) is a well-posed problem. The assembly of the corresponding system of linear equations in the basis coefficients is standard. The difference to a standard finite element method is that the evaluation of the approximative homogenized coefficient in the centroids requires the solution of the finite element problem (4.10) with respect to the mesh $\mathcal{T}_\delta$ of Y. In the fully periodic case where $A(x,y) = A(y)$ is independent of the slow variable, all cell problems coincide and only one of them needs to be solved.

4.7 Error analysis of HMM approximation of the homogenized solution

The following theorem quantifies the error of the method under certain smoothness assumptions on A.

Theorem 4.1 (Error of the FE-HMM). *Assume we have a coefficient*

$$A\in C^1(\bar{D}; W^{1,\infty}_{\#}(Y;\mathbb{R}^{d\times d}))$$

with

$$A_\varepsilon := A(\cdot,\tfrac{\cdot}{\varepsilon})\in \mathcal{M}(D,\alpha,\beta)$$

for any $\varepsilon > 0$. Then the error of the FE-HMM approximation $u_{0,H,\delta} \in V_H$ that solves (4.19) *can be bounded by*

$$\|u_0 - u_{0,H,\delta}\|_{H^1(D)} \leq C(\alpha,\beta)\Big[\|A\|^2_{C^0(D;W^{1,\infty}(Y;\mathbb{R}^{d\times d}))}\delta + \|A\|_{C^1(D;L^\infty(Y;\mathbb{R}^{d\times d}))}H\Big] \quad (4.20)$$

with some generic constant $C(\alpha,\beta) > 0$ independent of the mesh parameters H and δ.

Proof. We shall split the error $e := u_0 - u_{0,H,\delta}$ into three parts

$$e = \underbrace{(u_0 - u_{0,\delta})}_{=:e_\delta} + \underbrace{(u_{0,\delta} - \tilde{u}_{0,H,\delta})}_{=:e_H} + \underbrace{(\tilde{u}_{0,H,\delta} - u_{0,H,\delta})}_{=:\tilde{e}_H},$$

the error e_δ due to discretization of the cell problems, the error e_H due to discretization of the perturbed homogenized problem, and the error $\tilde{e}_H$ due to numerical quadrature (one-point) of the approximative homogenized coefficient. All errors will be estimated separately below.

(a) *Estimation of e_δ:* Straight-forward computations show

$$\begin{aligned}
\alpha\|\nabla e_\delta\|^2_{L^2(D)} &\leq \int_D \Big(A_{0,\delta}\nabla e_\delta\Big)\cdot \nabla e_\delta \, dx \\
&= \int_D \Big(A_{0,\delta}\nabla u_0\Big)\cdot \nabla e_\delta \, dx - \int_D \Big(A_{0,\delta}\nabla u_{0,\delta})\cdot \nabla e_\delta \, dx \\
&= \int_D \Big((A_{0,\delta} - A_0)\nabla u_0\Big)\cdot \nabla e_\delta \, dx \\
&\leq \|A_{0,\delta} - A_0\|_{L^\infty(D;\mathbb{R}^{d\times d})}\|\nabla u_0\|_{L^2(D)}\|\nabla e_\delta\|_{L^2(D)}.
\end{aligned}$$

The error in the homogenized coefficient can be estimated pointwise for any $x \in D$ and any component $(j,i) \in \{1,...,d\}^2$ by

$$\begin{aligned}
\Big|[A_0(x)]_{ji} - [A_{0,\delta}(x)]_{ji}\Big| &= \int_Y \Big(A(x,y)(e_i + \nabla_y w_i(x,y))\Big)\cdot e_j - \Big(A(x,y)(e_i + \nabla_y w_{i,\delta}(x,y))\Big)\cdot e_j \, dy \\
&= \int_Y \Big(A(x,y)\nabla_y(w_i(x,y) - w_{i,\delta}(x,y))\Big)\cdot e_j \, dy \\
&\leq \beta\|\nabla_y(w_i(x,\cdot) - w_{i,\delta}(x,\cdot))\|_{L^2(Y)} \\
&\overset{(4.12)}{\leq} C(\alpha,\beta)\|A(x,\cdot)\|^2_{W^{1,\infty}(Y;\mathbb{R}^{d\times d})}\delta.
\end{aligned}$$

The combination of the previous two estimates and the well-posedness of the homogenized problem yield

$$\|\nabla e_\delta\|_{L^2(D)} \leq C\Big(\alpha,\beta,\|A\|^2_{C^0(\bar{D};W^{1,\infty}(Y;\mathbb{R}^{d\times d}))}\Big)\delta\|f\|_{L^2(D)} \quad (4.21)$$

with some constant C independent of δ.

For the remaining part of this proof we introduce the short hand notation $a \lesssim b$ whenever there is a constant C that may depend on α, β and derivatives of A but not on mesh parameters δ, H and the microscopic length scale ε such that $a \leq Cb$. With this notation (4.21) reads

$$\|\nabla e_\delta\|_{L^2(D)} \lesssim \delta \|f\|_{L^2(D)}.$$

(b) *Estimation of* e_H*:* This error is the standard discretization error when approximating the solution of (4.15) by the Galerkin approximation that solves (4.16). This error has been quantified in Section 4.5.2 and we get

$$\|u_{0,\delta} - \tilde{u}_{0,H,\delta}\|_{L^2(D)} \lesssim H\|f\|_{L^2(D)}.$$

Here the hidden constant depends on $\|A_{0,\delta}\|_{W^{1,\infty}(D;\mathbb{R}^{d\times d})}$ which can be bounded in terms of $\|A\|_{C^1(\bar{D};L^\infty(Y;\mathbb{R}^{d\times d}))}$.

(c) *Estimation of* $\tilde{e}_H$*:* The same arguments as in (a) yield

$$\|\nabla \tilde{e}_H\|_{L^2(D)} \leq \alpha^{-1}\|A_{0,\delta} - A_{0,H,\delta}\|_{L^\infty(D;\mathbb{R}^{d\times d})}\|\nabla u_{0,\delta}\|_{L^2(D)}.$$

For any $T \in \mathcal{T}_H$, some $x \in T$ and $(i, j) \in \{1, ..., d\}^2$, we have that

$$\begin{aligned} |[A_{0,\delta}(x)]_{ij} - [A_{0,H,\delta}(x)]_{ij}| &= |[A_{0,\delta}(x)]_{ij} - [A_{0,\delta}(c_T)]_{ij}| \\ &\lesssim \|A_{0,\delta}\|_{C^1(D;\mathbb{R}^{d\times d})} \underbrace{|x - c_T|}_{\leq H} \\ &\lesssim H\|A\|_{C^1(\bar{D};L^\infty(Y;\mathbb{R}^{d\times d}))}. \end{aligned}$$

The combination of the previous estimates and the stability of Galerkin's method yield

$$\|\nabla \tilde{e}_H\|_{L^2(D)} \lesssim H\|f\|_{L^2(D)}.$$

The combination of the results of (a) - (c) readily yields the assertion. □

Assuming symmetry of A, the bound can be improved,

$$\|u_0 - u_{0,H,\delta}\|_{H^1(D)} \leq C(\alpha,\beta)\Big[\|A\|^2_{C^0(D;W^{1,\infty}(Y;\mathbb{R}^{d\times d}))}\delta^2 + \|A\|_{C^1(D;L^\infty(Y;\mathbb{R}^{d\times d}))}H\Big].$$

If we measure the error in the $L^2(D)$ norm, then Theorem 4.1 and Friedrichs' inequality lead to the trivial bound

$$\|u_0 - u_{0,H,\delta}\|_{L^2(D)} \leq C(\alpha,\beta,\|A\|_{C^1(\bar{D};W^{1,\infty}_\#(Y;\mathbb{R}^{d\times d}))}(H + \delta^2)$$

in the symmetric case. In order to get improved bounds, we need to assume additional regularity with respect to the slow variable. If $A \in C^2(\bar{D}; W^{1,\infty}_\#(Y;\mathbb{R}^{d\times d}))$, then a modification of the arguments of the proof of Theorem 4.1 and the use of duality arguments lead to rate-optimal bounds for the error of the HMM in the L^2 norm

$$\|u_0 - u_{0,H,\delta}\|_{L^2(D)} \leq C(\|A\|_{C^2(\bar{D};W^{1,\infty}_\#(Y;\mathbb{R}^{d\times d}))},\alpha,\beta)(\delta + H)^2. \tag{4.22}$$

Note that for this error bound the choice of the centroids as quadrature points is crucial.

4.8 Error of HMM with respect to the original oscillatory solution

In contrast to the previous section, where we have studied the error in approximating the homogenized solution u_0, we are now aiming at the error of the HMM in approximating the original oscillatory solution u_ε for some fixed $\varepsilon > 0$. Clearly, we will have to measure the error in the $L^2(D)$ norm as in the $H^1(D)$ norm not even the best-approximation in the finite element space is an appropriate approximation unless ε is resolved by the underlying mesh (which is not the regime we are interested in in homogenization problems). We have seen in the previous section that the error $u_0 - u_{0,H,\delta}$ of the HMM when compared with the homogenized solution is $O(H^2 + \delta^2)$; see (4.22). What remains is a quantification of the rate of convergence of u_ε to u_0 in $L^2(D)$. This quantification is independent of the actual discretization scheme and is merely a supplement of Chapter 3. We cite only a recent result of [8], which is the sharpest one for our purposes.

Theorem 4.2 (Quantified homogenization error). *Let the assumption of Theorem 3.3 (energy method) be satisfied and assume further that A is symmetric, ε-periodic and Hölder continuous with exponent $\lambda > 0$ and that the homogenized solution satisfies $u_0 \in H^2(D) \cap H_0^1(D)$. Then*

$$\|u_\varepsilon - u_0\|_{L^2(D)} \leq C(\alpha, \beta, \lambda, \varepsilon \|u_0\|_{H^2(D)}$$

with some uniform constant $C > 0$ independent of ε.

Proof. The result is part of [8, Theorem 3.4] and we refer to the original work for a proof. □

Corollary 4.1. *Let the assumptions of Theorem 4.1 and 4.2 be satisfied. Then*

$$\|u_\varepsilon - u_{0,H,\delta}\|_{L^2(D)} \leq C(\|A\|_{C^1(\bar{D}; W_\#^{1,\infty}(Y;\mathbb{R}^{d\times d}))}, \alpha, \beta)(\varepsilon + \delta^2 + H).$$

Moreover, if $A \in C^2(\bar{D}; W_\#^{1,\infty}(Y;\mathbb{R}^{d\times d}))$ is such that (4.22) *holds, then*

$$\|u_\varepsilon - u_{0,H,\delta}\|_{L^2(D)} \leq C(\|A\|_{C^2(\bar{D}; W_\#^{1,\infty}(Y;\mathbb{R}^{d\times d}))}, \alpha, \beta)(\varepsilon + \delta^2 + H^2).$$

References

1. A. Abdulle, W. E, B. Engquist, and E. Vanden-Eijnden. The heterogeneous multiscale method. *Acta Numer.*, 21:1--87, 2012.

2. J. H. Bramble and S. R. Hilbert. Bounds for a class of linear functionals with applications to Hermite interpolation. *Numer. Math.*, 16:362--369, 1970/1971.
3. S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer Verlag, New York, 1994.
4. W. E and B. Engquist. The heterogeneous multiscale methods. *Commun. Math. Sci.*, 1(1):87--132, 2003.
5. W. E and B. Engquist. The heterogeneous multi-scale method for homogenization problems. In *Multiscale methods in science and engineering*, volume 44 of *Lect. Notes Comput. Sci. Eng.*, pages 89--110. Springer, Berlin, 2005.
6. A. Ern and J.-L. Guermond. Finite element quasi-interpolation and best approximation. *ArXiv e-prints*, May 2015.
7. Ohannes A. Karakashian and Frederic Pascal. A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems. *SIAM J. Numer. Anal.*, 41(6):2374--2399 (electronic), 2003.
8. Carlos E. Kenig, Fanghua Lin, and Zhongwei Shen. Convergence rates in L^2 for elliptic homogenization problems. *Arch. Ration. Mech. Anal.*, 203(3):1009--1036, 2012.
9. Daniel B. Szyld. The many proofs of an identity on the norm of oblique projections. *Numer. Algorithms*, 42(3-4):309--323, 2006.