Daniel Peterseim

# Numerical Homogenization beyond Scale Separation and Periodicity

Lecture notes for participants of the

AMSI Winter School on Computational Modeling of Hetergogeneous Media

Queensland University of Technology, July 1-12, 2019.

Preliminary version of 21st June 2019.

# Preface

The objective of these lecture notes is to introduce quickly the reader to numerical homogenization. The choice of the material is rather personal and strongly influenced by our own work in this context. The manuscript is not meant to give a complete overview of numerical homogenization and its mathematical background but to make the reader familiar with the underlying ideas of a central approach in this context.

These notes where written during courses held at the University of Bonn in 2014 and 2016, the Hausdorff Institute of Mathematics Bonn in 2017, the University of Augsburg in 2017, the Technical University of Athens in 2018 and Mathematisches Forschungsinstitut Oberwolfach in 2019. Right now the notes are turned into a textbook. This is a joint project with Axel Målqvist who already contributed significant parts to these notes. Moreover, I would like to thank Denis Düsseldorf, Dietmar Gallistl, Roland Maier, Mira Schedensack, and Dora Varga for their help in producing and proof-reading earlier drafts. The manuscript is still under construction and has a very preliminary character. Some parts certainly need improvement and completion.

Augsburg, June 2019 *Daniel Peterseim*

# Contents

# Chapter 1
# Multiscale Problems

Many physical processes in microheterogeneous media such as modern composite and functional materials are described by partial differential equations (PDEs) with rough coefficients or domains with a complex microstructure. Given the complexity of these processes, the key to reliably simulate some relevant classes of such processes involves the construction of appropriate macroscopic (homogenized or effective) models. This chapter presents a few examples and gives a brief review of the literature on numerical techniques for solving these problems.

## 1.1 Multiscale Problems

Heterogeneous micro-structures on many non-separable scales and high contrast in physical properties of the constituents are key features for the superior behaviour of modern composite and multi-functional materials. However, these features cause major difficulties for their computer simulation. The resolution of all characteristic length scales is prohibitively expensive while the naive disregard of relevant microscopic information leads to questionable results, even on macroscopic scales of interest.

Homogenization methods try to remedy this dilemma. They account for the relevant microscopic information in a hierarchical, concurrent and adaptive fashion so that a reliable simulation of multiscale problems eventually becomes feasible in state-of-the-art computing environments. This book concerns the design of the related numerical algorithms and, equally important, the mathematics behind them to foresee and assess their reliability and efficiency in engineering and scientific applications.

Among the possible applications of the methods presented in this book is the mechanical analysis of multiphase materials such as composite and multifunctional materials. The manipulation of characteristics and relative volumes of its constituents allows one to equip engineered multiphase materials with some targeted portfolio of physical properties (e.g. light-weight, stiffness, strong electric and mag-

netic order, energy conversion). The development of novel multifunctional materials for the next generation of performance-tailored structures requires the topological optimisation of the micro-structures and, hence, understanding of how certain material properties (conductivity, permeability, etc.) depend on controllable variables (thermal conductivities of the constituents, relative volumes, particles shapes, coating and size). Transport processes in porous media, e.g. groundwater flow in unsaturated soils [59, 64], share the previous challenges in that the occurring permeabilities and hydraulic conductivities have rapidly changing features due to different types of soil, microscopic inclusions in the rock or porous subsurface rock formations. Any meaningful numerical simulation of relevant physical effects has to account for these highly heterogeneous fine scale structures in the whole computational domain. If pore scale effects become relevant or if domains spread over kilometers, the computational load easily exceeds computer capacity when standard finite element or finite volume methods are used.

## 1.2 A brief history of multiscale methods

The common starting point for the development of analytical and numerical techniques for multiscale problems is the steady state heterogeneous diffusion equation

$$-\nabla \cdot (A\nabla u) = f, \tag{1.1}$$

where $A$ is a positive definite, rapidly varying, diffusion matrix and $f$ is a given source. The diffusion matrix models e.g. heat or electrical conductivity in a composite material, permeability in the subsurface of the earth, or, in a vector valued setting, elastic properties of a composite material. Even for this seemingly simple model problem, direct numerical simulation, e.g. by the finite element method or the finite difference method, is challenging because the spatial variations of the diffusion matrix need to be well resolved by the computational mesh to achieve accurate solutions. This problem, which we study more closely in Chapter 2, has lead to the development of many numerical algorithms, often refereed to as multiscale methods, starting in the 1980's and 1990's.

In the particular case of periodicity and scale separation ($A(x) = A_\varepsilon(x) = A_1(x/\varepsilon)$ for some small period $\varepsilon > 0$ and some 1-periodic coefficient $A_1$) has been studied extensively using the theory of homogenization. For periodic data the solution to equation (1.1), as $\varepsilon \to 0$, solves a similar equation with a constant (effective) diffusion matrix $A_0$. The construction of $A_0$ involves the solution of a local (on the $\varepsilon$-scale) elliptic equation of similar form as the original one. The idea of solving local problems to compute an effective representation of the rapidly varying data (or the whole differential operator) is fundamental and underlies all numerical multiscale methods. Homogenization theory has directly inspired the development of the multiscale finite element method (MsFEM) [38, 22] and the heterogeneous

multiscale method (HMM) [20], and is used to analyze the accuracy of these methods. This development started in the 1990's and is still a very active research field.

An alternative framework, also developed during the 1990's, is the variational multiscale method [40] which uses a decomposition of the trial and test space of the weak form, using a (coarse scale) finite element interpolant. The full space is decomposed into a (coarse scale) finite element space and a (fine scale) remainder space, defined as the kernel of the interpolant. The fine scale effects are incorporated in the coarse scale equation by solving diffusion problems in the remainder space. Originally these fine scale problems where approximated by analytical techniques but later they where solved numerically on vertex patches [45, 44, 39, 52]. In order for this technique to be numerically useful the solutions to the fine scale problems, stated in the remainder space, need to decay exponentially for localized right hand side data. Numerical evidence of this can be seen already in [45] but no theoretical justification was given at the time and there where therefore no way to guarantee accuracy.

The question whether there are stable and accurate methods beyond the strong structural assumptions of analytical homogenization regarding scale separation or even periodicity remained open for a long time. Only recently, the existence of an optimal approximation of the low-regularity solution space by some arbitrarily coarse generalized finite element space (that represents the homogenized problem) was shown in [6] and [33]. However, the constructions therein include prohibitively expensive global solutions of the full fine scale problem or the solution of more involved eigenvalue problems. An efficient and feasible construction, solely based on the solution of localized microscopic cell problems, was first given and rigorously justified in [47] in 2014 and later optimized in [35], generalized in [34, 36] and reinterpreted in [41]. This approach builds upon the variational multiscale method and its orthogonal decomposition of coarse and fine scales as introduced in the earlier works [45, 44, 39, 52]. By the new multiscale method, often referred to as Localized Orthogonal Decomposition (LOD), these earlier contributions are given (with small modifications) a theoretical justification. The derivation, analysis, and application of LOD is the topic of this book. Although it is a fairly recent development, the methodology already inspired numerous new approaches. This field of research is currently very active with applications far beyond classical homogenization problems. Among the latest developments are [54, 55, 42, 37, 53], to mention only a few.

# Chapter 2
# Numerical Analyst's Review of Elliptic Homogenization

This chapter provides a brief illustration of homogenization problems and their treatment in analysis and numerics. The restriction to one spatial dimension allows fairly explicit and transparent proofs without any advanced arguments. We will take the unconventional perspective of a numerical analyst who is interested in the approximability of solutions and quantitative error estimates rather than qualitative limits for infinitesimal small parameters.

## 2.1 Oscillatory diffusion problems and pre-asymptotic effects

For the illustration of the critical scaling effects that motivate this book, we shall consider the simplest possible model problem, that is, a one-dimensional diffusion problem in a periodic laminate,

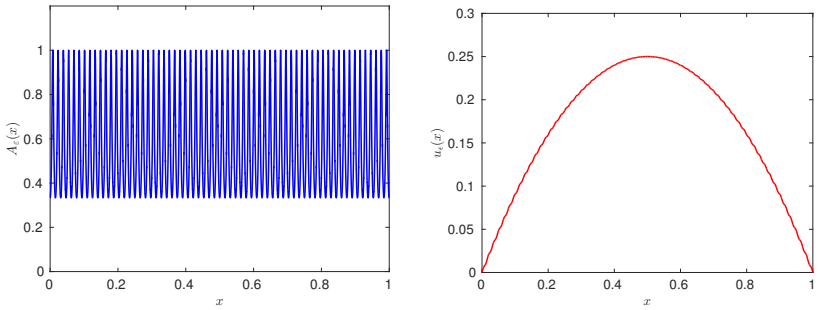
$$-\frac{d}{dx}\left(A_\varepsilon(x)\frac{d}{dx}u_\varepsilon(x)\right) = f(x) \quad \text{in } (0,1), \tag{2.1a}$$

$$u_\varepsilon(0) = u_\varepsilon(1) = 0, \tag{2.1b}$$

with some smooth forcing term $f$ and a uniformly positive, smooth, periodic diffusion coefficient $A_\varepsilon$ with some small parameter $\varepsilon > 0$ that reflects its period length. The problem admits a unique solution $u_\varepsilon$ in the Sobolev space $H^1_0(0,1)$ of square integrable functions with square integrable weak derivative and with vanishing boundary values (in the sense of traces). The unique solution $u_\varepsilon$ is as well characterized by the variational formulation

$$\int_0^1 A_\varepsilon u_\varepsilon' v' \,\mathrm{d}x = \int_0^1 f v \,\mathrm{d}x \quad \text{for all } v \in H^1_0(0,1). \tag{2.2}$$

We shall study the particular instance of problem (2.2) where data $f \equiv 1$ and cofficient $A_\varepsilon$ is given by




(a) Highly oscillatory diffusivity $A_\varepsilon$ defined in (2.3) for $\varepsilon = 2^{-6}$.

(b) Corresponding solution $u_\varepsilon$ of (2.2) for $\varepsilon = 2^{-6}$.

**Fig. 2.1** Illustration of model problem (2.1) for $\varepsilon = 2^{-6}$.

$$A_\varepsilon(x) := \left(2 + \cos\left(2\pi\frac{x}{\varepsilon}\right)\right)^{-1} \tag{2.3}$$

for some small parameter $\varepsilon > 0$ such that $\varepsilon^{-1} \in \mathbb{N}$; cf. Figure 2.1a. In this one-dimensional setting, the corresponding unique solution

$$u_\varepsilon = x - x^2 + \varepsilon\left(\frac{1}{4\pi}\sin(2\pi\frac{x}{\varepsilon}) - \frac{1}{2\pi}x\sin(2\pi\frac{x}{\varepsilon}) - \frac{\varepsilon}{4\pi^2}\cos(2\pi\frac{x}{\varepsilon}) + \frac{\varepsilon}{4\pi^2}\right) \tag{2.4}$$

of (2.1) is easily computed and allows us to study the performance of numerical techniques.

The numerical solution of second order elliptic partial differential equations in variational form is very well established. Nowadays the most popular scheme is the Galerkin finite element method. For symmetric problems such as our model problem, the Galerkin method seeks the best approximation $u_{\varepsilon,h}$ of $u_\varepsilon$ (with respect to the scalar product on the left-hand side of (2.2)) within some finite-dimensional subspace $V_h \subset H_0^1(0,1)$. The simplest choice is to use conforming first-order finite elements ($P_1$-FEM) on a uniform mesh

$$\mathcal{T}_h := \{[jh, (j+1)h] \,|\, j = 0, \ldots, 1/h\}$$

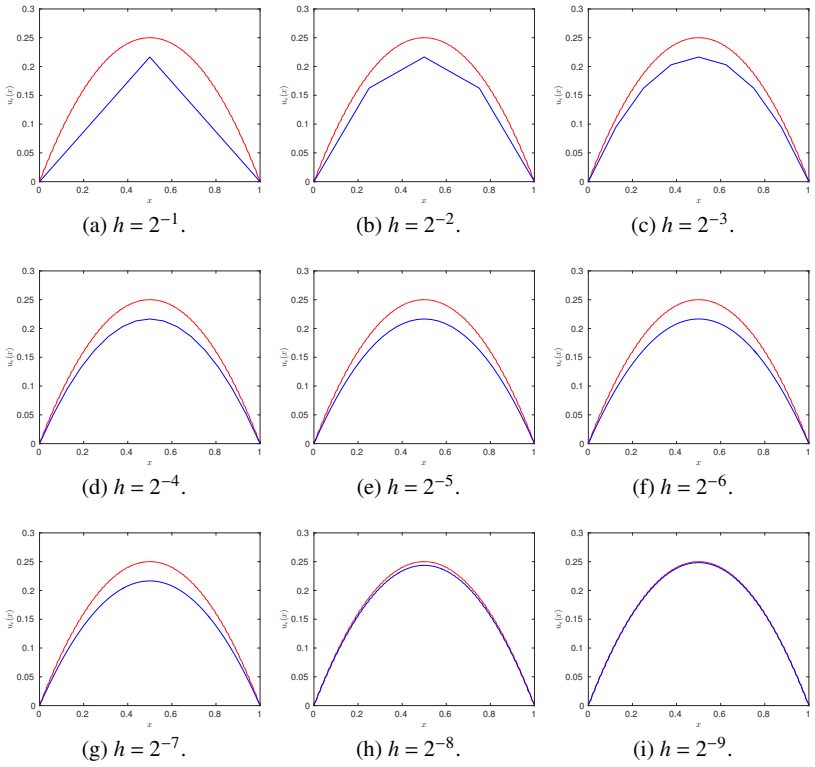
of the unit interval with mesh-size parameter $0 < h < 1$ (such that $h^{-1} \in \mathbb{N}$). In this case, the approximation space reads

$$V_h = \{w \in C^0(0,1) \,|\, \forall T \in \mathcal{T}_h,\ w|_T \text{ is affine and } w(0) = w(1) = 0\}. \tag{2.5}$$

The finite element approximation $u_{\varepsilon,h} \in V_h$ is uniquely characterized by the discrete variational problem

$$\int_0^1 A_\varepsilon u'_{\varepsilon,h} v'_h \,\mathrm{d}x = \int_0^1 f v_h \,\mathrm{d}x \quad \text{for all } v_h \in V_h. \tag{2.6}$$

By a choice of basis of $V_h$, this problem may be rephrased as a system of linear algebraic equations in the coefficients of a suitable basis representation of $u_{\varepsilon,h}$.



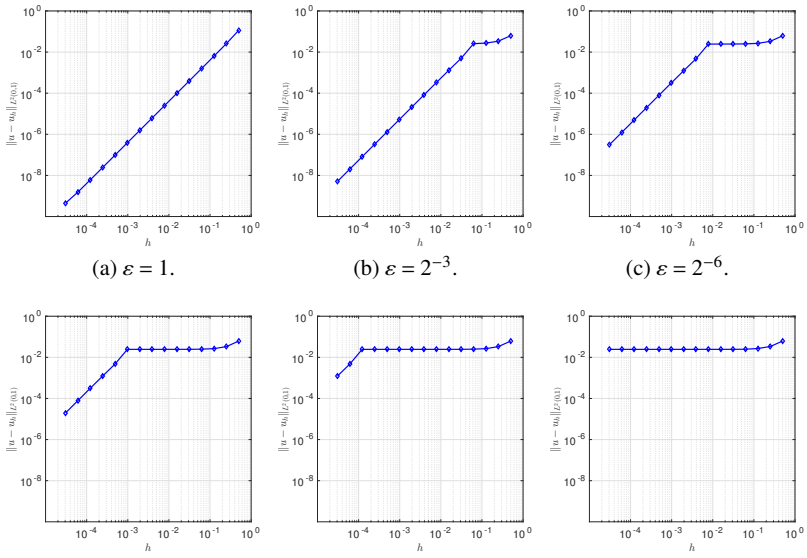
(a) $h = 2^{-1}$. (b) $h = 2^{-2}$. (c) $h = 2^{-3}$.

(d) $h = 2^{-4}$. (e) $h = 2^{-5}$. (f) $h = 2^{-6}$.

(g) $h = 2^{-7}$. (h) $h = 2^{-8}$. (i) $h = 2^{-9}$.

**Fig. 2.2** Finite element approximation of model problem (2.1) for $\varepsilon = 2^{-6}$.

We shall study the performance of this method for several choices of the modeling parameter $\varepsilon$ (assumed to be given here) and the mesh size parameter $h$ (to be chosen). Figure 2.2 depicts the finite element approximation on different scales of numerical resolution $h$ for fixed $\varepsilon = 2^{-6}$. The FE solutions show very different behavior in different regimes of numerical resolution. In the case of under-resolution $h \gtrsim \varepsilon$, FEM is not capable of capturing the solution, neither its microscopic oscillations nor its macroscopic behavior. The FEM solution rather seems to converge to some other function. (Guess which one!) This regime is called pre-asymptotic regime. Only if $h$ is sufficiently small, i.e., $h \lesssim \varepsilon$, the method suddenly switches to the expected asymptotic behavior of quadratic convergence (in the space of square integrable functions $L^2(0,1)$). Figure 2.3 shows that the sharp phase transition between the pre-asymptotic and asymptotic regime is truly linked to the scale $h \approx \varepsilon$. Thus, the performance of this FEM (which is representative for all standard methods) suffers critically from very small microstructures represented by the parameter $\varepsilon$. In many



(a) $\varepsilon = 1$. (b) $\varepsilon = 2^{-3}$. (c) $\varepsilon = 2^{-6}$.

(d) $\varepsilon = 2^{-9}$. (e) $\varepsilon = 2^{-12}$. (f) $\varepsilon = 2^{-15}$.

**Fig. 2.3** Finite element approximation of model problem (2.1): $L^2$-error vs. mesh size $h$ for several values of the diffusion parameter $\varepsilon$.

relevant multi-dimensional applications, the fine scales (represented by $\varepsilon$ here) are so small that this asymptotic regime is never reached, even on large computers. The aim of this book is to present advanced numerical techniques to reduce such crucial scale-dependent pre-asymptotic effects in finite element and related methods.

## 2.2 Effective coefficient and periodic homogenization

Classical homogenization is a tool of mathematical modeling that seeks a simplified model that is able to capture the macroscopic responses of the problem. Note that the solution $u_\varepsilon$ explicitly given in (2.4) is composed of some macroscopic ($\varepsilon$-independent) part

$$u_0 = x - x^2$$

and some microscopic (highly oscillatory and small $L^2$-norm) remainder

$$u_\varepsilon - u_0 = \varepsilon\left(\frac{1}{4\pi}\sin(2\pi\frac{x}{\varepsilon}) - \frac{1}{2\pi}x\sin(2\pi\frac{x}{\varepsilon}) - \frac{\varepsilon}{4\pi^2}\cos(2\pi\frac{x}{\varepsilon}) + \frac{\varepsilon}{4\pi^2}\right)$$

that tends to zero (in $L^2$) with $\varepsilon$. In other words,

$$u_\varepsilon \to u_0 \text{ strongly in } L^2(0,1) \text{ as } \varepsilon \to 0, \tag{2.7}$$

whereas the sequence is only bounded in $H^1(0,1)$ but not strongly convergent as oscillations get faster and faster. Moreover, one observes that $u_0$ is the solution of the Poisson problem

$$-\frac{d}{dx}\left(A_0 \frac{d}{dx} u_0(x)\right) = f(x) \quad \text{in } (0,1), \tag{2.8a}$$

$$u(0) = u(1) = 0, \tag{2.8b}$$

with some positive constant $A_0$ – the so-called effective or homogenized coefficient. Is this just by coincidence for the particular $f \equiv 1$ or is there is some general mechanism behind? Let $A_1 \in L^\infty_\#(0,1)$ be a uniformly positive, 1-periodic coefficient and define $\varepsilon$-periodic coefficients $A_\varepsilon$ by $A_\varepsilon(x) := A_1(\frac{x}{\varepsilon})$. The main question of periodic homogenization then reads: Is there an effective (constant) coefficient $A_0 > 0$ such that the solutions $u_\varepsilon$ of (2.1) converge to the solution $u_0$ of problem (2.8) uniformly with respect to $f \in L^2(0,1)$? If yes, Problem (2.8) is denoted homogenized (or effective) problem.[1] In addition to this theoretical question, there is the equally important question of computability of the effective coefficient.

For the simple model problem of this section, both questions can be answered in a positive and satisfying way in one stroke. For the time being, we assume that $A_0 > 0$ is some real number and that the tentative macroscopic part $u_0$ solves (2.8). We shall have a look at the $L^2$-error between $u_\varepsilon$ and $u_0$. We restrict ourselves tacitly to values of $\varepsilon$ that are related to integer frequencies, i.e., $\varepsilon^{-1} \in \mathbb{N}$.

Since the error $(u_\varepsilon - u_0) \in H^1_0(0,1) \subset L^2(0,1)$, there exists a unique $z \in H^1_0(0,1)$ such that

$$\int_0^1 A_0 z' w' \,\mathrm{d}x = \int_0^1 (u_\varepsilon - u_0) w \,\mathrm{d}x \quad \text{for all } w \in H^1_0(0,1).$$

The choice $w = u_\varepsilon - u_0$ as a test function yields that

$$\|u_\varepsilon - u_0\|^2_{L^2(0,1)} = \int_0^1 A_0 z' (u_\varepsilon - u_0)' \,\mathrm{d}x.$$

This relation between the $L^2$ error and the variational form is known as Aubin-Nitsche duality trick [5].

The solutions $u_0$ and $u_\varepsilon$ are linked by the equality of fluxes

$$A_\varepsilon u_\varepsilon' = A_0 u_0' \quad \text{in } H^{-1}(0,1),$$

that is,

$$\int_0^1 A_\varepsilon u_\varepsilon' v' \,\mathrm{d}x = \int_0^1 A_0 u_0' v' \,\mathrm{d}x \qquad \text{for all } v \in H^1_0(0,1).$$

[1] Keep in mind that, in general, the structure and the type of the homogenized problem can be very different from the structure of the original problem. It will not be the case in the present setting though.

This and some algebraic manipulations lead to

$$\|u_\varepsilon - u_0\|^2_{L^2(0,1)} = \int_0^1 A_0 z' \frac{A_0 - A_\varepsilon}{A_0 A_\varepsilon} A_\varepsilon u'_\varepsilon \,\mathrm{d}x = \sum_{T\in\mathcal{T}_\varepsilon} \int_T A_0 z' \frac{A_0 - A_\varepsilon}{A_0 A_\varepsilon} A_\varepsilon u'_\varepsilon \,\mathrm{d}x.$$

We have divided the integral into integrals over the periods

$$T \in \mathcal{T}_\varepsilon := \{[(j-1)\varepsilon, j\varepsilon] \mid j = 1, 2, \ldots, N\}$$

of the coefficient $A_\varepsilon$. Subtracting and adding mean values of the fluxes $A_0 z'$ and $A_\varepsilon u'_\varepsilon$ on these periods leads to

$$\begin{aligned}\|u_\varepsilon - u_0\|^2_{L^2(0,1)} = \sum_{T\in\mathcal{T}_\varepsilon} \Bigg( &\int_T \Big(A_0 z' - \fint_T A_0 z' \,\mathrm{d}x\Big) \frac{A_0 - A_\varepsilon}{A_0 A_\varepsilon} A_\varepsilon u'_\varepsilon \,\mathrm{d}x \\ &+ \fint_T A_0 z' \,\mathrm{d}x \int_T \frac{A_0 - A_\varepsilon}{A_0 A_\varepsilon} \Big(A_\varepsilon u'_\varepsilon - \fint_T A_\varepsilon u'_\varepsilon \,\mathrm{d}x\Big) \mathrm{d}x \\ &+ \fint_{T\mathrm{d}x} A_0 z' \,\mathrm{d}x \int_T \frac{A_0 - A_\varepsilon}{A_0 A_\varepsilon} \,\mathrm{d}x \fint_T A_\varepsilon u'_\varepsilon \,\mathrm{d}x \Bigg).\end{aligned}$$

This error representation allows us, by several applications of the Cauchy-Schwarz and the Poincaré inequality (see Theorems A.22,A.24), to estimate the first two summands by multiples of $\varepsilon$,

$$\begin{aligned}\|u_\varepsilon - u_0\|^2_{L^2(0,1)} \le \sum_{T\in\mathcal{T}_\varepsilon} \Big( &\fint_T A_0 z' \,\mathrm{d}x \int_T \frac{A_0 - A_\varepsilon}{A_0 A_\varepsilon} \,\mathrm{d}x \fint_T A_\varepsilon u'_\varepsilon \,\mathrm{d}x \Big) \\ &+ \varepsilon\pi^{-1} \left\| \frac{A_0 - A_\varepsilon}{A_0 A_\varepsilon} \right\|_{L^\infty(T)} \Big( \|A_0 z''\|_{L^2(0,1)} \|A_\varepsilon u'_\varepsilon\|_{L^2(0,1)} \\ &\qquad + \|A_0 z'\|_{L^2(0,1)} \|(A_\varepsilon u'_\varepsilon)'\|_{L^2(0,1)} \Big).\end{aligned}$$

The first term on the right-hand side tends to zero (as $\varepsilon \to 0$) if and only if it is actually zero. This is achieved by the unique choice

$$A_0 := \left( \fint_T A_\varepsilon^{-1} \,\mathrm{d}x \right)^{-1} = \left( \int_0^1 A_1^{-1} \,\mathrm{d}x \right)^{-1} = \left( \int_0^1 A_\varepsilon^{-1} \,\mathrm{d}x \right)^{-1}.$$

Since

$$\begin{aligned} -A_0 z'' &= u_\varepsilon - u_0 \quad \text{in the sense of } L^2(0,1), \\ \|A_0 z'\|_{L^2(0,1)} &\le \pi^{-1} \|u_\varepsilon - u_0\|_{L^2(0,1)}, \\ -(A_\varepsilon u'_\varepsilon)' &= f \quad \text{in the sense of } L^2(0,1), \text{ and} \\ \|A_\varepsilon u'_\varepsilon\|_{L^2(0,1)} &\le \pi^{-1} \sqrt{\beta/\alpha} \|f\|_{L^2(0,1)}, \end{aligned}$$

with $\alpha := \operatorname{ess\,inf}_{0<x<1} A_1(x) > 0$ and $\beta := \operatorname{ess\,sup}_{0<x<1} A_1(x) \ge \alpha$, we finally get

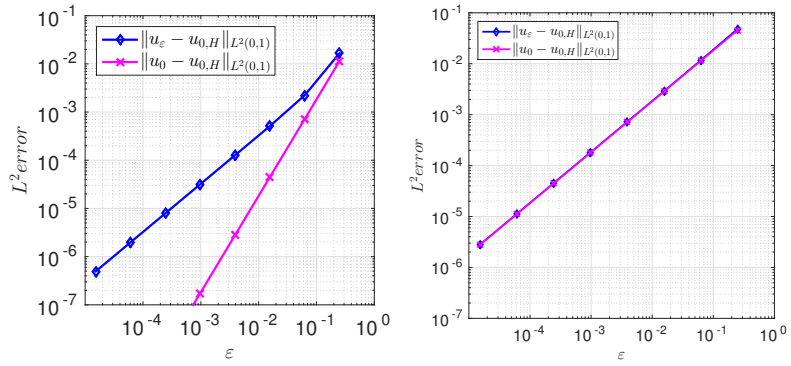
$$\|u_\varepsilon - u_0\|_{L^2(0,1)} \leq \varepsilon \frac{2}{\alpha\pi^2}\left(1 + \sqrt{\beta/\alpha}\right)\|f\|_{L^2(0,1)}. \tag{2.9}$$

The previous calculations show that the desired effective coefficient $A_0 > 0$ exists indeed. Moreover, the corresponding macroscopic solution approximates the true solution $u_\varepsilon$ with an accuracy proportional to $\varepsilon$ in $L^2(0,1)$. Note that the effective coefficient is the harmonic mean of $A_\varepsilon$ rather than the simple average. This observation answers the above question of periodic homogenzation in a quantitative way. As $A_0 \in [\alpha,\beta]$ is easily computed by (numerical) quadrature to high accuracy in this simple model problem, we also have access to reliable and accurate numerical approximations of $u_0$ by standard schemes such as the P1-FEM introduced in the previous section. Since the coefficient is a global constant, the Aubin-Nitsche duality trick, Céa's lemma and standard interpolation error estimates show that for any $f \in L^2(0,1)$ the Galerkin finite element approximation $u_{0,H} \in V_H$ of $u_0$ computed on a uniform mesh $\mathcal{T}_H$ of width $H > 0$ satisfies

$$\|u_0 - u_{0,H}\|_{L^2(0,1)} \leq \frac{1}{A_0\pi^2} H^2 \|f\|_{L^2(0,1)}.$$

Hence, for a given fixed value of $\varepsilon$, a finite element computation on the discretization scale $H = \sqrt{\varepsilon}$ would yield an approximation of the macroscopic part of $u_\varepsilon$ on the same order of accuracy as $u_0$ itself. In practical applications, it may still be too expensive to compute on the scale $\sqrt{\varepsilon}$. In this case, the numerical discretization parameter $H$ should be chosen according to the available computational resources, accepting that the simulation commits some larger but still acceptable error.

Given the data of Section 2.1, the errors $\|u_0 - u_{0,H}\|_{L^2(0,1)}$ and $\|u_\varepsilon - u_{0,H}\|_{L^2(0,1)}$ are depicted in Figure 2.4 for several values of $\varepsilon$ and two choices of the coupling between $H$ and $\varepsilon$ to confirm the previous discussion of the theoretical results. These



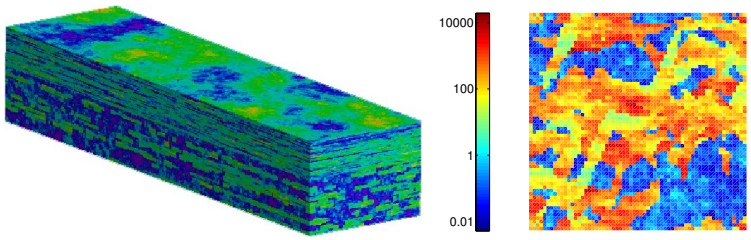

(a) $H = \varepsilon$. (b) $H = \sqrt{\varepsilon}$.

**Fig. 2.4** Finite element approximation of homogenized model problem (2.8): $L^2$-error vs. diffusion parameter $\varepsilon$ for two choices of the mesh size $H = \varepsilon$ (a) and $H = \sqrt{\varepsilon}$ (b).

concepts of periodic homogenization may be generalized to higher space dimensions [9] but the characterization of the effective coefficient is usually not explicit anymore. The approach is still constructive under the assumption of periodicity in the sense that the effective coefficient is related to an average over a function that solves some monoscale PDE. Among the most popular constructive analytical tools are the energy method (or method of oscillating test functions) [49, 50, 16, 17], two-scale convergence [51, 2] and periodic unfolding [15, 14]. The numerical incarnation of these approaches is the Heterogeneous Multiscale Method (HMM) [20, 21, 1]. In this book, we will not follow this path but aim to treat more general problems beyond periodicity. We will comment on the periodic case in Chapter 5 where we will recover some central results of periodic homogenization from the more general error bounds for numerical homogenization from Chapter 4.

## 2.3 Numerical homogenization of arbitrary rough coefficients

The mathematical theory of homogenization can treat very general non-periodic coefficients in the framework of $G$- or $H$-convergence [49, 62, 18]. However, apart from being non-constructive in many cases, homogenization in the classical analytical sense considers a sequence of operators $-\mathrm{div}(A_\varepsilon \nabla \cdot)$ and aims to characterize the limit as $\varepsilon$ tends to zero. In many realistic applications, e.g. in geophysics, (cf. Figure 2.5), such a sequence of models can hardly be identified or may not be available at all.




(a) 3$d$ data. (b) 2$d$ slice with scale.

**Fig. 2.5** Strongly heterogeneous data from SPE10 benchmark; see `www.spe.org/web/csp/`.

That is why we are interested in the computation of effective representations of very rough unstructured coefficients. In the context of our model problem and the concept of a weak solution, $L^\infty(0,1)$ is the most general space of possible coefficients with the additional requirement of uniform positivity. For this section, we assume that $A \in L^\infty(0,1)$ and that there exists constants $\alpha,\beta$ such that

$$0 < \alpha \leq \text{ess } inf_{0<x<1} A(x) \leq \text{ess } sup_{0<x<1} A(x) \leq \beta < \infty. \tag{2.10}$$

The set of admissible coefficients will be denoted

$$\mathcal{M}([0,1],\alpha,\beta) := \{A \in L^\infty(0,1) |\ A \text{ satisfies (2.10)}\}. \tag{2.11}$$

Note that $A$ is fairly free to vary within the bounds $\alpha$ and $\beta$ and that we do not assume any frequencies of variation or smoothness.

Consider the model problem (2.1) with $A_\varepsilon$ replaced by such a general $A \in \mathcal{M}([0,1],\alpha,\beta)$ and some forcing term $f \in L^2(0,1)$. The problem admits a unique solution $u$ in the Sobolev space $H^1_0(0,1)$ that is as well characterized by the variational formulation

$$\int_0^1 Au'v'\,\mathrm{d}x = \int_0^1 fv\,\mathrm{d}x \quad \text{for all } v \in H^1_0(0,1). \tag{2.12}$$

The aim of this section is to revisit the derivation of the previous section to see if the assumption of periodicity was really essential. We consider a general mesh of $N+2$ points

$$0 := x_0 < x_1 < x_2 < \ldots < x_{N+1} =: 1.$$

We shall introduce the corresponding finite element mesh

$$\mathcal{T}_H := \{T = [x_j, x_{j+1}] |\ j = 0,1,\ldots,N\}.$$

Although the mesh may be fairly general in terms of its local mesh size, we shall refer only to the global mesh size parameter $H := \max_{T\in\mathcal{T}_H} |T|$. The discretization parameter $H$ is related to the target scale of interest or observation and we are interested to compute an effective coefficient that represents the solution of the model problem on all scales larger or equal to the local mesh size. This is a much more pragmatic question than asking for a constant coefficient that represents the limit for some (possibly artificial) infinitesimal small parameter. More precisely, we are looking for an admissible coefficient $A_H$ that is piecewise constant with respect to the mesh $\mathcal{T}_H$, i.e.,

$$A_H \in \mathcal{M}(\mathcal{T}_H,\alpha,\beta) := \{B \in \mathcal{M}([0,1],\alpha,\beta) \mid \forall T \in \mathcal{T}_H : B|_T \text{ is constant}\}. \tag{2.13}$$

We will follow closely the derivation of the previous section. Revisiting the arguments shows that periodicity was solely used to argue that $A_0$ is a global constant whereas the other arguments carry over to the present setting when $\epsilon$ is replaced by $H$. Let us recall some details.

For the time being, we assume that $A_H \in \mathcal{M}(\mathcal{T}_H,\alpha,\beta)$ and that the tentative macroscopic part $u_H$ solves

$$\int_0^1 A_H u_H' v'\,\mathrm{d}x = \int_0^1 fv\,\mathrm{d}x \quad \text{for all } v \in H^1_0(0,1). \tag{2.14}$$

We shall have a look at the $L^2$-error between $u$ and $u_H$. Since the error $(u-u_H) \in H^1_0(0,1) \subset L^2(0,1)$, there exists a unique $z \in H^1_0(0,1)$ such that

$$\int_0^1 A_H z' w' \,\mathrm{d}x = \int_0^1 (u - u_H) w \,\mathrm{d}x \quad \text{for all } w \in H_0^1(0,1).$$

The choice $w = (u - u_H)$ as a test function yields that

$$\|u - u_H\|_{L^2(0,1)}^2 = \int_0^1 A_H z' (u - u_H)' \,\mathrm{d}x$$

and the same arguments as in the previous section lead to

$$\begin{aligned}\|u - u_H\|_{L^2(0,1)}^2 \leq \sum_{T \in \mathcal{T}_H} \Big( \fint_T A_H z' \,\mathrm{d}x \int_T \frac{A_H - A}{A_H A} \,\mathrm{d}x \fint_T A u' \,\mathrm{d}x \Big) \\ + \pi^{-1} H \left\| \frac{A_H - A}{A_H A} \right\|_{L^\infty(0,1)} \big( \|(A_H z')'\|_{L^2(0,1)} \|A u'\|_{L^2(0,1)} \\ + \|A_H z'\|_{L^2(0,1)} \|(A u')'\|_{L^2(0,1)} \big)\end{aligned}$$

The first term on the right-hand side is eliminated by the unique choice

$$A_H|_T := \left( \fint_T A^{-1} \,\mathrm{d}x \right)^{-1} \tag{2.15}$$

for any $T \in \mathcal{T}_H$. Since

$$\begin{aligned} -(A_H z')' &= u - u_H \quad \text{in the sense of } L^2(0,1), \\ \|A_H z'\|_{L^2(0,1)} &\leq \pi^{-1} \sqrt{\beta/\alpha} \|u - u_H\|_{L^2(0,1)}, \\ -(A u')' &= f \quad \text{in the sense of } L^2(0,1), \text{ and} \\ \|A u'\|_{L^2(0,1)} &\leq \pi^{-1} \sqrt{\beta/\alpha} \|f\|_{L^2(0,1)}, \end{aligned}$$

we finally get

$$\|u - u_H\|_{L^2(0,1)} \leq H \frac{4}{\alpha \pi^2} \sqrt{\beta/\alpha} \|f\|_{L^2(0,1)}. \tag{2.16}$$

Homogenization in the classical sense of a global constant coefficient is, hence, achieved whenever one is able to find a mesh $\mathcal{T}_H$ such that the numbers $A_H|_T$ coincide for all $T \in \mathcal{T}_H$. In the periodic case, this happens for any equidistant mesh that is in resonance with the frequency of the coefficient (i.e., $H$ is an integer multiple of $\varepsilon$). In the general case, this cannot be expected or is very hard to achieve and $A_H$ is only $\mathcal{T}_H$-piecewise constant.

Note that $u_H$ may now be replaced with its Galerkin projection onto the P1-FE space on the same mesh $\mathcal{T}_H$ without any harm. Similar as in the previous section, we may also consider its Galerkin approximation on an even coarser mesh of width $\sqrt{H}$. However, on this scale, $A_H$ is not a constant in each element and such an approach may suffer from possible oscillations of $A_H$ on the scale $H$.

## 2.4 A different approach to numerical homogenization

Another approach to the numerical homogenization of (2.12) (or (1.1) in general) is that of the approximation of the solution space $H^1_0(0,1)$ by a finite-dimensional space as in Section 2.1 but without undesired scale-dependent pre-asymptotic effects. In what follows, we shall illustrate that this is possible.

Given positive constants $\beta \geq \alpha > 0$, some admissible coefficient $A \in \mathcal{M}([0,1],\alpha,\beta)$ and some forcing term $f \in L^2(0,1)$, we wish to approximate the unique function $u \in H^1_0(D)$ that satisfies the variational problem (2.12).

While in the two previous sections the aim was to approximate $A$ by some effective coefficient $A_H$ and some corresponding effective problem that is easily solved by means of standard finite elements, we are now heading for a generalized finite element method that encodes the unresolvable fine-scale information in its shape functions. This is a discrete approach in the sense that the resulting effective problem will be a discrete one.

As in Section 2.3, we consider a fairly general mesh $\mathcal{T}_H := \{T = [x_j, x_{j+1}] \,|\, j = 0,1,\ldots,N\}$ represented by $N+2$ mesh points

$$0 := x_0 < x_1 < x_2 < \ldots < x_{N+1} =: 1.$$

The global mesh size parameter is $H := \max_{T\in\mathcal{T}_H} |T|$ and we use a capital letter to emphasize that $H$ may be arbitrarily coarse and possibly larger than characteristic length scales of the coefficient $A$, if any.

Our goal is to design a finite-dimensional space $\tilde{V}_H \subset V := H^1_0(0,1)$ (linked to the mesh $\mathcal{T}_H$) with a local basis and high-approximation properties regardless of variations of $A$. In particular we want the space to be accurate in the pre-asymptotic regime of the standard FEM observed in Section 2.1. Our starting point will be the standard finite element space

$$V_H = \{v_H \in H^1_0(0,1) \,|\, \forall T \in \mathcal{T}_H : v_H|_T \in \mathbb{P}_1\} \tag{2.17}$$

of continuous $\mathcal{T}_H$-piecewise affine functions that vanish at the boundary of the unit interval previously defined in (2.5). We shall also characterize the functions of the solution space $V = H^1_0(0,1)$ that are not well captured by $V_H$. Define

$$W_H := \{w \in H^1_0(0,1) \,|\, \forall j = 1,\ldots,N_H : w(x_j) = 0\}. \tag{2.18}$$

We will refer to this space as the fine scale or microscopic space. Its elements oscillate at frequencies larger than $H^{-1}$. Observe that any function $v \in V$ can be cast in the form $v_H \in V_H$ plus $w_H \in W_H$, where $v_H$ is the nodal interpolation of $v$ at the vertices $x_j$ and $w_H$ is the error of interpolation.[2] In other words,

$$V = V_H \oplus W_H.$$

[2] Recall that point evaluation is well posed for univariate $H^1_0$ functions in the sense of the Sobolev embedding $H^1(0,1) \hookrightarrow C([0,1])$ (cf. Theorem A.19).

This decomposition is orthogonal in $H_0^1(0,1)$, i.e., for any $v_H \in V_H$ and any $w_H \in W_H$ it holds

$$\int_0^1 v_H'(x) w_H'(x)\,\mathrm{d}x = \sum_{T\in\mathcal{T}_H} (v_H|_T)' \int_T w_H'(x)\,\mathrm{d}x = 0.$$

However, the experiment of Section (2.1) clearly indicates that this orthogonality has no impact in the context of the model problem (2.12) which is related to a different scalar product.

Instead, the new approach to numerical homogenization of this section is based on the orthogonalization of this decomposition with respect to the scalar product

$$a(\cdot,\cdot) := \int_0^1 A(\cdot)'(\cdot)'\,\mathrm{d}x$$

induced by the model problem (2.12). Keeping the characterization of fine scales $W_H$ fixed, this orthogonalization characterizes a new coarse space $\tilde{V}_H$ by

$$V = \tilde{V}_H \oplus W_H \quad \text{and} \quad \tilde{V}_H \perp_a W_H.$$

A Galerkin method based on $\tilde{V}_H$ computes the $a$-orthogonal projection $\tilde{u}_H \in \tilde{V}_H$ of the unknown solution $u \in V$ onto $\tilde{V}_H$, i.e., $\tilde{u}_H$ is the unique function in $\tilde{V}_H$ that satisfies

$$a(\tilde{u}_H, \tilde{v}_H) = \int_0^1 f(x)\tilde{v}_H(x)\,\mathrm{d}x \quad \text{for all } \tilde{v}_H \in \tilde{V}_H. \tag{2.19}$$

By Galerkin orthogonality,

$$a(u-\tilde{u}_H,\tilde{v}_H)=0 \ \text{ for all } \tilde{v}_H\in\tilde{V}_H,$$

the error $(u-\tilde{u}_H)\in W_H$ is a fine scale function. Hence, the error of this method vanishes in all mesh points, i.e., $\tilde{u}_H$ interpolates $u$ in the mesh points $x_j$ ($j=0,\dots,N+1$). This, Friedrichs' inequality (cf. Theorems A.23,A.24), Galerkin orthogonality, symmetry of the bilinear form $a$ and the Cauchy-Schwarz inequality yield

$$\begin{aligned}
\|u-\tilde{u}_H\|^2_{L^2(0,1)} &\le \frac{H^2}{\pi^2}\|(u-\tilde{u}_H)'\|^2_{L^2(0,1)}\\
&\le \alpha^{-1}\frac{H^2}{\pi^2}\int_0^1 A(u-\tilde{u}_H)'(u-\tilde{u}_H)'\,\mathrm{d}x\\
&= \alpha^{-1}\frac{H^2}{\pi^2}\int_0^1 Au'(u-\tilde{u}_H)'\,\mathrm{d}x\\
&= \alpha^{-1}\frac{H^2}{\pi^2}\int_0^1 f(u-\tilde{u}_H)\,\mathrm{d}x\\
&\le \alpha^{-1}\frac{H^2}{\pi^2}\|f\|_{L^2(0,1)}\|u-\tilde{u}_H\|_{L^2(0,1)},
\end{aligned}$$

and, hence, the error estimate

$$\|u-\tilde{u}_H\|_{L^2(0,1)}\le\alpha^{-1}\frac{H^2}{\pi^2}\|f\|_{L^2(0,1)}. \tag{2.20}$$

This means that the error of the method is proportional to the discretization parameter $H$ squared, unconditionally for all $H$ and independent of the coefficient $A$. In contrast to the homogenized solution and the standard finite element approximation, the approximation $\tilde{u}_H$ encodes also fine scale information. A truly coarse approximation would be the finite element part of $\tilde{u}_H$, that is, its nodal interpolation $I_H\tilde{u}_H\in V_H$ defined by $I_H\tilde{u}_H(x_j)=\tilde{u}_H(x_j)(=u(x_j))$ for all $j=0,1,\dots,N_H$. Note that $I_H\tilde{u}_H$ still enjoys the favorable error estimate

$$\|u-I_H\tilde{u}_H\|_{L^2(0,1)}\le\alpha^{-1}\frac{H}{\pi}\|f\|_{L^2(0,1)}. \tag{2.21}$$

Another remark concerns the treatment of the right-hand side. Note that the new approach affects not only the discrete differential operator but also the right-hand side because the test functions are modified. Replacing the right-hand side $\tilde{v}_H\mapsto\int_0^1 f\tilde{v}_H\,\mathrm{d}x$ in (2.19) with $\tilde{v}_H\mapsto\int_0^1 fI_H(\tilde{v}_H)\,\mathrm{d}x$ removes this problem and leads to a modified method

$$a(\tilde{u}_H,\tilde{v}_H)=\int_0^1 f(x)I_H\tilde{v}_H(x)\,\mathrm{d}x \ \text{ for all } \tilde{v}_H\in\tilde{V}_H. \tag{2.22}$$

Note that the solutions of (2.19) and (2.22) do not coincide but their $H^1$ distance can be controlled by the term $H\|f\|_{L^2(0,1)}$ so that the error bounds (2.20) and (2.21) remain valid in terms of the rate of convergence $H^2$ and $H$, respectively.

It remains to find a local basis of the space $\tilde{V}_H$ so that the discretization leads to a sparse linear system that can be solved efficiently. Starting from the nodal basis

$$\{\Lambda_j \in V_H \,|\, \Lambda_j(x_i) = \delta_{ij} \text{ for } i,j = 1,2,\dots,N_H - 1\}$$

of $V_H$, the Schmidt-type orthogonalization yields that

$$\tilde{V}_H = \operatorname{span}\{\Lambda_j^{\mathrm{ms}} := \Lambda_j - \phi_j \,|\, j = 1,2,\dots,N_H - 1\},$$

where the correction $\phi_j \in W_H$ is such that

$$a(\phi_j, w) = a(\Lambda_j, w) \quad \text{for all } w \in W_H. \tag{2.23}$$

Since the right hand side vanishes for test functions $w \in W_H$ which do not have support in the elements $T_j, T_{j+1}$ adjacent to $x_j$, (2.23) is equivalent to solve the local problems

$$a(\phi_j|_{T_j}, w) = H^{-1} \int_{T_j} Aw' \,\mathrm{d}x \quad \text{for all } w \in H_0^1(T_j), \tag{2.24}$$

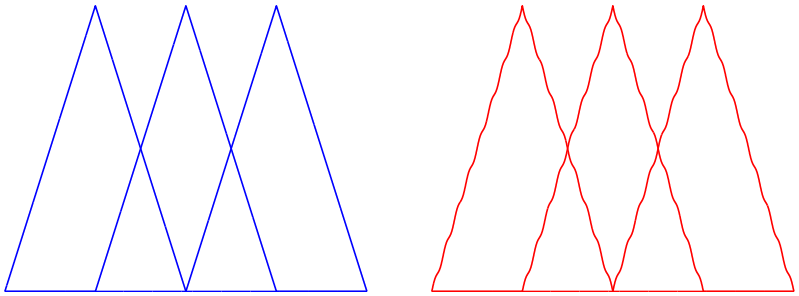
$$a(\phi_j|_{T_{j+1}}, w) = -H^{-1} \int_{T_{j+1}} Aw' \,\mathrm{d}x \quad \text{for all } w \in H_0^1(T_{j+1}). \tag{2.25}$$

This implies that $\operatorname{supp}\Lambda_j^{\mathrm{ms}} \subset \operatorname{supp}\Lambda_j$ for all $j$. Moreover, since $\phi_j|_{T_{j+1}} = -\phi_j|_{T_j}$, the $\phi_j$ can be computed by solving (2.24) for all $T \in \mathcal{T}_H$. The local problems (2.24) are denoted corrector problems. For periodic coefficients and if the mesh size $H$ is an integer multiple of the period (!), only one of these problems for an arbitrary element $T \in \mathcal{T}_H$ needs to be solved. In the one-dimensional case the corrector problems are easily solved analytically by hand. It turns out that for any $j = 1,2,\dots,N_H$,

$$\Lambda_j^{\mathrm{ms}}(x) := \begin{cases} \dfrac{\int_{x_{j-1}}^{x} A^{-1}(s)\,\mathrm{d}s}{\int_{x_{j-1}}^{x_j} A^{-1}(s)\,\mathrm{d}s}, & \text{if } x \in T_j, \\ 1 - \dfrac{\int_{x_j}^{x} A^{-1}(s)\,\mathrm{d}s}{\int_{x_j}^{x_{j+1}} A^{-1}(s)\,\mathrm{d}s}, & \text{if } x \in T_{j+1}, \\ 0, & \text{else.} \end{cases} \tag{2.26}$$

See Figure 2.6 for a visualization of this perturbed nodal basis given the oscillatory coefficient from (2.3) with $\varepsilon = 2^{-5}$. In the literature, this method and its variants are known under several names, e.g., Generalized FEM (GFEM) [7], Variational Multiscale Method [40], Multiscale FEM (MsFEM) [38] or Residual Free Bubbles [12]. The previous derivation is based on the interpretation of [47]; see also [35, 58].

We shall have a closer look at the relation of the current approach with the method of the previous Section 2.3. Recall that the nodal values $\tilde{u}_H(x_j)$ of the Galerkin approximation $\tilde{u}_H = \sum_{j=1}^{N} \tilde{u}_H(x_j)\Lambda_j^{\mathrm{ms}}$ are the unique solution of the system of $N$



(a) Classical nodal basis. (b) Perturbed nodal basis.

**Fig. 2.6** Classical nodal basis on uniform mesh $\mathcal{T}_H$ ($H = .25$) and corrected nodal basis for numerical homogenization (coefficient $A_\varepsilon$ as in (2.3) with $\varepsilon = 2^{-5}$).

linear equations

$$\sum_{k=1}^{N}\left(\int_0^1 A(\Lambda_j^{\mathrm{ms}})'(\Lambda_k^{\mathrm{ms}})'\,\mathrm{d}x\right)\tilde{u}(x_k) = \int_0^1 f\Lambda_j\,\mathrm{d}x, \quad j = 1,\dots,N.$$

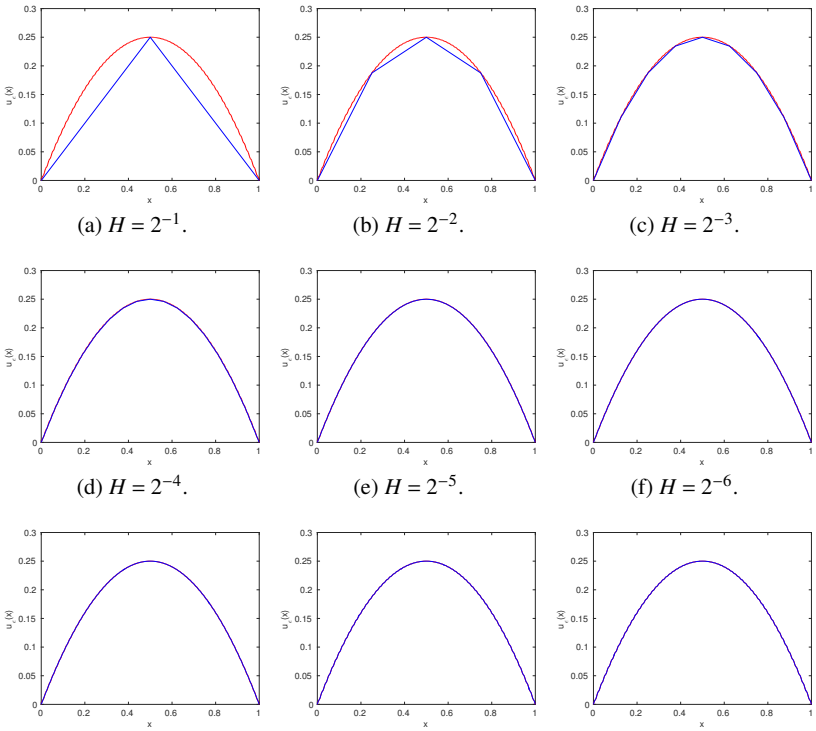
Using the explicit representation of $\Lambda_j^{\mathrm{ms}}$, this system is easily rewritten as

$$\sum_{k=1}^{N}\left(\int_0^1 A_H\Lambda_j'\Lambda_k'\,\mathrm{d}x\right)\tilde{u}(x_k) = \int_0^1 f\Lambda_j\,\mathrm{d}x, \quad j = 1,\dots,N,$$

where $A_H$ is exactly the homogenized coefficient defined in (2.15). This means that the method of this section is equivalent to computing the homogenized coefficient with respect to $\mathcal{T}_H$ as in the previous section followed by a P1-FE approximation of the corresponding homogenized solution characterized by (2.14) on the same mesh $\mathcal{T}_H$.

As in Section 2.1, we shall study the performance of this method for several choices of the modeling parameter $\varepsilon$ (typically given) and the mesh size parameter $H$ (to be chosen). We do not reconstruct fine scale information, so that we expect the error to behave as predicted by (2.21). The Figures 2.7 and 2.8 summarize the results of the numerical homogenization on different scales of numerical resolution $H$.

The previous derivations and numerical results indicate the possible superiority of numerical homogenization over analytical techniques with regard to its applicability beyond periodicity and scale separation. However, we shall warn the reader that all previous derivations – related to numerical and analytical homogenization – hold only in one space dimension. We have used, e.g. , that any $L^2(0,1)$ function is a gradient or that point evaluation for $H^1$ functions is a well-defined and stable operation. The main goal of this book will be to generalize the previous approaches to two- and three-dimensional settings. In this regard, the re-interpretation of nu-



(a) $H = 2^{-1}$. (b) $H = 2^{-2}$. (c) $H = 2^{-3}$.

(d) $H = 2^{-4}$. (e) $H = 2^{-5}$. (f) $H = 2^{-6}$.

(g) $H = 2^{-7}$. (h) $H = 2^{-8}$. (i) $H = 2^{-9}$.

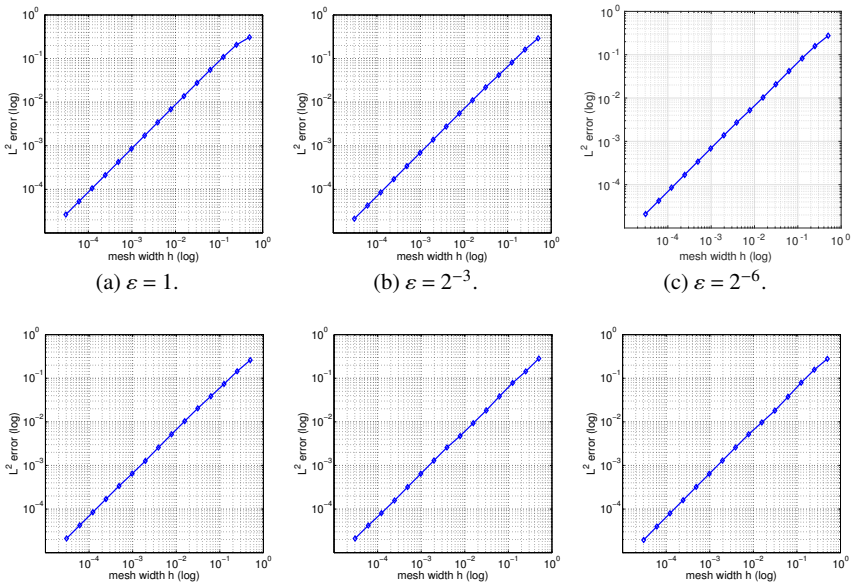
**Fig. 2.7** Numerical homogenization of model problem (2.1) for $\varepsilon = 2^{-6}$.

merical homogenization of this section will be of great value as it allows such a generalization even for $L^\infty$ coefficients. This will be the topic of Chapters 3–5.

## 2.5 The case of random coefficients

In some of the applications mentioned earlier it is very unlike that the coefficient $A$ that represents porosity or permeabilty in geophysical applications is known explicitely. The coefficient is rather the result of measurements that underlie errors or it is the result of measurements combined with inverse modeling. In any case, it is very likely that the data $A$ is uncertain and the question is how uncertainties on the fine scale change the macroscopic responses of the processes.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with set of events $\Omega$, $\sigma$-algebra $\mathcal{F} \subset 2^\Omega$ and probability measure $\mathbb{P}$. Let $\mathbf{A}$ be an $\mathcal{M}([0,1], \alpha, \beta)$-valued random field and let, for the sake of readability, $f \in L^2(0,1)$ be deterministic. Among the possible examples is the random medium, where in each cell of a uniform mesh of width $\varepsilon$ (for some




(a) $\varepsilon = 1$. (b) $\varepsilon = 2^{-3}$. (c) $\varepsilon = 2^{-6}$.

(d) $\varepsilon = 2^{-9}$. (e) $\varepsilon = 2^{-12}$. (f) $\varepsilon = 2^{-15}$.

**Fig. 2.8** Numerical homogenization of model problem (2.1): $L^2$-error of $I_H \tilde{u}_H$ vs. mesh size $H$ for several values of the diffusion parameter $\varepsilon$. No pre-asymptotic effects observable.

small parameter $\varepsilon$), the conductivity is an independent random number identically distributed in the interval $[\alpha,\beta]$.

Consider the model problem

$$\left\{ \begin{array}{r} -\frac{d}{dx}\left(\mathbf{A}(\omega)(x)\frac{d}{dx}\mathbf{u}(\omega)(x)\right) = f(x) \quad \text{in } (0,1), \\ \mathbf{u}(\omega)(0) = \mathbf{u}(\omega)(1) = 0, \end{array} \right\} \quad \text{for almost all } \omega \in \Omega. \tag{2.27}$$

The weak formulation of (2.27) seeks an $H_0^1(0,1)$-valued random field $\mathbf{u}$ such that for almost all $\omega \in \Omega$

$$\int_0^1 \mathbf{A}(\omega)\mathbf{u}(\omega)'v'\,\mathrm{d}x = \int_0^1 fv\,\mathrm{d}x \quad \text{for all } v \in H_0^1(0,1). \tag{2.28}$$

The reformulation of this problem in the Hilbert space $L^2(\Omega; H_0^1(0,1))$ of $H_0^1(0,1)$-valued random fields with finite second moments leads to a coercive variational problem that seeks $\mathbf{u} \in L^2(\Omega; H_0^1(0,1))$ such that

$$\int_\Omega \int_0^1 A(\omega)\mathbf{u}(\omega)'\mathbf{v}(\omega)'\,\mathrm{d}x\,\mathrm{d}\mathbb{P}(\omega) = \int_\Omega \int_0^1 f\mathbf{v}(\omega)\,\mathrm{d}x\,\mathrm{d}\mathbb{P}(\omega) \tag{2.29}$$

holds for all $\mathbf{v} \in L^2(\Omega; H_0^1(0,1))$. It is easily checked that (2.29) is a well-posed problem in the sense of the Lax-Milgram theorem with a coercive and bounded bilinear form

$$a : L^2(\Omega; H_0^1(0,1)) \times L^2(\Omega; H_0^1(0,1)) \to \mathbb{R},$$
$$(\mathbf{u}, \mathbf{v}) \mapsto \int_\Omega \int_0^1 A(\omega)\mathbf{u}(\omega)'\mathbf{v}(\omega)' \,\mathrm{d}x\,\mathrm{d}\mathbb{P}(\omega)$$

and a bounded linear functional

$$\mathbf{v} \mapsto \int_\Omega \int_0^1 f(x)\mathbf{v}(\omega)(x) \,\mathrm{d}x\,\mathrm{d}\mathbb{P}(\omega)$$

on $L^2(\Omega; H_0^1(0,1))$. This shows that, for any deterministic $f \in L^2(0,1)$[3], there exists a unique solution $\mathbf{u} \in L^2(\Omega; H_0^1(0,1))$ with

$$\|\mathbf{u}\|_{L^2(\Omega;H_0^1(0,1))} := \left( \int_\Omega \int_0^1 (\mathbf{u}(\omega)'(x))^2 \,\mathrm{d}x\,\mathrm{d}\mathbb{P}(\omega) \right)^{1/2} \leq \pi^{-1}\alpha^{-1}\|f\|_{L^2(0,1)}.$$

Though it would be possible, we disregard the possibility of more general $f \in H^{-1}(0,1)$ or randomness in the right-hand side $f$ in this section.

As in the previous sections we consider the mesh $\mathcal{T}_H := \{T = [x_j, x_{j+1}] \,|\, j = 0, 1, \ldots, N\}$ represented by $N+2$ mesh points

$$0 := x_0 < x_1 < x_2 < \ldots < x_{N+1} =: 1.$$

We have seen in the previous sections that, for any event $\omega \in \Omega$, there exists an effective deterministic coefficient $\mathbf{A}_H(\omega) \in \mathcal{M}([0,1], \alpha, \beta)$ such that, for any $f \in L^2(0,1)$, the solution $\mathbf{u}_H(\omega) \in H_0^1(0,1)$ of the deterministic problem

$$\int_0^1 \mathbf{A}_H(\omega)\mathbf{u}_H(\omega)'v' \,\mathrm{d}x = \int_0^1 fv \,\mathrm{d}x \quad \text{for all } v \in H_0^1(0,1)$$

correctly captures the expected macroscopic behavior of the random field $\mathbf{u}(\omega)$ for the event $\omega$ in the sense that

$$\|\mathbf{u}(\omega) - \mathbf{u}_H(\omega)\|_{L^2(0,1)} \leq C_{\alpha,\beta} H \|f\|_{L^2(0,1)}$$

with $C_{\alpha,\beta} := \frac{4}{\alpha\pi^2}\sqrt{\beta/\alpha}$, cf. (2.16). This and Jensen's inequality imply that

$$\|\mathbb{E}(\mathbf{u}) - \mathbb{E}(\mathbf{u}_H)\|_{L^2(0,1)} \leq \mathbb{E}(\|\mathbf{u} - \mathbf{u}_H\|_{L^2(0,1)}) \leq C_{\alpha,\beta} H \|f\|_{L^2(0,1)},$$

where $\mathbb{E}\mathbf{u} := \int_\Omega \mathbf{u}(\omega) \,\mathrm{d}\mathbb{P}(\omega)$ (resp. $\mathbb{E}\mathbf{u}_H$ and $\mathbb{E}(\|\mathbf{u} - \mathbf{u}_H\|_{L^2(0,1)})$) denotes the expectation of $\mathbf{u}$ (resp. $\mathbf{u}_H$ and the random variable $\|\mathbf{u} - \mathbf{u}_H\|_{L^2(0,1)}$). This means that the expectation of $\mathbf{u}$ is well approximated by $\mathbb{E}(\mathbf{u}_H)$. However, the approximation of $\mathbb{E}(\mathbf{u}_H)$

[3] Actually, there is a unique solution for all $f$ in the dual space of $L^2(\Omega; H_0^1(0,1))$.

by a sampling procedure requires the generation of many sample coefficients and each sample requires (e.g. Monte Carlo) the computation of an effective coefficient and the solution of the corresponding deterministic differential equation. In general, it is not clear that the statistical properties such as the variance of the $\mathcal{M}(\mathcal{T}_H,\alpha,\beta)$-valued random field $\mathbf{A}_H$ are any better than those of the original $\mathcal{M}([0,1],\alpha,\beta)$-valued random field $\mathbf{A}$. So we cannot expect that a Monte Carlo method with the spatially homogenized field $\mathbf{A}_H$ requires a smaller number of samples than a Monte Carlo method for the original problem with the field $\mathbf{A}$. Without any structural assumptions on $\mathbf{A}$, we cannot even expect a significant reduction of computing times for a single sample as the computation of $\mathbf{A}_H(\omega)$ requires the previous generation of $\mathbf{A}(\omega)$ which might oscillate on scales much smaller than $H$ so that this step dominates the overall complexity.

However, under further assumptions on the distribution of $\mathbf{A}$, stronger results are possible. We shall give one simple example. Assume that $\mathbf{A}$ represents a random medium, where in each cell $t$ of a uniform mesh $\mathcal{T}_\varepsilon$ of width $H \gg \varepsilon > 0$ (such that $H/\varepsilon \in \mathbb{N}$), the conductivity $\mathbf{A}|_t = \mathbf{A}|_*$ has the same probability distribution over the interval $[\alpha,\beta]$ and the conductivities for any two different cells are mutually independent, i.e., the random field $\mathbf{A}$ is independent and identically distributed (i.i.d.). With this simplifying assumption, we have that, for almost all $\omega \in \Omega$ and all $T \in \mathcal{T}_H$,

$$\begin{aligned}\mathbf{A}_H(\omega)|_T &= \frac{|T|}{\int_T 1/\mathbf{A}(\omega)(x)\,\mathrm{d}x} = \frac{|T|}{\sum_{t\in\mathcal{T}_\varepsilon : t\subset T}\int_t 1/\mathbf{A}|_t(\omega)\,\mathrm{d}x}\\ &= \frac{|T|}{\varepsilon\sum_{t\in\mathcal{T}_\varepsilon : t\subset T} 1/\mathbf{A}|_t(\omega_t)} = \frac{H}{\varepsilon\sum_{i=1}^{H/\varepsilon} 1/\mathbf{A}|_*(\omega_i)} = \left(\mathbb{E}_{H/\varepsilon}[1/\mathbf{A}_*(\cdot)]\right)^{-1},\end{aligned}$$

where $\mathbb{E}_N[X] = \frac{1}{N}\sum_{i=1}^N X_i$ denotes the emperical or sample mean of the random variable $X$. The law of large numbers (or the theory of Monte Carlo methods) tells us that $\mathbf{A}_H$ is almost a deterministic coefficient in the sense that its variance scales like $\sqrt{\varepsilon/H}$ (the reciprocal square root of the number of samples) and so is $\mathbf{u}_H$. Hence, for sufficiently small $\varepsilon$ (relative to $H$), we may replace $\mathbf{A}_H$ with its expectation $\mathbb{E}[\mathbf{A}_H]$, denoted as $A_H$. This quantity is a global constant and the corresponding solution $u_H$ approximates $I_H\mathbb{E}\mathbf{u} = \mathbb{E}I_H\mathbf{u}$ for all observation scales $H > 0$ in the sense that $u_H$ and $\mathbb{E}I_H\mathbf{u}$ conincide in the limit $\varepsilon \to 0$. If the distribution is uniform over $[\alpha,\beta]$, one easily computes

$$\lim_{\varepsilon\to 0} A_H = \lim_{\varepsilon\to 0} 1/\mathbb{E}(1/\mathbf{A}(\cdot)(x)) = \frac{\beta-\alpha}{\log(\beta)-\log(\alpha)}.$$

Such a medium, hence, achieves homogenization in the sense that the expected behavior of the solution is well-captured by any suitable approximate solution of the deterministic problem

$$\int_0^1 A_H u_H' v'\,\mathrm{d}x = \int_0^1 f v\,\mathrm{d}x \quad \text{for all } v \in H_0^1(0,1) \tag{2.30}$$

In this simple one-dimensional setting, this result recovers classical (and far more general) results from stochastic homogenization [43, 56, 65] and the recent approaches [10], [29, 30, 27, 28, 31, 19, 32], and [4, 3].

Chapter **??** will study the numerical approach to numerical stochastic homogenization for more interesting cases of distributions and corresponding numerical homogenization methods that allow to identify an effective deterministic coefficient $A_H \in \mathcal{M}([0,1],\alpha,\beta)$ such that, for any $f \in L^2(0,1)$, the solution $u_H \in H^1_0(0,1)$ of (2.30) correctly captures the expected macroscopic behavior of the random field $\mathbf{u}$ in the sense that

$$\|\mathbb{E}\mathbf{u} - u_H\|_{L^2(0,1)} \leq C_{\alpha,\beta} H \|f\|_{L^2(0,1)}.$$

with some generic constant $C_{\alpha,\beta}$.

# Chapter 3
# Decompositions of Scales in Elliptic Problems

In this Chapter we study an elliptic model problem with a rough diffusion matrix, posed on a bounded domain in $\mathbb{R}^d$. We do not assume periodicity or scale separation in the diffusion matrix. In the spirit of Section 2.4 we construct a finite dimensional test space that is ideal for numerical homogenization. A key component in the construction is the use of a quasi-interpolation operator. The kernel of that operator defines the fine scales of the problem and its $a$-orthogonal complement defines the ideal function space used for numerical homogenization. There are no analytical expressions for the basis functions that span this space for $d > 1$. Instead the basis has to be computed numerically.

## 3.1 Model Problem with Rough Diffusion

Classical homogenization theory relies on strong structural assumptions, such as periodicity and scale separation, on the diffusion coefficient. In practical applications however, it is often impossible to model material properties encoded in the coefficient by a locally periodic coefficient of the form $A(x) = A(x, \frac{x}{\varepsilon})$ with a 1-periodic $A(x, \cdot)$. Often, we are not even able to identify a parameter $\varepsilon$ that represents microscopic oscillations. In those cases we are still interested in coarse representations of the partial differential operator (in which $A$ is the diffusion coefficient) that allows the efficient simulation on some macroscopic scale of interest.

We consider the following Poisson type boundary value problem with rough diffusion

$$-\nabla \cdot (A\nabla u) = f \tag{3.1}$$

in a bounded domain $D \subset \mathbb{R}^d$ with homogeneous Dirichlet boundary condition. The diffusion matrix $A$ is allowed to be strongly heterogeneous, highly varying, and non-periodic. The heterogeneities and oscillations of the coefficient may appear on several non-separable scales. More specifically we only assume the diffusion matrix $A \in \mathcal{M}_{\text{sym}}(D, \alpha, \beta)$ to be symmetric and uniformly elliptic with

$$0 < \alpha = \operatorname*{ess\,inf}_{x\in D} \inf_{v\in\mathbb{R}^d\setminus\{0\}} \frac{(A(x)v)\cdot v}{v\cdot v}, \tag{3.2}$$

$$\infty > \beta = \operatorname*{ess\,sup}_{x\in D} \sup_{v\in\mathbb{R}^d\setminus\{0\}} \frac{(A(x)v)\cdot v}{v\cdot v}. \tag{3.3}$$

We let the right hand side $f \in L^2(D)$ and seek a weak solution to the model problem. Find $u \in V := \mathrm{H}^1_0(D)$ such that

$$a(u,v) := \int_D (A\nabla u)\cdot\nabla v = \int_D fv =: F(v) \quad \text{for all } v \in V. \tag{3.4}$$

The existence and uniqueness of a solution to this problem is guaranteed by the Riesz representation theorem since $a(\cdot,\cdot)$ is a scalar product and $F(\cdot)$ is a linear functional on the Hilbert space $V$.

The great challenge in computing a numerical approximation of $u$, as seen in Section 2.1, is that the computational mesh has to resolve the variations in the diffusion in order to achieve an accurate solution. Following the ideas presented in Section 2.4 we construct a generalized finite element space that allows for an accurate representation of the solution $u$. We start by constructing the classical finite element method and quasi interpolation operators onto finite element spaces which will play an important role in the construction of the numerical method.

## 3.2 Finite Element Spaces

This section presents some preliminaries on finite element meshes and spaces. We consider two discretization scales $H > h > 0$. Let $\mathcal{T}_H$ (resp. $\mathcal{T}_h$) denote corresponding regular (in the sense of [13]) finite element meshes of $D$ into closed simplices with mesh-size functions $0 < H \in L^\infty(D)$ defined by $H|_T = \operatorname{diam}(T) =: H_T$ for all $T \in \mathcal{T}_H$ (resp. $0 < h \in L^\infty(D)$ defined by $h|_t = \operatorname{diam}(t) =: h_t$ for all $t \in \mathcal{T}_h$). The mesh sizes may vary in space but we will not exploit the possible mesh adaptivity.

The error bounds, typically, depend on the maximal mesh sizes $\|H\|_{L^\infty(D)}$. If no confusion seems likely, we will use $H$ also to denote the maximal mesh size instead of writing $\|H\|_{L^\infty(D)}$. For the sake of simplicity we assume that $\mathcal{T}_h$ is derived from $\mathcal{T}_H$ by some regular, possibly non-uniform, mesh refinement. However, this condition is not essential, see [46, 36].

As usual in finite element analysis, the error analysis depends on the constant $\gamma > 0$ which represents the shape regularity of the finite element mesh $\mathcal{T}_H$;

$$\gamma := \max_{T\in\mathcal{T}_H} \gamma_T \quad \text{with} \quad \gamma_T := \frac{\operatorname{diam}(T)}{\operatorname{diam}(B_T)} \ \text{ for } T \in \mathcal{T}_H, \tag{3.5}$$

where $B_T$ denotes the largest ball contained in $T$.

The first-order conforming finite element space corresponding to $\mathcal{T}_H$ is given by

$$V_H := \{v \in V \mid \forall T \in \mathcal{T}_H, v|_T \text{ is a polynomial of total degree} \leq 1\}. \tag{3.6}$$

Let $\mathcal{N}_H$ denote the set of interior vertices of $\mathcal{T}_H$. For every vertex $z \in \mathcal{N}_H$, let $\Lambda_z \in V_H$ denote the corresponding nodal basis function (tent/hat function) determined by nodal values

$$\Lambda_z(z) = 1 \text{ and } \Lambda_z(y) = 0 \quad \text{for all } y \neq z \in \mathcal{N}_H.$$

These nodal basis functions form a basis of $V_H$. The dimension of $V_H$ equals the number of interior vertices,

$$N_H := \dim V_H = |\mathcal{N}_H|.$$

Let $V_h \supset V_H$ denote some conforming finite element space corresponding to the fine mesh $\mathcal{T}_h$. It can be the space of continuous piecewise affine functions on the fine mesh or any other (generalized) finite element space that contains $V_H$, e.g., the space of continuous $p$-th order piecewise polynomials as in [61]. By $N_h := \dim V_h$ we denote the dimension of $V_h$. For standard choices of $V_h$, this dimension is proportional to the number of interior vertices in the fine mesh $\mathcal{T}_h$.

## 3.3 Quasi-interpolation

We wish to generalize the idea of decomposing the space $V$ into two $a$-orthogonal subspaces from Section 2.4. In one dimension this was achieved by letting the fine scale space $W$ be the kernel of the nodal interpolant and the multiscale space $V_H^{\text{ms}}$ be its $a$-orthogonal complement in $V$. Since the nodal interpolant is not well defined for functions in $V$ in dimensions $d > 1$ this approach does not immediately apply. We will instead use a quasi-interpolant with some desired properties.

We let $\mathcal{I}_H : V \to V_H$ be a quasi-interpolation operator that acts as a stable quasi-local projection in the sense that $\mathcal{I}_H \circ \mathcal{I}_H = \mathcal{I}_H$ and that for any $T \in \mathcal{T}_H$ and all $v \in V$ it, for a generic constant $C_I$, holds

$$H^{-1}\|v - \mathcal{I}_H v\|_{L^2(T)} + \|\nabla \mathcal{I}_H v\|_{L^2(T)} \leq C_I \|\nabla v\|_{L^2(\mathsf{N}(T))}, \tag{3.7}$$

where

$$\mathsf{N}(S) = \bigcup \left\{ K \in \mathcal{T}_H : K \cap \bar{S} \neq \emptyset \right\}$$

refers to the union of $S$ and the adjacent elements. The constant $C_I$ depends on the shape regularity parameter $\gamma$ of the finite element mesh $\mathcal{T}_H$ (see (3.5) above) but not on $H_T$.

Note that there exists a constant $C_{\text{ol}} > 0$ that only depends on $\gamma$ such that the number of elements covered by $\mathsf{N}(T)$ is uniformly bounded (w.r.t. $T$) by $C_{\text{ol}}$,

$$\max_{T \in \mathcal{T}_H} |\{K \in \mathcal{T}_H \mid K \subset \mathsf{N}(T)\}| \leq C_{\text{ol}}. \tag{3.8}$$

One possible construction of an interpolant which fulfills the properties is $\mathcal{I}_H := E_H \circ \Pi_H$, where $\Pi_H$ is the piecewise $L^2$ projection onto $P^1(\mathcal{T}_H)$ and $E_H$ is the averaging operator that maps $P_1(\mathcal{T}_H)$ to $V_H$ by assigning to each interior vertex the arithmetic mean of the corresponding function values of the adjacent elements, that is, for any $v \in P_1(\mathcal{T}_H)$ and any free vertex $z \in \mathcal{N}_H$,

$$(E_H(v))(z) = \frac{1}{\text{card}\{K \in \mathcal{T}_H : z \in K\}} \sum_{T \in \mathcal{T}_H : z \in T} v|_T(z).$$

For this choice, the proof of (3.7) follows by combining the well-established approximation and stability properties of $\Pi_H$ and $E_H$, see for instance [23]. This is by no means a unique choice. In [47] a weighted Clément interpolant was used. This choice turned out to be particularly useful for eigenvalue computations [46]. We will get back to this particular Clément type interpolant in Chapter **??**. Also problems of high contrast data may need a more carefully tuned interpolation operator as shown in Section **??**. This diversity of interpolants used in practise reflects the importance of having a correct representation of the fine scales (which in this approach is defined as the kernel of the interpolation operator).

## 3.4 Orthogonalization of scales and ideal numerical homogenization

By generalizing Section 2.4 to higher dimensions we will now presents a numerical approach to homogenization that is not based on the mathematical theory of homogenization but the availability of operator-dependent subspaces with a quasi-local basis and approximation properties independent of oscillations and roughness of the diffusion coefficient. The method does not rely on symmetry of $A$. Since some arguments are more illustrative in the symmetric case and to stay as close as possible to the $1d$ template, we will still assume symmetry of $A$.

We consider the weak form (3.4) of the model problem (3.1) posed on the domain $D \subset \mathbb{R}^d$ and let the finite element mesh $\mathcal{T}_H$ be some regular mesh of $D$. Here the mesh size parameter represents the scale of interest that can be chosen independent of characteristic length scales of $A$. As in Section 2.4, we shall characterize the functions in $V$ that are not well-captured by finite element shape functions. Note, however, that a characterization by nodal values as in (2.18) is not possible in dimension $d > 1$. This is where the quasi-interpolation operator $\mathcal{I}_H : V \to V_H$, that is based on volume averaging, comes into play. Define

$$W := \{w \in V \,|\, \mathcal{I}_H w = 0\} = \text{kern}\mathcal{I}_H, \tag{3.9}$$

the space of (microscopic) fine-scale functions. (Observe that we could have written $W = \text{kern}\mathcal{I}_H^{\text{nodal}}$ in the one-dimensional case with the nodal interpolation operator $\mathcal{I}_H^{\text{nodal}}$.)

The remaining steps of the derivation widely coincide with Section 2.4. Observe that the solution space $V$ can be decomposed as

$$V = V_H \oplus W \tag{3.10}$$

and, for any $v \in V$, $\mathcal{I}_H v \in V_H$ and $(1-\mathcal{I}_H)v \in W$ are the unique elements of $V_H$ and $W$ such that

$$v = \mathcal{I}_H v + (1-\mathcal{I}_H)v.$$

Moreover, by equation (3.7), the decomposition is stable in the sense that

$$\|\mathcal{I}_H v\| + \|(1-\mathcal{I}_H)v\| \leq 2C_I\|v\|,$$

where $\|\cdot\|$ refers to either the $L^2(D)$ norm or the $H^1(D)$ norm. (In contrast to the $1d$ case with nodal interpolation, $\mathcal{I}_H$ is only an oblique projection with respect to both $L^2(D)$ and $H^1_0(D)$.)

We now construct the orthogonal compliment to $W$ in the space $V$ using the scalar product

$$a(u,v) := \int_D (A\nabla u)\cdot\nabla v \,\mathrm{d}x \tag{3.11}$$

associated with the problem (3.4).
Keeping $W$ fixed, we characterize a new space $V_H^{\mathrm{ms}} \subset V$ as the subspace that satisfies

$$V = V_H^{\mathrm{ms}} \oplus W \quad \text{and} \quad a(V_H^{\mathrm{ms}}, W) = 0,$$

i.e.,

$$V_H^{\mathrm{ms}} := \{v_H^{\mathrm{ms}} \in V \mid \forall w \in W : a(v_H^{\mathrm{ms}}, w) = 0\}. \tag{3.12}$$

The Galerkin method with subspace $V_H^{\mathrm{ms}}$ applied to (3.4) seeks $u_H^{\mathrm{ms}} \in V_H^{\mathrm{ms}}$ such that

$$a(u_H^{\mathrm{ms}}, v) = F(v) \tag{3.13}$$

for all $v \in V_H^{\mathrm{ms}}$. By Galerkin orthogonality

$$a(u - u_H^{\mathrm{ms}}, v) = 0$$

for all $v \in V_H^{\mathrm{ms}}$, the error $u - u_H^{\mathrm{ms}}$ of this method is a fine-scale function, i.e.,

$$\mathcal{I}_H u = \mathcal{I}_H u_H^{\mathrm{ms}}. \tag{3.14}$$

With equation (3.7) this readily implies that

$$\begin{aligned}\|u - u_H^{\mathrm{ms}}\|_{L^2(D)} &= \|(1-\mathcal{I}_H)(u - u_H^{\mathrm{ms}})\|_{L^2(D)} \\ &\leq C_I H \|\nabla(u - u_H^{\mathrm{ms}})\|_{L^2(D)}.\end{aligned}$$

Moreover,

$$\begin{aligned}\|\nabla(u-u_H^{\mathrm{ms}})\|_{L^2(D)}^2 &\leq \alpha^{-1}a(u-u_H^{\mathrm{ms}},u-u_H^{\mathrm{ms}})\\ &= \alpha^{-1}a(u,u-u_H^{\mathrm{ms}})\\ &= \alpha^{-1}\int_D f(u-u_H^{\mathrm{ms}})\,\mathrm{d}x\\ &\leq \alpha^{-1}\|f\|_{L^2(D)}\|u-u_H^{\mathrm{ms}}\|_{L^2(D)}.\end{aligned}$$

The combination of the previous two estimates yields the following theorem.

**Theorem 3.1 (Error of the ideal method).** *Let $u$ solve equation* (3.4) *and $u_H^{\mathrm{ms}}$ solve equation* (3.13)*. Then it holds,*

$$\|\nabla(u-u_H^{\mathrm{ms}})\|_{L^2(D)} \leq C_I\alpha^{-1}H\|f\|_{L^2(D)}, \tag{3.15}$$

$$\|u-u_H^{\mathrm{ms}}\|_{L^2(D)} \leq C_I^2\alpha^{-1}H^2\|f\|_{L^2(D)}. \tag{3.16}$$

We refer to the approximation $u_H^{\mathrm{ms}}$ as the ideal multiscale approximation since it preserves optimal order a priori error bounds without assumptions on regularity of the solution (beyond $H_0^1(D)$) and with constants independent of the variation (derivatives) in the diffusion $A$.

This is in agreement with the $1d$ result (2.20). However, the multidimensional case is structurally very different from the $1d$ case with nodal interpolation when it comes to the practical feasibility of the method. We return to this issue in the next chapter. First we discuss modifications of the original formulation with different advantages.

## 3.5 Modifications of the original method

We start by reformulating the method to a finite element method with modified bilinear form. For this purpose, let $-Q_H : V \to W$ denote the $a$-orthogonal projection onto the closed subspace $W \subset V$. We will refer to $Q_H$ as the *correction operator*. Note that its complementary projection

$$(1-(-Q_H)) = (1+Q_H)$$

maps $V_H$ onto $V_H^{\mathrm{ms}}$ and is invertible with inverse $\mathcal{I}_H$. We can, hence, identify any $v_H^{\mathrm{ms}} \in V_H^{\mathrm{ms}}$ with its finite element component $v_H = \mathcal{I}_H v_H^{\mathrm{ms}}$ and vice versa $v_H^{\mathrm{ms}} = (1+Q_H)v_H$. This allows us to reformulate the method (3.13) as follows: Find $u_H \in V_H$ such that

$$a((1+Q_H)u_H,(1+Q_H)v_H) = F((1+Q_H)v_H) \tag{3.17}$$

for all $v_H \in V_H$. Replacing the problem-dependent right-hand side (the evaluation of $Q_H$ requires the solution of variational problems based on the bilinear form $a$) with $v_H \mapsto F(v_H)$ yields the variant of (3.13): Find $\bar{u}_H \in V_H$ such that

$$a((1+Q_H)\bar{u}_H,(1+Q_H)v_H) = F(v_H) \tag{3.18}$$

for all $v_H \in V_H$.

**Lemma 3.1 (Error of the coarse scale approximation).** *The discrete problem* (**??**) *admits a unique solution $\bar{u}_H \in V_H$ for any $F \in L^2(D)$ and the error is bounded by*

$$\|u - \bar{u}_H\|_{L^2(D)} \leq C\Big(\min_{v_H \in V_H} \|u - v_H\|_{L^2(D)} + H\|f\|_{L^2(D)}\Big).$$

*Proof.* Observe that the modified bilinear form

$$a((1+Q_H)\bullet, (1+Q_H)\bullet) : V_H \times V_H \to \mathbb{R}$$

satisfies, for any $v_H \in V_H$,

$$\begin{aligned} a((1+Q_H)v_H, (1+Q_H)v_H) \geq \quad & \alpha \|\nabla(1+Q_H)v_H\|^2_{L^2(D)} \\ \geq \quad & \frac{\alpha}{C_I^2}\|\nabla v_H\|^2_{L^2(D)} \end{aligned} \tag{3.19}$$

because of $\mathcal{I}_H(1+Q_H)v_H = v_H$ and the $H^1_0(D)$-stability of $\mathcal{I}_H$. This proves well-posedness of (**??**) and, in particular, unique solvability.
To prove the error estimate, observe that the solution $u_H$ of (3.17) (which is well-posed by (3.19) as well) is exactly $u_H = \mathcal{I}_H u = \mathcal{I}_H u_H^{\mathrm{ms}}$. This implies, for any $v_H \in V_H$,

$$\begin{aligned} \|u - u_H\|_{L^2(D)} = \quad & \|(1-\mathcal{I}_H)(u - v_H)\|_{L^2(D)} \\ \leq \quad & C_I \|u - v_H\|_{L^2(D)} \end{aligned} \tag{3.20}$$

The error $(u_H - \bar{u}_H)$ can be estimated with (3.19),

$$\begin{aligned} \frac{\alpha}{C_I^2}\|\nabla(\bar{u}_H - u_H)\|^2_{L^2(D)} &\leq a((1+Q_H)(\bar{u}_H - u_H), (1+Q_H)(\bar{u}_H - u_H)) \\ &= \int_D f \underbrace{Q_H(u_H - \bar{u}_H)}_{=(1-\mathcal{I}_H)Q_H(u_H - \bar{u}_H)} \mathrm{d}x \\ &\leq \|f\|_{L^2(D)} C_I H \|\nabla(\bar{u}_H - u_H)\|_{L^2(D)}. \end{aligned}$$

Hence, Friedrichs' inequality yields

$$\|\bar{u}_H - u_H\|_{L^2(D)} \leq \frac{C_F C_I^3}{\alpha} H \|f\|_{L^2(D)}.$$

Since

$$\|\bar{u}_H - u\|_{L^2(D)} + \leq \|u_H - u\|_{L^2(D)} + \|\bar{u}_H - u_H\|_{L^2(D)}$$

the assertion follows.

A sharper version of this result can be derived in the following way, [25].

**Theorem 3.2.** *The solutions $u \in V$ to* (3.4) *and $\bar{u}_H \in V_H$ to* (3.18) *for right-hand side $f \in L^2(\Omega)$ satisfy the following error estimate*

$$\frac{\|u - \bar{u}_H\|_{L^2(\Omega)}}{\|f\|_{L^2(\Omega)}} \lesssim H^2 + wcba(A, \mathcal{T}_H).$$

# Chapter 4
# Numerical Homogenization Beyond Periodicity and Scale Separation

In this chapter we will derive and analyze the Localized Orthogonal Decomposition method based on the ideal method for numerical homogenization presented in the previous chapter. The crucial steps are the construction and truncation of the corrected basis functions. This localization leads to a sparse matrix representation on the macro scale. The error committed will be controlled by the exponential decay in the correctors. We will also discuss how this method can be reinterpreted as a domain decomposition method. Furthermore, we study the challenging problem of high contrast diffusion, how it affects the decay, and how it can be handled by modifying the interpolation operator used to define the fine scales.

## 4.1 Exponential Decay of the Finescale Green's Function

A first step towards turning the ideal method into an efficient numerical method is to construct a basis for the space $V_H^{\mathrm{ms}}$. We construct the basis in the same way as in Section 2.4,

$$V_H^{\mathrm{ms}} = \operatorname{span}\{\Lambda_x^{\mathrm{ms}} := (1 + Q_H)\Lambda_x \,|\, x \in N_H\}.$$

In order to get a feasible numerical method, it is crucial to have a sparse (i.e. local) basis representation. The function $Q_H\Lambda_x$ will have global support in general as opposed to $\Lambda_x$. In this section we justify a localization procedure that leads to a (quasi-)local variant of the method (3.13). To this end, we introduce the following *element corrector* for any $T \in \mathcal{T}_H$ and unit vector $e_i$ ($i \in \{1, ..., d\}$): Let $w_{T,i} \in W$ solve

$$a(w_{T,i}, v) = -\int_T (Ae_i) \cdot \nabla v \,\mathrm{d}x \tag{4.1}$$

for all $v \in W$. We claim that, for all $v_H \in V_H$,

$$Q_H v_H = \sum_{T \in \mathcal{T}_H} \sum_{i=1}^{d} \left.\frac{\partial v_H}{\partial x_i}\right|_T w_{T,i}. \tag{4.2}$$

Indeed, we have, for any $v \in W$,

$$\begin{aligned} a\Big(\sum_{T\in\mathcal{T}_H}\sum_{i=1}^{d}\frac{\partial v_H}{\partial x_i}\Big|_T w_{T,i}, v\Big) &= \sum_{T\in\mathcal{T}_H}\sum_{i=1}^{d}\frac{\partial v_H}{\partial x_i}\Big|_T a(w_{T,i}, v) \\ &\overset{(4.1)}{=} -\sum_{T\in\mathcal{T}_H}\sum_{i=1}^{d}\frac{\partial v_H}{\partial x_i}\Big|_T \int_T (Ae_i)\cdot\nabla v \,\mathrm{d}x \\ &= -a(v_H, v) \end{aligned}$$

and this is exactly the equation to be satisfied by the negative of the Galerkin projection onto $W$. We see from (4.2) that $Q_H v_H$ is built from a sum of contributions that are solutions to (4.1), a problem with locally supported right-hand side. When we compute a numerical approximation of $w_{T,i}$, by discretizing the space $W$ with a fine mesh $\mathcal{T}_h$, we see clearly that it decays exponentially away form the element $T$, see Figure (4.1).

In the following, we shall prove that $w_{T,i}$ decays exponentially fast away from $T$. For a proper quantification of the decay we introduce the following element neighborhoods (or patches) for a subdomain $S \subset D$,

$$\mathsf{N}^\ell(S) = \mathsf{N}(\mathsf{N}^{\ell-1}(S)) \text{ for } \ell \geq 2 \text{ and } \mathsf{N}^1(S) := \mathsf{N}(S).$$

where we recall the definition of an element patch from Chapter 3, $\mathsf{N}(S) = \{K \in \mathcal{T}_H : K \cap \bar{S} \neq \emptyset\}$. Due to the many constants in the proof we introduce the notation $a \lesssim b$ which abbreviates $a \leq Cb$ for some constant $C$, that is independent of the mesh-size and $\ell$ but may depend on the contrast of the coefficient $A$.

**Theorem 4.1 (Exponential decay).** *Let $T \in \mathcal{T}_H$, $i \in \{1, ..., d\}$ and let $w_{T,i} \in W$ solve (4.1). Then there exists $c = c(\alpha,\beta) > 1$ (independent of $T$) such that, for any $\ell \geq 5$,*

$$\|\nabla w_{T,i}\|_{L^2(D\setminus\mathsf{N}^\ell(T))} \lesssim exp(-c\ell)|T|^{1/2}.$$

*Proof.* We consider Lipschitz continuous cutoff functions $\eta_\ell \in W^{1,\infty}(D;[0,1])$ with

$$\begin{aligned} \eta_\ell &\equiv 0 \quad \text{in } \mathsf{N}^\ell(T) \\ \eta_\ell &\equiv 1 \quad \text{in } D\setminus\mathsf{N}^{\ell+1}(T) \\ \|\nabla\eta_\ell\|_{L^\infty(D)} &\leq C_\eta H^{-1}. \end{aligned} \tag{4.3}$$

Let $\nu = \eta_{\ell-3}$ and observe that

$$\begin{aligned} \mathrm{supp}(\nu) &= D\setminus\mathsf{N}^{\ell-3}(T) \\ \mathrm{supp}(\nabla\nu) &= \mathsf{N}^{\ell-2}(T)\setminus\mathsf{N}^{\ell-3}(T) =: R. \end{aligned}$$

We abbreviate $w := w_{T,i}$ and calculate with the product rule

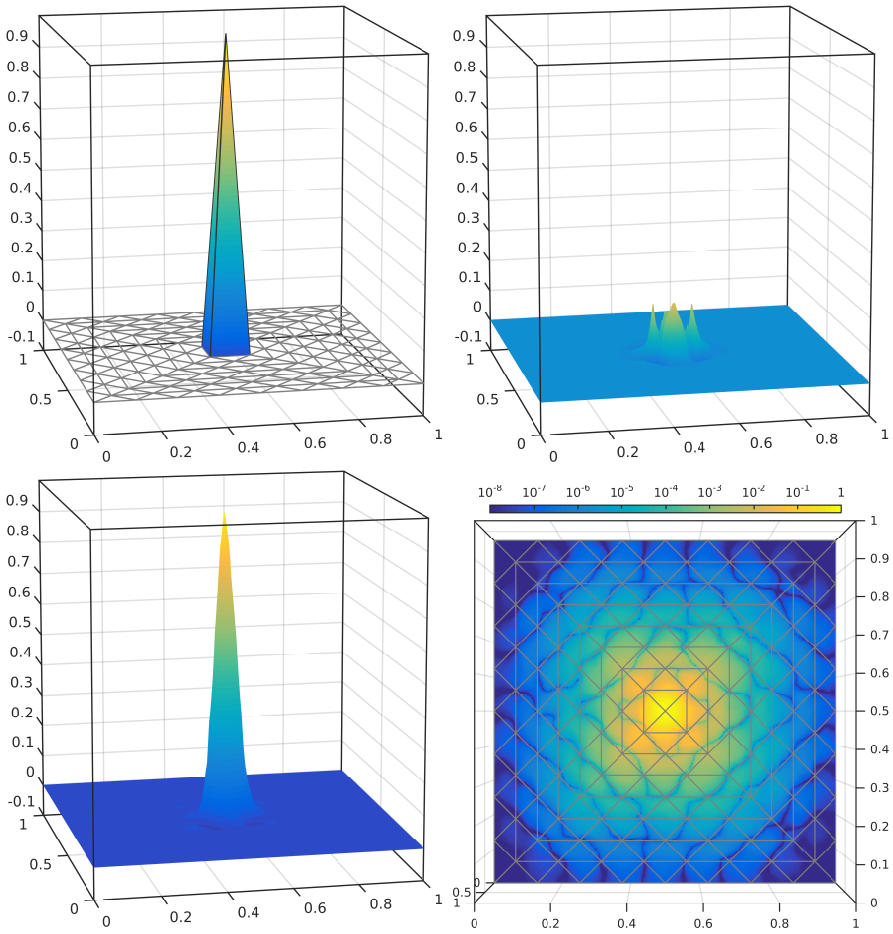


**Fig. 4.1** Standard nodal basis function $\Lambda_x$ with respect to the coarse mesh $\mathcal{T}_H$ (top left), corresponding ideal corrector $Q_H\Lambda_x$ (top right), and corresponding test basis function $\phi_x^{\mathrm{ms}} = (1+Q_H)\Lambda_x$ (bottom left). The bottom right figure shows a top view on the modulus of test basis function $\phi_x^{\mathrm{ms}} = (1+Q_H)\Lambda_x$ with logarithmic color scale to illustrate the exponential decay property. The underlying rough diffusion coefficient $A$ is depicted in Fig. **??**.

$$\begin{aligned}\|\nabla w\|^2_{L^2(D\setminus \mathsf{N}^\ell(T))} &\lesssim \|A^{1/2}\nabla w\|^2_{L^2(D\setminus \mathsf{N}^\ell(T))} \overset{A \text{ spd, } v\geq 0}{\leq} \langle A\nabla w, v\nabla w\rangle_{L^2(D)} \\ &\leq \underbrace{|\langle A\nabla w, \nabla(1-\mathcal{I}_H)(v w)\rangle_{L^2(D)}|}_{=:M_1} + \underbrace{|\langle A\nabla w, \nabla \mathcal{I}_H(v w)\rangle_{L^2(D)}|}_{=:M_2} + \underbrace{|\langle A\nabla w, w\nabla v\rangle_{L^2(D)}|}_{=:M_3}.\end{aligned}$$

We proceed by estimating $M_1$, $M_2$, $M_3$.

**$\mathbf{M_1}$** Since $v := (1-\mathcal{I}_H)(vw) \in W$ we have by (4.1)

$$M_1 = \left| \int_T (Ae_i) \cdot \nabla v \, \mathrm{d}x \right|,$$

but as the support of $v$ lies outside of $T$, we may conclude that $M_1 = 0$.

**$\mathbf{M_2}$** Since $w \in W$, we have $\operatorname{supp}(\mathcal{I}_H(vw)) \subset \mathsf{N}(R)$. Hence, with $\operatorname{supp} \nabla v = R$,

$$M_2 \lesssim C_I \|\nabla w\|_{L^2(\mathsf{N}(R))} \Big( \|v \nabla w\|_{L^2(\mathsf{N}^2(R))} + \|\nabla v\|_{L^\infty(D)} \|w\|_{L^2(R)} \Big).$$

With the bound (4.3) on $\nabla v$ and with

$$\|w\|_{L^2(R)} = \|w - \mathcal{I}_H w\|_{L^2(R)} \leq C_I H \|\nabla w\|_{L^2(\mathsf{N}(R))},$$

we conclude

$$M_2 \lesssim \|\nabla w\|^2_{L^2(\mathsf{N}^2(R))}.$$

**$\mathbf{M_3}$** Similarly, we have with (4.3) that

$$M_3 \lesssim \|\nabla w\|^2_{L^2(\mathsf{N}^2(R))}.$$

Altogether, there is a constant $\tilde{C} > 0$ such that

$$\|\nabla w\|^2_{L^2(D \setminus \mathsf{N}^\ell(T))} \leq \tilde{C} \|\nabla w\|^2_{L^2(\mathsf{N}^2(R))}. \tag{4.4}$$

Since

$$\mathsf{N}^2(R) = \mathsf{N}^\ell(T) \setminus \mathsf{N}^{\ell-5}(T),$$

we get

$$\|\nabla w\|^2_{L^2(D \setminus \mathsf{N}^\ell(T))} + \|\nabla w\|^2_{L^2(\mathsf{N}^2(R))} = \|\nabla w\|^2_{L^2(D \setminus \mathsf{N}^{\ell-5}(T))}$$

and with (4.4) it follows that

$$(1 + \tilde{C}^{-1}) \|\nabla w\|^2_{L^2(D \setminus \mathsf{N}^\ell(T))} \leq \|\nabla w\|^2_{L^2(D \setminus \mathsf{N}^{\ell-5}(T))}.$$

A repeated application of this argument with $\gamma = (1 + \tilde{C}^{-1})^{-1} < 1$ results in

$$\begin{aligned} \|\nabla w\|^2_{L^2(D \setminus \mathsf{N}^\ell(T))} &\leq \gamma^{\lfloor \ell/5 \rfloor} \|\nabla w\|^2_{L^2(D)} && (4.5) \\ &\overset{\text{stability of (4.1)}}{\lesssim} \gamma^{\lfloor \ell/5 \rfloor} \|e_i\|^2_{L^2(T)} && (4.6) \\ &\lesssim \gamma^{\lfloor \ell/5 \rfloor} |T|. && (4.7) \end{aligned}$$

Since $\gamma^{\lfloor \ell/5 \rfloor} \leq 2\exp(-c\ell)$ for some $c > 0$, this is the assertion.

The decay motivates a localized version of (4.1). Define the localized form

$$a_{\mathsf{N}^\ell(T)}(v, w) := \int_{N^\ell(T)} (A \nabla v) \cdot \nabla w \, \mathrm{d}x$$

based on the element patches $\mathsf{N}^\ell(T)$ and define $w_{T,i}^{(\ell)} \in W(N^\ell(T)) = \{v \in V : \mathcal{I}_H v = 0,\ \mathrm{supp}(v) \subset N^\ell(T)\}$ as the solution to the 'cell problem'

$$a_{\mathsf{N}^\ell(T)}(w_{T,i}^{(\ell)}, v) = -\int_T (Ae_i)\cdot\nabla v \,\mathrm{d}x \tag{4.8}$$

for all $v \in W(\mathsf{N}^\ell(T))$, where $W(\mathsf{N}^\ell(T))$ is the kernel of $\mathcal{I}_H$ when restricted to $H_0^1(\mathsf{N}^\ell(T))$ (with values in the 'restricted' finite element space). We extend these localized correctors by zero to the domain $D$ and define, for $v_H \in V_H$,

$$Q_H^{(\ell)} v_H := \sum_{T\in\mathcal{T}_H}\sum_{i=1}^{d} \frac{\partial v_H}{\partial x_i}\Big|_T w_{T,i}^{(\ell)}. \tag{4.9}$$

*Remark 4.1.* $Q_H^{(\ell)}\Lambda_z$ has now support in the nodal patch $\mathsf{N}^{\ell+1}(z)$ given by

$$\mathsf{N}^1(z) = \bigcup\Big\{T \in \mathcal{T}_H : z \in T\Big\}$$

and

$$\mathsf{N}^{\ell+1}(z) = \bigcup\Big\{T \in \mathcal{T}_H : T \cap \mathsf{N}^\ell(z) \neq \emptyset\Big\}$$

**Corollary 4.1 (Local truncation error).** *Under the same assumptions as in Theorem 4.1 it holds,*

$$\|\nabla(w_{T,i} - w_{T,i}^{(\ell)})\|_{L^2(D)} \lesssim exp(-c\ell)\,|T|^{1/2}.$$

*Proof.* We pick $\nu = 1 - \eta_{\ell-1}$, where $\eta_\ell$ is defined as in (**??**) with

$$\begin{aligned}\mathrm{supp}(\nu) &= \mathsf{N}^\ell(T)\\ \mathrm{supp}(\nabla\nu) &= \mathsf{N}^{\ell-1}(T)\setminus\mathsf{N}^{\ell-2}(T).\end{aligned}$$

We abbreviate $w := w_{T,i}$ and in the same spirit let $w^{(\ell)} := w_{T,i}^{(\ell)}$. The following relation holds

$$a(w - w^{(\ell)}, (1-\mathcal{I}_H)(\nu w)) = 0, \tag{4.10}$$

since $(1-\mathcal{I}_H)(\nu w) \in W(\mathsf{N}^\ell(T)) \subset W$. Therefore,

$$\begin{aligned}\|A^{1/2}\nabla(w - w^{(\ell)})\|_{L^2(D)} &\le \|A^{1/2}\nabla(w - (1-\mathcal{I}_H)(\nu w))\|_{L^2(D)}\\ &\le \beta^{1/2}\|\nabla(1-\mathcal{I}_H)((1-\nu)w)\|_{L^2(D)}.\end{aligned}$$

Since $\mathcal{I}_H$ is stable in $H^1$ we get, using the Poincaré inequality

$$\|A^{1/2}\nabla(w - w^{(\ell)})\|_{L^2(D)} \lesssim \|\nabla((1-\nu)w)\|_{L^2(D)} \le \|w\nabla\nu\|_{L^2(D)} + \|\nu\nabla w\|_{L^2(D)}. \tag{4.11}$$

Since $w \in W$ we can subtract the interpolant $\mathcal{I}_H w$ in the first term and use the interpolation inequality (3.7) together with the properties in (4.10) to get,

$$\|A^{1/2}\nabla(w-w^{(\ell)})\|_{L^2(D)} \lesssim \|\nabla w\|_{L^2(D\setminus \mathsf{N}^{\ell-1}(T))} \lesssim \exp(-c\ell)|T|^{1/2}, \tag{4.12}$$

by Theorem 4.1. The Corollary follows using the lower bound of $A$.

**Lemma 4.1 (Global truncation error).** *Under the same assumptions as in Theorem 4.1 it holds for any $v \in V_H$,*

$$\|\nabla(Q_H v - Q_H^{(\ell)} v)\|_{L^2(D)} \lesssim \ell^{(d-1)/2} exp(-c\ell)\|\nabla v\|_{L^2(D)}.$$

*Proof.* For $v_H \in V_H$ let $w_T = \sum_{i=1}^d \frac{\partial v_H|_T}{\partial x_i} w_{T,i}$ and $w_T^{(\ell)} = \sum_{i=1}^d \frac{\partial v_H|_T}{\partial x_i} w_{T,i}^{(\ell)}$. We start by bounding $w_T - w_T^{(\ell)}$ in terms of $v_H$. Using Theorem 4.1 we get

$$\|\nabla w_T\|_{L^2(D\setminus N^\ell(T))} \leq \sum_{i=1}^d |\frac{\partial v_H}{\partial x_i}| \|\nabla w_{T,i}\|_{L^2(D\setminus N^\ell(T))} \lesssim \exp(-c\ell)\|\nabla v_H\|. \tag{4.13}$$

Corollary 4.1 with $w := w_T$ gives

$$\|A^{1/2}\nabla(w_T - w_T^{(\ell)})\|_{L^2(D)} \lesssim \|\nabla w_T\|_{L^2(D\setminus \mathsf{N}^\ell(T))} \lesssim \exp(-c\ell)\|A^{1/2}\nabla v_H\|_{L^2(T)}, \tag{4.14}$$

using (4.13) in the last step of equation (4.12), and the uniform bounds on $A$.

Now we let $e = \sum_{T\in\mathcal{T}_H} w_T - w_T^{(\ell)} = Q_H v - Q_H^{(\ell)} v$ and introduce a new cut off function $\nu = 1 + \eta_{\ell+1} - \eta_{\ell-2}$ and observe that

$$\begin{aligned} \operatorname{supp}(1-\nu) &= \mathsf{N}^{\ell+2}(T)\setminus \mathsf{N}^{\ell-2}(T) \\ \operatorname{supp}(\nabla\nu) &\subset \mathsf{N}^{\ell+2}(T)\setminus \mathsf{N}^{\ell-2}(T). \end{aligned}$$

We have that

$$a(w_T - w_T^{(\ell)}, (1-\mathcal{I}_H)(\nu e)) = 0,$$

since

$$a(w_T, (1-\mathcal{I}_H)(\nu e)) = \int_T A\nabla v_H \cdot \nabla(1-\mathcal{I}_H)(\nu e)\, dx$$

due to $(1-\mathcal{I}_H)(\nu e) \in W$ and

$$a(w_T^{(\ell)}, (1-\mathcal{I}_H)(\nu e)) = \int_T A\nabla v_H \cdot \nabla(1-\mathcal{I}_H)(\nu e)\, dx$$

since $(1-\mathcal{I}_H)\eta_{\ell+1}$ and $w_T^{(\ell)}$ have disjoint support and $(1-\mathcal{I}_H)(1-\eta_{\ell-2}) \in W(N^\ell(T))$.

We can now proceed with the following calculation

$$\|A^{1/2}\nabla(Q_H v - Q_H^{(\ell)} v)\|_{L^2(D)}^2 = \sum_{T\in\mathcal{T}_H} a(e, w_T - w_T^{(\ell)}) \tag{4.15}$$

$$= \sum_{T\in\mathcal{T}_H} a((1-\mathcal{I}_H)(e-ve), w_T - w_T^{(\ell)}) \tag{4.16}$$

$$\lesssim \sum_{T\in\mathcal{T}_H} \|\nabla e\|_{L^2(N^{\ell+2}(T)\setminus N^{\ell-2}(T))} \|\nabla(w_T - w_T^{(\ell)})\|_{L^2(D)}. \tag{4.17}$$

We get

$$\|A^{1/2}\nabla(Q_H v - Q_H^{(\ell)} v)\|_{L^2(D)}^2 \lesssim \exp(-c\ell)\|\nabla v_H\|_{L^2(D)} \left( \sum_{T\in\mathcal{T}_H} \|\nabla e\|^2_{L^2(N^{\ell+2}(T)\setminus N^{\ell-2}(T))} \right)^{1/2} \tag{4.18}$$

$$\lesssim \ell^{(d-1)/2}\exp(-c\ell)\|\nabla v_H\|_{L^2(D)} \|A^{1/2}\nabla(Q_H v - Q_H^{(\ell)} v)\|_{L^2(D)}, \tag{4.19}$$

since each element $T$ appears in $\ell^{d-1}$ rings $N^{\ell+2}(T)\setminus N^{\ell-2}(T))$ on a quasi uniform mesh $\mathcal{T}_H$. The theorem follows.

## 4.2 The Localized Orthogonal Decomposition method

The results of the previous section motivate the following localized variant of the method (3.13): Find $u_{H,(\ell)}^{\mathrm{ms}} \in V_{H,\ell}^{\mathrm{ms}}$ such that

$$a(u_{H,(\ell)}^{\mathrm{ms}}, v) = F(v) \tag{4.20}$$

for all $v \in V_{H,\ell}^{\mathrm{ms}}$. Another way of expressing this is: Find $u_{H,\ell} \in V_H$ such that

$$a((1+Q_H^{(\ell)})u_{H,(\ell)}, (1+Q_H^{(\ell)})v_H) = F((1+Q_H^{(\ell)})v_H) \tag{4.21}$$

for all $v_H \in V_H$. We have simply replaced the corrector $Q_H$ by its localized approximation based on the cell problems (4.8). The following theorem states that the results of Lemma 3.1 are widely preserved provided that the *oversampling* or *localization parameter* $\ell \approx |\log H|$.

**Theorem 4.2 (A priori error bound).** *The Localized Orthogonal Decomposition method* (4.20) *admits a unique solution* $u_{H,(\ell)}^{\mathrm{ms}} \in V_{H,\ell}^{\mathrm{ms}}$ *(for any* $F \in L^2(D)$*) and the error is bounded by*

$$\|\nabla(u - u_{H,(\ell)}^{\mathrm{ms}})\|_{L^2(D)} \le C(\alpha,\beta)(H + \ell^{(d-1)/2} exp(-c\ell))\|f\|_{L^2(D)}. \tag{4.22}$$

*Hence, the choice* $\ell \approx |\log H|$ *recovers the convergence rate of the ideal method.*

*Proof.* Let $v_H \in V_H$. From Lemma 4.1 we have

$$\|\nabla(Q_H - Q_H^{(\ell)})v_H\|_{L^2(D)} \lesssim \ell^{(d-1)/2}\exp(-c\ell)\|\nabla v_H\|_{L^2(D)}.$$

We let $u = u_H^{\mathrm{ms}} + w$ with $u_H^{\mathrm{ms}} \in V_H^{\mathrm{ms}}$ and $w \in W$. The ideal estimate, Theorem 3.1, shows that $\|\nabla w\|_{L^2(D)} \lesssim H\|f\|_{L^2(D)}$. Galerkin orthogonality gives $a(u - u_{H,(\ell)}^{\mathrm{ms}}, v) = 0$ for all $v \in V_{H,\ell}^{\mathrm{ms}}$. We apply this with $v := (1 + Q_H^{(\ell)})\mathcal{I}_H u$ to get,

$$\begin{aligned}
\|\nabla(u - u_{H,(\ell)}^{\mathrm{ms}})\|_{L^2(D)} &\lesssim \|A^{1/2}\nabla(u - u_{H,(\ell)}^{\mathrm{ms}})\|_{L^2(D)} \\
&\leq \|A^{1/2}\nabla(u - v)\|_{L^2(D)} \\
&\lesssim \|A^{1/2}\nabla w\|_{L^2(D)} + \|A^{1/2}\nabla(u_H^{\mathrm{ms}} - v)\|_{L^2(D)} \\
&\lesssim H\|f\|_{L^2(D)} + \|A^{1/2}\nabla((1 + Q_H - 1 + Q_H^{(\ell)})\mathcal{I}_H u)\|_{L^2(D)} \\
&\lesssim H\|f\|_{L^2(D)} + \ell^{(d-1)/2}\exp(-c\ell)\|\nabla\mathcal{I}_H u\|_{L^2(D)} \\
&\lesssim (H + \ell^{(d-1)/2}\exp(-c\ell))\|f\|_{L^2(D)}.
\end{aligned}$$

The final step towards a fully practical method regards the discretization of the cell problems (4.8). For any $T \in \mathcal{T}_H$ and any $\ell \in \mathbb{N}$, let $\mathcal{T}_h(\mathsf{N}^\ell(T))$ denote a regular mesh of the patch $\mathsf{N}^\ell(T)$ and let $V_h(\mathsf{N}^\ell(T))$ denote the corresponding finite element space that satisfies homogeneous Dirichlet boundary condition on $\partial\mathsf{N}^\ell(T)$. We assume that $\mathcal{T}_h(\mathsf{N}^\ell(T))$ is the result of $\log_2 \frac{H}{h}$ uniform refinements of $\mathcal{T}_H(\mathsf{N}^\ell(T))$. The restriction of $V_h(\mathsf{N}^\ell(T))$ to the space of fine scale functions results in the discrete approximation space

$$W_h(\mathsf{N}^\ell(T)) \subset W(\mathsf{N}^\ell(T))$$

for the numerical solution of the cell problems (4.8). Define *approximate localized correctors* $w_{T,i,h}^{(\ell)} \in W_{H,h}(\mathsf{N}^\ell(T))$ as unique solutions to the discrete cell problems

$$a_{\mathsf{N}^\ell(T)}(w_{T,i,h}^{(\ell)}, v_h) = -\int_T (Ae_i)\cdot\nabla v_h \,\mathrm{d}x, \quad \forall v_h \in W_h(\mathsf{N}^\ell(T)). \tag{4.23}$$

This leads to a further modification of the correction operator $Q_H$. For any $v_H \in V_H$, define

$$Q_{H,h}^{(\ell)} v_H := \sum_{T\in\mathcal{T}_h}\sum_{i=1}^{d}\left(\frac{\partial v_H|_T}{\partial x_i}\right) w_{T,i,h}^{(\ell)}, \tag{4.24}$$

where $w_{T,i,h}^{(\ell)}$ has been extended to $D$ by zero outside $\mathsf{N}^\ell(T)$. The practical quasi-local method then seeks $u_{H,(\ell)}^{\mathrm{ms},h} \in V_{H,\ell}^{\mathrm{ms},h} = \{(1 + Q_{H,h}^{(\ell)})v_H : v_H \in V_H\}$ such that

$$a(u_{H,(\ell)}^{\mathrm{ms},h}, v) = F(v) \tag{4.25}$$

for all $v_H \in V_{H,\ell}^{\mathrm{ms},h}$. Under the assumption that there is a global fine mesh $\mathcal{T}_h$ of the whole domain $D$ such that all local meshes $\mathcal{T}_h(\mathsf{N}^\ell(T))$ are submeshes of $\mathcal{T}_h$, Theorem 4.2 remains valid if we replace the solution $u$ by a reference solution $u_h \in V_h$ on the global fine mesh, i.e.,

$$a(u_h, v_h) = F(v_h)$$

for all $v_h \in V_h$. (Note that this reference solution is never computed.) Using standard arguments for Galerkin methods yields an error estimate for the method (4.25).

**Theorem 4.3.** *The practical quasi-local method* (4.25) *is well-posed and satisfies*

$$\|\nabla(u_h - u^{\mathrm{ms},h}_{H,(\ell)})\|_{L^2(D)} \leq C(H + \ell^{(d-1)/2} exp(-c\ell))\|f\|_{L^2(D)}.$$

*Moreover, for* $\ell \approx |\log H|$,

$$\|\nabla(u - u^{\mathrm{ms},h}_{H,(\ell)})\|_{L^2(D)} \leq C\Big(\|\nabla(u - u_h)\|_{L^2(D)} + H\|f\|_{L^2(D)}\Big).$$

*If* $h$ *is sufficiently small, this yields at least convergence of order* $O(H)$.

To illustrate the estimates we present a numerical experiment. Let $D$ be the unit square and let $f \equiv 1$. Consider the coefficient $A$ that is piecewise constant with respect to a uniform Cartesian grid of width $2^{-6}$. Its values are randomly chosen between 1 and 10; see Fig. **??**. We consider uniform meshes $\mathcal{T}_H$ of size $H = 2^{-1}, 2^{-2}, \ldots, 2^{-5}$ of $D$ that do not resolve the rough coefficient $A$ appropriately. The reference mesh $\mathcal{T}_h$ has width $h = 2^{-9}$. Since no analytical solution is available, the standard finite element approximation $u_h \in V_h$ on the reference mesh $\mathcal{T}_h$ serves as the reference solution. Doing this, we assume that $u_h$ is sufficiently accurate and, necessarily, that $\mathcal{T}_h$ resolves the discontinuities of $A$. The corrector problems are also solved on this scale of numerical resolution.

# Chapter 5
# Effective Coefficients and Connections to Periodic Homogenization

This chapter raises the question whether there is a link between the numerical homogenization method of this chapter and the mathematical theory of homogenization. Is it possible to reconstruct effective diffusion tensors from this method? The answer is somehow yes. This section shows that the modified bilinear form in (4.21) may be re-interpreted as an effective integral operator acting on finite element spaces. Under certain assumptions, it is even possible to link it to a partial differential operator with some effective diffusion tensor that is piecewise constant with respect to the coarse mesh $\mathcal{T}_H$; similar to Section 2.3.

## 5.1 Quasi-local effective coefficient

We re-interpret the left-hand side of (4.21) as a non-local operator acting on standard finite element functions. More precisely, we show the equivalence between a slight modification of (4.21) and some integral equation. The observations of this section apply as well to the fully discrete method of (4.25) but, for the sake of simplicity, the discretization of the corrector problems is disregarded in this section, i.e., $h = 0$.

In the first step, we trade the symmetry of the bilinear form in (4.21) for a slightly simpler method in terms of computation. The modified variant seeks $u_{H,\ell} \in V_H$ such that

$$a(u_{H,\ell}, (1 + Q_H^{(\ell)})v_H) = F(v_H) \tag{5.1}$$

for all $v_H \in V_H$. We have removed the corrector in the first argument of the bilinear form. Note that, in the ideal case, this has no effect due to the orthogonality of $W_H$ and $(1 + Q_H)V_H$. For finite $\ell$, the perturbation of the symmetric version can be controlled in terms of $e^{-c\ell}$. Moreover, the well-posedness of 5.1 can be ensured under the mild condition $\ell \gtrsim 1$.

Now, consider any $u_H, v_H \in V_H$. We have

$$a(u_H, (1 + Q_H^{\ell})v_H) = \int_D \nabla u_H \cdot (A \nabla v_H)\, dx + \int_D \nabla u_H \cdot (A Q_H^{\ell} \nabla v_H)\, dx.$$

The second term can be expanded with (4.9) as

$$\begin{aligned}
&\int_D \nabla u_H \cdot (A \nabla Q_H^\ell v_H)\,dx \\
&\quad = \sum_{T\in\mathcal{T}_H} \sum_{k=1}^{d} (\partial_k v_H|_T) \int_D \nabla u_H \cdot (A\nabla w_{T,k}^\ell)\,dx \\
&\quad = \sum_{K,T\in\mathcal{T}_H} \int_K \nabla u_H \cdot \left( \sum_{k=1}^{d} \fint_K (A(y)\nabla w_{T,k}^\ell(y))\,dy\, (\partial_k v_H|_T) \right) dx \\
&\quad = \sum_{K,T\in\mathcal{T}_H} |K||T|\, \nabla u_H|_K \cdot (\mathcal{K}_{T,K} \nabla v_H|_T)
\end{aligned}$$

for the matrix $\mathcal{K}_{T,K}^\ell$ defined for any $K, T \in \mathcal{T}_H$ by

$$(\mathcal{K}_{T,K}^\ell)_{j,k} := \frac{1}{|T||K|} e_j \cdot \int_K A \nabla w_{T,k}^\ell\,dx.$$

Define the piecewise constant matrix field over $\mathcal{T}_H \times \mathcal{T}_H$, for $T, K \in \mathcal{T}_H$ by

$$\mathcal{A}_H^\ell|_{T,K} := \frac{\delta_{T,K}}{|K|} \fint_T A\,dx + \mathcal{K}_{T,K}^\ell$$

(where $\delta$ is the Kronecker symbol) and the bilinear form $\mathfrak{a}^\ell$ on $V_H \times V_H$ by

$$\mathfrak{a}^\ell(v_H, z_H) := \int_D \int_D \nabla v_H(y) \cdot (\mathcal{A}_H^\ell(x,y) \nabla z_H(x))\,dy\,dx \quad \text{for any } v_H, z_H \in V_H.$$

We obtain for all $v_H, z_H \in V_H$ that

$$a(v_H, (1 + Q_H^\ell) z_H) = \mathfrak{a}^\ell(v_H, z_H). \tag{5.2}$$

*Remark 5.1 (notation).* For simplices $T, K \in \mathcal{T}_H$ with $x \in T$ and $y \in K$, we will sometimes write $\mathcal{K}^\ell(x,y)$ instead of $\mathcal{K}_{T,K}^\ell$ (with analogous notation for $\mathcal{A}^\ell$).

Next, we state the equivalence of two multiscale formulations.

**Proposition 5.1.** *A function $u_H^\ell \in V_H$ solves* (5.1) *if and only if it solves*

$$\mathfrak{a}^\ell(u_{H,\ell}, v_H) = F(v_H). \tag{5.3}$$

*Proof.* This follows directly from the representation (5.2).

*Remark 5.2.* For $d = 1$ and $I_H$ the standard nodal interpolation operator, the corrector problems localize to one element and the presented multiscale approach coincides with various known methods (homogenization, MSFEM). The resulting effective coefficient $\mathcal{A}_H^\ell$ is diagonal and, thus, local. This is no longer the case for $d \geq 2$.

## 5.2 Local effective coefficient in the periodic case

We shall now connect the quasi-local effective coefficient to the effective coefficient in the theory of (periodic) homogenization. The first step is to approximate the quasi-local coefficient by a local one. The exponential decay motivates to approximate the non-local bilinear form $\mathfrak{a}^\ell(\cdot,\cdot)$ by a quadrature-like procedure: Define the piecewise constant coefficient $A_H^\ell \in P_0(\mathcal{T}_H;\mathbb{R}^{d\times d})$ by

$$A_H^\ell|_T := \fint_T A\,dx - \sum_{K\in\mathcal{T}_H} |K|\mathcal{K}_{T,K}^\ell.$$

and the bilinear form $\tilde{a}^\ell$ on $V\times V$ by

$$\tilde{a}^\ell(u,v) := \int_D \nabla u\cdot(A_H^\ell\nabla v)\,dx.$$

*Remark 5.3.* In analogy to classical periodic homogenization, the local effective coefficient $A_H^\ell$ can be written as

$$\begin{aligned}(A_H^\ell)_{j,k}|_T &= |T|^{-1}\int_{N^\ell(T)} e_j\cdot(A(\chi_T e_k - \nabla w_{T,k}^\ell))\\ &= |T|^{-1}\int_{N^\ell(T)} (e_j-\nabla w_{T,j}^\ell)\cdot(A(\chi_T e_k - \nabla w_{T,k}^\ell))\end{aligned}$$

for the characteristic function $\chi_T$ of $T$ and the slightly enlarged averaging domain $N^\ell(T)$.

The localized multiscale method is to seek $\tilde{u}_{H,\ell}\in V_H$ such that

$$\tilde{a}^\ell(\tilde{u}_{H,\ell},v_H) = F(v_H)\quad \text{for all } v_H\in V_H. \tag{5.4}$$

The unique solvability of (5.4) is not guaranteed a priori in general. For non-periodic coefficients it must be checked a posteriori whether positive spectral bounds $\alpha_H,\beta_H$ on $A_H^\ell$ exist. Throughout this paper we assume that such bounds exist, that is, we assume that there exist positive numbers $\alpha_H,\beta_H$ such that

$$\alpha_H|\xi|^2 \le \xi\cdot(A_H^\ell(x)\xi) \le \beta_H|\xi|^2 \tag{5.5}$$

for all $\xi\in\mathbb{R}^d$ and almost all $x\in D$.

While the applicability of the local effective coefficient $A_H^\ell$ remains open in the general case, we shall justify the its use in the periodic setting. We show that the procedure in its idealized form with $\ell=\infty$ recovers the classical periodic homogenization limit. We denote by $V := H^1_\#(D)/\mathbb{R}$ the space of periodic $H^1$ functions with vanishing integral mean over $D$. We assume $D$ to be a polytope allowing for periodic boundary conditions. We adopt the notation of Chapters 3–(4), in particular $W_H\subseteq V$ is the kernel of the quasi-interpolation $I_H$, $V_H$ is the space of piecewise affine glob-

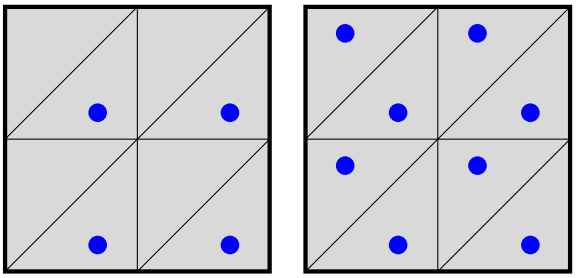


**Fig. 5.1** Periodic coefficients with respect to a square grid and triangulations: non-matching (left) and matching (right).

ally continuous functions of $V$, and $Q_H^\ell$, $a$, $\tilde{a}^\ell$, $\mathfrak{a}^\ell$, $\mathcal{A}_H^\ell$, $A_H^\ell$, $\mathcal{K}^\ell$ are defined as above. We assume that the domain $D$ matches with integer multiples of the period. We assume the triangulation $\mathcal{T}_H$ to match with the periodicity pattern. For simplicial partitions this implies further symmetry assumptions. In particular, periodicity with respect to a uniform rectangular grid is not sufficient. Instead we require further symmetry within the triangulated macro-cells, see Example 5.1 for an illustration. This property will be required in the proof of Propositon 5.2 below. In particular, not every periodic coefficient may meet this requierement. Also, generating such a triangulation requires knowledge about the length of the period.

*Example 5.1.* Figure 5.1 displays a periodic coefficient and a matching triangulation.

In the periodic setting, the following properties of $A_H^\ell$ can be derived. First, it is not difficult to prove that the coefficient $A_H^\ell$ is globally constant. The following result states that, in the idealized case $\ell = \infty$, the coefficient $A_H^\ell$ is even independent of the mesh-size $H$ and coincides with the classical homogenization limit, where for any $j = 1,\dots,d$, the corrector $\hat{q}_j \in H^1_\#(D)/\mathbb{R}$ is the solution to

$$\operatorname{div} A(\nabla \hat{q}_j - e_j) = 0 \text{ in } D \text{ with periodic boundary conditions.} \tag{5.6}$$

**Proposition 5.2.** *Let $A$ be periodic and let $\mathcal{T}_H$ be uniform and aligned with the periodicity pattern of $A$ and let $V$, $W$ be spaces with periodic boundary conditions. Then, for any $T \in \mathcal{T}_H$, the idealized coefficient $A_H^{(\infty)}|_T$ coincides with the homogenized coefficient from the classical homogenization theory. In particular, $A_H^{(\infty)}$ is globally constant and independent of $H$.*

*Proof.* Let $T \in \mathcal{T}_H$ and $j,k \in \{1,\dots,d\}$. The definitions of $A_H^{(\infty)}|_T$ and $\mathcal{K}^{(\infty)}$ lead to

$$\begin{aligned} \fint_T A_{jk}\,dx - (A_H^{(\infty)}|_T)_{jk} &= |T|^{-1} \sum_{K\in\mathcal{T}_H} \int_K e_j \cdot (A\nabla w_{T,k})\,dx \\ &= |T|^{-1} \int_D e_j \cdot (A\nabla w_{T,k})\,dx. \end{aligned} \tag{5.7}$$

The sum over all element correctors defined by $q_k := \sum_{T\in\mathcal{T}_H} w_{T,k}$ solves

$$a(w,q_k) = (\nabla w, Ae_k)_{L^2(D)} \quad \text{for all } w \in W. \tag{5.8}$$

The definitions of $w_{T,k}$ and $q_k$ and the symmetry of $A$ lead to

$$\begin{aligned} |T|^{-1}\int_D e_j\cdot(A\nabla w_{T,k})\,dx &= |T|^{-1}\int_D \nabla q_j\cdot(A\nabla w_{T,k})\,dx \\ &= \fint_T e_k\cdot(A\nabla q_j)\,dx. \end{aligned} \tag{5.9}$$

Let $v \in V$. We have $(v - I_H v) \in W$ and therefore by (5.8) that

$$\begin{aligned} \int_D \nabla v\cdot(A(\nabla q_j - e_j))\,dx &= \int_D (\nabla I_H v)\cdot(A(\nabla q_j - e_j))\,dx \\ &= \sum_{K\in\mathcal{T}_H}\int_K (\nabla I_H v)\,dx\cdot\fint_K A(\nabla q_j - e_j)\,dx \end{aligned}$$

where for the last identity it was used that $\nabla I_H v$ is constant on each element. By periodicity we have that $\fint_K A(\nabla q_j - e_j)\,dx = \fint_D A(\nabla q_j - e_j)\,dx$ for any $K \in \mathcal{T}_H$. Therefore, for all $v \in V$,

$$\int_D \nabla v\cdot(A(\nabla q_j - e_j))\,dx = \int_D (\nabla I_H v)\,dx\cdot\fint_D A(\nabla q_j - e_j)\,dx = 0$$

due to the periodic boundary conditions of $I_H v$. Hence, the difference $\nabla q_j - e_j$ satisfies (5.6). This is the corrector problem from classical homogenization theory and, thus, the proof is concluded by the above formulae (5.7)–(5.9). Indeed, by symmetry of $A$,

$$(A_H^{(\infty)}|_T)_{jk} = \fint_T A_{jk}\,dx - \fint_T e_k\cdot(A\nabla q_j)\,dx = \fint_T (e_j - \nabla q_j)\cdot Ae_k\,dx.$$

*Remark 5.4.* For Dirichlet boundary conditions, the method is different from the classical periodic homogenization as it takes the boundary conditions into account.

Unfortunately, a more precise discussion is beyond the scope of this lecture and we refer to [25] for the details, in particular quantified homogenization error estimates resulting from this theory and also error bounds beyond the periodic setting.

# Appendix A
# Functional analytic preliminaries

The discussion of well-posedness of PDEs as well as the analysis of variational discretization schemes strongly rely on tools from functional analysis. In this chapter, we will briefly recall some of these tools.

## A.1 Abstract linear spaces

### *A.1.1 Normed linear spaces and inner product spaces*

**Definition A.1 (vector space).** A set $X$ together with the mappings $+ : X \times X \to X$ and $\cdot : \mathbb{R} \times X \to X$ is called *(real) vector space*, if the following conditions are satisfied.

1. $(X,+)$ is a commutative group.
2. The scalar-vector-multiplication is associative, i.e., $\alpha(\beta x) = (\alpha\beta)x$ for all $\alpha,\beta \in \mathbb{R}$ and $x \in X$.
3. *Distributivity* holds in the sense that:

$$\alpha(x+y) = \alpha x + \alpha y \quad \text{and} \quad (\alpha+\beta)x = \alpha x + \beta x$$

for all $\alpha,\beta \in \mathbb{R}$ and $x,y \in X$.

**Definition A.2 (convexity).** A set $M \subseteq X$ is called *convex*, if for every $x,y \in M$ and for all $\lambda \in (0,1)$,

$$\lambda x + (1-\lambda)\, y \in M.$$

**Definition A.3 (subspace).** A set $M \subseteq X$ is a *subspace of $X$*, if it holds:

1. $0 \in M$ and
2. $\forall x,y \in M\ \forall \alpha \in \mathbb{R} \quad \alpha x + y \in M$.

The second property implies that supspaces are convex.

**Definition A.4 (scalar/inner product (spaces)).** Given a vector space $X$, a function

$$\langle \cdot, \cdot \rangle : X \times X \to \mathbb{R}$$

is called *scalar product*, if for all $x, y, z \in X$ and for all $\alpha \in \mathbb{R}$

1. $\langle \alpha x + y, z \rangle = \alpha \langle x, z \rangle + \langle y, z \rangle$ (linearity),
2. $\langle x, y \rangle = \langle y, x \rangle$ (symmetry),
3. $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0$ if and only if $x = 0$ (positive definiteness).

A vector space $X$ with scalar product $\langle \cdot, \cdot \rangle$ is called *pre-Hilbert space* or *inner product space* and is written as $(X, \langle \cdot, \cdot \rangle)$.

Inner product spaces allow a notion of orthogonality.

**Definition A.5 (orthogonality).** Two vectors $x, y \in X$ are *orthogonal* if

$$\langle x, y \rangle = 0.$$

One also writes $x \perp y$ or even $x \perp_X y$ with emphasis on the inner-product space.

*Example A.1 (inner product spaces).*

1. $\mathbb{R}^n$ with the standard Euclidean scalar product defined by

$$\langle x, y \rangle := x \cdot y := x^{\mathrm{T}} y \quad \text{for } x, y \in \mathbb{R}^n,$$

2. the space of quadratic summable sequences

$$\ell^2 := \left\{ \left(x_j\right)_{j \in \mathbb{N}} \in \mathbb{R}^{\mathbb{N}} \,\middle|\, \sum_{j=1}^{\infty} x_j^2 < \infty \right\}$$

   with scalar product

$$\langle (x_j), (y_j) \rangle := \sum_{j=1}^{\infty} x_j y_j \quad \text{for } (x_j), (y_j) \in \mathbb{R}^{\mathbb{N}},$$

3. The space of quadratic summable functions

$$L^2(D) := \left\{ f : D \to \mathbb{R} \,\middle|\, f \text{ Lebesgue measurable and } \int_D |f|^2 \, \mathrm{d}x < \infty \right\}$$

   with $L^2$-scalar product

$$\langle f, g \rangle := \int_D f(x) g(x) \, \mathrm{d}x \quad \text{for } f, g \in L^2(D).$$

The spaces $L^p$ or $\ell^p$ for $p \neq 2$ are not inner product spaces.

**Definition A.6 (norm, normed linear space).** Given a vector space $X$, a function

$$\|\cdot\| : X \to \mathbb{R}$$

is called *norm* if for all $x, y \in X$ and for all $\alpha \in \mathbb{R}$ it holds

1. $\|\alpha x\| = |\alpha| \|x\|$;
2. $\|x + y\| \leq \|x\| + \|y\|$;
3. $\|x\| = 0$ implies $x = 0$.

The pair $(X, \|\cdot\|)$ is called *normed linear space* (NLS).

A seminorm is a norm with the property (c) removed.

*Remark A.1.* For every inner product space $(X, \langle\cdot,\cdot\rangle)$, the function

$$\|\cdot\| : X \to [0, \infty), \quad x \mapsto \sqrt{\langle x, x\rangle}$$

defines a norm. This norm is called *induced by the scalar product* $\langle\cdot,\cdot\rangle$ or *the norm associated to the scalar product* $\langle\cdot,\cdot\rangle$. Hence, every inner product space canonically defines a normed linear space.

**Theorem A.1 (Cauchy-Schwarz inequality).** *Let $(X, \langle\cdot,\cdot,\rangle)$ be an inner product space and $\|\cdot\|$ be the induced norm. Then, for all $x, y \in X$, the inequality*

$$\langle x, y\rangle \leq \|x\| \, \|y\|$$

*holds. Equality,*

$$\langle x, y\rangle = \|x\| \, \|y\| ,$$

*holds if and only if $x = 0$ or $y = \lambda x$ for some $\lambda \geq 0$.*

**Theorem A.2 (Parallelogram equality).** *Let $(X, \langle\cdot,\cdot,\rangle)$ be an inner product space and $\|\cdot\|$ be the induced norm. For every $x, y \in X$ it holds*

$$\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2 . \tag{A.1}$$

*Conversely, if $\|\cdot\|$ is a norm that satisfies the parallelogram equality, then there exists a scalar product $\langle\cdot,\cdot\rangle$ which induces $\|\cdot\|$.*

*Remark A.2 (Inner Product Spaces in $\mathbb{R}^d$.).* The space $(\mathbb{R}^n, \langle\cdot,\cdot\rangle)$ is an inner product space, if and only if there is a symmetric positive definite matrix $A \in \mathbb{R}^{d\times d}$, such that for all $x, y \in \mathbb{R}^d$ it holds

$$\langle x, y\rangle = x^T A y.$$

### *A.1.2 Hilbert and Banach spaces*

Throughout this section, $(X, \|\cdot\|)$ is a normed linear space.

**Definition A.7 (Cauchy sequence).** A sequence $\left(x_j\right) \in X^{\mathbb{N}}$ is called *Cauchy sequence (CS) in X*, if for every $\varepsilon > 0$ exists an $n_\varepsilon \in \mathbb{N}$, such that for all $j,k \geq n\varepsilon$ it holds

$$\left\|x_j - x_k\right\| < \varepsilon.$$

**Definition A.8 (convergence, limit).** A sequence $\left(x_j\right) \in X^{\mathbb{N}}$ is called *convergent*, if there exists some $x \in X$, such that for every $\varepsilon > 0$ there exists some $n_\varepsilon \in \mathbb{N}$, such that for all $n \geq n_\varepsilon$ it holds

$$\|x - x_n\| < \varepsilon.$$

In this case, $x$ is called *limit* of $(x_j)$ and written as

$$x = \lim_{j\to\infty} x_j \quad \text{or} \quad (x_j) \to x\,.$$

*Remark A.3 (uniqueness of the limit).* The limit of a convergent sequence is unique.

**Definition A.9 (complete spaces).** A normed linear space $(X, \|\cdot\|)$ is *complete* if every Cauchy sequence $(x_j) \in X^{\mathbb{N}}$ in $X$ has a limit in $X$. A complete inner product space is called *Hilbert space* (HS). A complete normed linear space is called *Banach space* (BS).

*Remark A.4.* 1. Every Hilbert space is a Banach space.
2. Every inner product space can be completed to a Hilbert space and this extension is unique up to different names.
3. Every normed linear space can be completed to a Banach space and this extension is unique up to different names.
4. Any normed linear space (resp. inner product space) can be considered as dense subspaces of a Hilbert space (resp. Banach space).

**Definition A.10 (closed sets).** A set $M \subseteq X$ is called *closed*, if the limit of every convergent sequence $\left(x_j\right) \in M^{\mathbb{N}}$ is also in $M$.

**Definition A.11 (complete sets).** A set $M \subseteq X$ is called *complete*, if every Cauchy sequence in $M$ is convergent with a limit in $M$.

*Remark A.5.* In Banach spaces, every closed set is complete and vice versa.

### *A.1.3 Best approximation in Hilbert spaces*

In this subsection, we will introduce orthogonal projections onto convex subsets in Hilbert spaces. The following concepts and results will be used for studying the errors of finite element approximations in a very abstract form.

**Definition A.12 (distance and best approximation).** Let $(X, \|\cdot\|)$ be a normed linear space and $K \subset X$ be a nonvoid subset. Then for every $x \in X$ the *distance of x and K* is given by

$$\operatorname{dist}(x,K) := \operatorname{dist}_{\|\cdot\|}(x,K) := \inf_{y\in K} \|x-y\| \,.$$

The (possibly empty) set

$$\mathcal{P}_K(x) := \{y \in K \,|\, \|y-x\| = \operatorname{dist}(x,K)\}$$

is called *set of best approximations of $x$ in $K$* or *proxima of $x$ in $K$*.

**Theorem A.3.** *Given an inner product space $(X,\langle\cdot,\cdot\rangle)$ and a convex, nonvoid subset $K \subset X$, every $x \in X$ and $y \in K$ satisfy*

$$y \in \mathcal{P}_K(x) \quad \Longleftrightarrow \quad \forall z \in K,\ \langle x-y, z-y\rangle \le 0\,.$$

*Proof.* $\Longrightarrow$) Some elementary algebra with norms and scalar products shows for all $x,y,w \in X$ that

$$\begin{aligned}\|x-w\|^2 - \|x-y\|^2 \\ \|x-y+y-w\|^2 - \|x-y\|^2 &= \|x-y\|^2 + 2\langle x-y, y-w\rangle + \|y-w\|^2 - \|x-y\|^2 \\ &= \|y-w\|^2 - 2\langle x-y, w-y\rangle\,.\end{aligned}$$

For $y \in \mathcal{P}_K(x)$ and $w \in K$, the left-hand side is non-negative. Given any $z \in K$ and $0 < \lambda \le 1$, $w = \lambda z + (1-\lambda)y \in K$. Therefore

$$0 \le \|x-\lambda z-(1-\lambda)y\|^2 - \|x-y\|^2 = -2\lambda\langle x-y, z-y\rangle + \lambda^2\|y-z\|^2\,.$$

After divison by $\lambda > 0$, this results in

$$\langle x-y, z-y\rangle \le \frac{1}{2}\lambda\|y-z\|^2\,.$$

For $\lambda \searrow 0$ the right-hand side tends to zero. This proves the asserted inequality.
$\Longleftarrow$) For every $z \in K$, a Cauchy inequality shows

$$\begin{aligned}\|x-y\|^2 &= \langle x-y, x-z\rangle + \underbrace{\langle x-y, z-y\rangle}_{\le 0} \\ &\le \|x-y\|\,\|x-z\|\,.\end{aligned}$$

Consequently,

$$\|x-y\| \le \|x-z\| \quad \text{for all } z \in K,$$

and hence $y \in \mathcal{P}_K(x)$.

The following result provides uniqueness of the best approximation in Hilbert spaces.

**Theorem A.4 (Chebyshev property).** *Let $K \subset X$ be a nonvoid, closed, convex subset in a Hilbert space $(X,\langle\cdot,\cdot\rangle)$. Then, for every $x \in X$, the set of its best approximations*

$$\mathcal{P}_K(x) = \{P_K(x)\}$$

*contains exactly one element $P_K(x)$. The hereby defined mapping*

$$P_K : X \to K, \quad x \mapsto P_K(x)$$

*is idempotent (i.e., $P_K^2 := P_K \circ P_K = P_K$), non-expansive (i.e., $P_K$ is Lipschitz continuous with a Lipschitz constant $\leq 1$) and monotone (i.e., $0 \leq \langle P_K x - P_K y, x - y\rangle$ for all $x, y \in X$).*

*Proof. 1. Proof of uniqueness*. According to Theorem A.3 all $x \in X$ and $y_1, y_2 \in \mathcal{P}_K(x)$ satisfy

$$\langle x - y_1, y_2 - y_1\rangle \leq 0 \quad \text{and} \quad \langle x - y_2, y_1 - y_2\rangle \leq 0\,.$$

The sum of these inequalities gives

$$\|y_2 - y_1\|^2 \leq 0\,,$$

hence $y_1 = y_2$.

*2. Proof of existence*. $d := \operatorname{dist}(x, K)$ is the infimum of the set

$$\{\|x - y\| \mid y \in K\}$$

so there exists a sequence $(y_j)_{j\in\mathbb{N}} \subset K$ such that

$$d^2 \leq \left\|x - y_j\right\|^2 \leq d^2 + 1/j \quad \text{for any } j \in \mathbb{N}\,.$$

According to the parallelogram equality (A.1) on page 51, it holds, for $j, k \in \mathbb{N}$, that

$$\left\|y_j - y_k\right\|^2 + \left\|y_j + y_k - 2x\right\|^2 = 2\left\|y_j - x\right\|^2 + 2\|y_k - x\|^2\,.$$

The right-hand side is

$$\text{RHS} \leq 4d^2 + 2/j + 2/k.$$

Since $K$ is convex, $\frac{1}{2}(y_j + y_k) \in K$, and

$$\text{LHS} = \left\|y_j - y_k\right\|^2 + 4\left\|(y_j + y_k)/2 - x\right\|^2 \geq \left\|y_j - y_k\right\|^2 + 4d^2..$$

The previous estimates prove

$$\left\|y_j - y_k\right\|^2 \leq 2/j + 2/k\,.$$

Hence $(y_j)_{j\in\mathbb{N}}$ is a Cauchy sequence in the closed set $K$, thus converges towards a limit point $y \in K$ which satisfies

$$d \le \|x-y\| \le \left\|x-y_j\right\| + \left\|y-y_j\right\| \le \sqrt{d^2+1/j} + \left\|y-y_j\right\| .$$

For $j \longrightarrow \infty$, the right-hand side tends to $d$ and shows

$$\|x-y\| = d \text{ and so } y \in \mathcal{P}_K(x).$$

*3. Proof of idempotence*. Obviously $P_K(x) = x$ for $x \in K$. This proves the claimed $P_K^2(x) = P_K(x)$ for all $x \in X$.

*4. Proof of monotonicity*. Theorem A.3 shows for every $x, y \in X$ and their best approximations $P_K(x), P_K(y) \in K$ that

$$\begin{aligned} \langle x - P_K(x), P_K(y) - P_K(x) \rangle &\le 0 , \\ \langle y - P_K(y), P_K(x) - P_K(y) \rangle &\le 0 . \end{aligned}$$

The sum of these inequalities gives

$$\begin{aligned} \|P_K(y) - P_K(x)\|^2 &= \langle P_K(y) - P_K(x), P_K(y) - P_K(x) \rangle \\ &\le \langle y - x, P_K(y) - P_K(x) \rangle . \end{aligned} \tag{A.2}$$

This proves monotonicity of $P_K$.

*5. Proof of non-expansiveness*. The application of (A.2) and a Cauchy-Schwarz inequality lead to

$$\|P_K(y) - P_K(x)\|^2 \le \|y-x\| \, \|P_K(y) - P_K(x)\| .$$

Consequently,

$$\|P_K(y) - P_K(x)\| \le \|y-x\| ,$$

i.e., $P_K$ is non-expansive.

An immediate consequence of Theorem A.4 is the separation theorem in Hilbert spaces. Notice that this is a very special case of the famous separation principle in Banach spaces which follows from Hahn-Banach extension theorem.

**Corollary A.1.** *Let $K$ be some closed convex nonvoid set in the Hilbert space $X$ and $x \in X \setminus K$. Then there exist some direction $\ell \in X$ and some real numbers $\alpha$ and $\beta$ such that*

$$\langle \ell, y \rangle \le \alpha < \beta = \langle \ell, x \rangle \quad \textit{for all } y \in K .$$

*Proof.* Given the best approximation $z := P_K(x)$ of $x$ in $K$ there holds for any $y \in K$ that

$$\langle x - z, y - z \rangle \le 0 .$$

With $\ell := x - z \in X$ this is equivalent to

$$\langle \ell, y\rangle \le \langle \ell, z\rangle = \langle \ell, x\rangle - \|\ell\|^2 =: \alpha < \beta := \langle \ell, x\rangle$$

**Exercise A.1.** Prove that closed convex sets in Hilbert spaces are weakly closed.

The following corollary considers the case where the convex subset $K$ is a subspace and the best approximation is characterised by orthogonality.

**Corollary A.2.** *Let $M$ be a closed subspace of the Hilbert space $(X, \langle\cdot,\cdot\rangle)$. Then $P = P_M$ is a linear, continuous mapping onto $M$, and for all $x \in X$ it holds*

$$(x - P(x)) \perp M\,.$$

*Furthermore, $Q := 1 - P$ is a linear, continuous mapping onto the* orthogonal complement

$$M^\perp := \{x \in X \,|\, x \perp M\} \text{ of } M \text{ in } X\,.$$

*For any $x \in X$ there exist unique vectors*

$$p_x = P(x) \in M \quad \text{and} \quad q_x = Q(x) \in M^\perp \quad \text{with } x = p_x + q_x\,.$$

*Proof. Proof of orthogonality*. Given $x \in X$ and $z \in M$ set $p_x := Px$ and consider $w = p_x \pm z \in M$ in the characterising inequality of the best approximation. Then it holds

$$\langle x - p_x, w - p_x\rangle \le 0\,.$$

For $w := p_x + z$ this reads

$$\langle x - p_x, z\rangle \le 0.$$

For $w := p_x - z$ this reads

$$\langle x - p_x, -z\rangle \le 0.$$

Alltogether,

$$\langle x - p_z, z\rangle = 0\,.$$

Since $z$ is arbitrary, $q_x := x - p_x \perp M$.

*Proof of linearity.* Let $\alpha, \beta \in \mathbb{R}$ and $x, y \in X$ with $x = p_x + q_x$, $y = p_y + q_y$ where $p_x, p_y \in M$ and $q_x, q_y \in M^\perp$. Then we have

$$\alpha x + \beta y - (\alpha p_x + \beta p_y) = \alpha q_x + \beta q_y \in M^\perp\,.$$

Hence the characterisation of Theorem A.3 shows for $\alpha p_x + \beta p_y \in M$ that

$$\alpha p_x + \beta p_y \in \mathcal{P}_M(\alpha x + \beta y) = \{P_M(\alpha x + \beta y)\}\,,$$

which reads

$$\alpha P_M x + \beta P_M y = P_M(\alpha x + \beta y).$$

An important conclusion is that

$$X = M \oplus M^{\perp}$$

for any closed subspace $M$ of $X$ and its orthogonal complement $M^{\perp}$ and this is stable by orthogonality

$$\|x\|^2 = \|Px\|^2 + \|Qx\|^2 .$$

### *A.1.4 Dual spaces and Riesz representation*

**Definition A.13 (Dual space).** Given a normed linear space $X$, the vector space

$$X^{\star} := \{F : X \to \mathbb{R} \,|\, F \text{ linear and continuous}\}$$

with canonical addition and (outer) multiplication is called the *(continous) dual space of* $X$. Any $F \in X^{\star}$ has a norm

$$\|F\|_{X^{\star}} := \sup_{x \in X \setminus \{0\}} \frac{F(x)}{\|x\|_X} .$$

*Example A.2.* Let $x \in X$ be a vector of an inner product space $(X, \langle \cdot, \cdot \rangle)$ then, by a Cauchy inequality, the mapping

$$\langle x, \cdot \rangle : X \to \mathbb{R}, y \mapsto \langle x, y \rangle$$

is an element of the dual space $X^{\star}$. The Riesz representation theorem states that all linear functionals in a Hilbert space are of that form.

**Theorem A.5 (Riesz representation theorem).** *Let* $(X, \langle \cdot, \cdot \rangle)$ *be a Hilbert space. Then the Riesz mapping*

$$\mathcal{R} : X \to X^{\star}, x \mapsto \langle x, \cdot \rangle$$

*is a norm isomorphism. In particular, for every* $F \in X^{\star}$ *there exists a unique* $x =: \mathcal{R}^{-1} F \in X$ *with* $\langle x, \cdot \rangle = F$ *and* $\|x\|_X = \|F\|_{X^{\star}}$. *The vector* $x := \mathcal{R}^{-1} F$ *is called* Riesz representation *of* $F \in X^{\star}$.

*Proof. Proof of isometry*. The mapping $\mathcal{R}$ inherits its linearity from the scalar product. The Cauchy-Schwarz inequality and the discussion of the equality in there leads to the identity

$$\|\mathcal{R}x\|_{X^{\star}} = \sup_{y \in X, \|y\|_X = 1} |\langle x, y \rangle| = \|x\|_X \qquad \text{for all } x \in X.$$

Hence $\|\mathcal{R}\| = 1$ and $\mathcal{R}$ is an isometry.

*Proof of injectivity*. Since $\|\mathcal{R}x\|_{X^{\star}} = \|x\|_X$ for all $x$, $\mathcal{R}x = 0$ implies $x = 0$. Hence, $\mathcal{R}$ is injective.

*Proof of surjectivity*. Let $F \in X^\star \setminus \{0\}$ be a non-vanishing element of the dual space. Then the closed subspace $M = \ker(F)$ is a genuine subset of $X$ and hence $M^\perp \setminus \{0\} \neq \emptyset$. For a $z \in M^\perp \setminus \{0\}$ define

$$x = F(z)\, z / \|z\|^2 \in X\,.$$

Due to the linearity of $F$, every $y \in X$ satisfies

$$F(z)y - F(y)z \in M\,.$$

Finally the identity

$$\langle x, y\rangle \|z\|^2 = \langle F(z)z, y\rangle = \langle z, F(z)y\rangle = \langle z, F(y)z\rangle = F(y)\|z\|^2$$

proves

$$F = \langle x, \cdot\rangle.$$

## A.2 Lebesgue spaces and test functions

The content of this subsection on higher analysis is partly copied from the book of Evans [24] to which we refer for details, references, and proofs.

Lebesgue's measure theory provides a powerful integration theory in $\mathbb{R}^d$ and is preferred over the Riemann integral, since it provides certain "completeness" properties, i.e., appropriate limits of integrable functions are integrable, a property that the Riemann integral does not have. We recall a few basic facts and definitions.

**Definition A.14 (measurable sets).** The measurable subsets of $\mathbb{R}^d$ form the smallest countable additive $\sigma$ algebra that includes all open and closed sets. The Lebesgue measure $|M|$ of a measurable set $M \subset \mathbb{R}^d$ extends the $d$-dimensional volume of balls and each subset of a measurable set of measure zero is measurable and of measure zero.

**Definition A.15 (measurable functions).** A function $f : \mathbb{R}^d \to \mathbb{R}$ is called a measurable function if

$$f^{-1}(\omega) \quad \text{is a measurable set}$$

for every open subset $\omega \subset \mathbb{R}$.

The following result illustrates that measurable functions are continuous up to small sets.

**Theorem A.6 (Theorem of Lusin).** *Given a Lebesgue measurable set $D \subset \mathbb{R}^d$ with Lebesgue measure $|D| < \infty$ and a bounded and measurable function $f : D \to \mathbb{R}$, and $\varepsilon > 0$, there exists a compact subset $K \subset D$ with $|D \setminus K| < \varepsilon$ such that $f|_K \in C(K)$.*

**Definition A.16 (summable function).** A measurable function is summable in $D$, written $f \in L^1(D)$, if

$$\int_D |f| \, \mathrm{d}x < \infty.$$

A measurable function is locally summable, written as $f \in L^1_{\mathrm{loc}}(D)$, if it is summable on all compact subsets $\omega \subset\subset D$, i.e., $f|_\omega \in L^1(\omega)$.

**Definition A.17 (almost everywhere (a.e.)).** Two functions $f, g : D \to \mathbb{R}$ are said to be equal almost everywhere, written $f = g$ a.e., if the set $\{f \neq g\} := \{x \in D \,|\, f(x) \neq g(x)\}$ has measure zero, i.e.,

$$f = g \text{ a.e. if } |\{f \neq g\}| = 0\,.$$

In the context of Lebesgue functions, $f$ and $g$ are identified if they coincide almost everywhere.

We identify two functions $f$ and $g$ that satisfy $\|f - g\|_{L^p(D)} = 0$ and say $f = g$ almost everywhere (a.e.). For example, take $n = 1$, $D = (-1, 1)$ and functions

$$f(x) = \begin{cases} 1 \text{ for } x \geq 0, \\ 0 \text{ for } x < 0, \end{cases} \quad \text{and} \quad g(x) = \begin{cases} 1 \text{ for } x > 0, \\ 0 \text{ for } x \leq 0. \end{cases}$$

**Theorem A.7 (Lebesgue differentiation theorem).** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be locally summable.*

*1. Then for a.e. point $x_0 \in \mathbb{R}^d$,*

$$\lim_{r \to 0} |B(x_0, r)|^{-1} \int_{B(x_0, r)} f \, \mathrm{d}x = f(x_0)\,.$$

*2. In fact, almost every point $x_0 \in \mathbb{R}^d$ is a Lebesgue point of $f$, i.e.,*

$$\lim_{r \to 0} |B(x_0, r)|^{-1} \int_{B(x_0, r)} |f(x) - f(x_0)| \, \mathrm{d}x = 0.$$

Given an open subset $D \subset \mathbb{R}^d$ and $1 \leq p < \infty$, define

$$\|f\|_{L^p(D)} := \left( \int_D |f|^p \right)^{1/p},$$

and, for $p = \infty$, define

$$\begin{aligned} \|f\|_{L^\infty(D)} &:= \text{ess } sup f \\ &:= \inf \{\eta > 0 \,|\, |\{x \in D \,|\, |f(x)| > \eta\}| = 0\}\,. \end{aligned}$$

Then, for $p \in \mathbb{N} \cup \{\infty\}$, $\|\cdot\|_{L^p(D)}$ defines a semi-norm on the space

$$\widehat{L^p}(D) := \{f : D \to \mathbb{R} \text{ measurable } \mid \|f\|_{L^p(D)} < \infty\},$$

in particular it fulfills the triangle inequality.

**Theorem A.8 (Minkowski inequality).** *Assume* $1 \le p \le \infty$ *and* $u, v \in \widehat{L}^p(D)$*. Then it holds*

$$\|u+v\|_{L^p(D)} \le \|u\|_{L^p(D)} + \|v\|_{L^p(D)}\,.$$

To obtain a normed vector space we factorise $\widehat{L}(D)$ by the kernel of $\|\cdot\|_{L^p(D)}$,

$$\begin{aligned}\ker(\|\cdot\|_{L^p(D)}) &:= \{u \in \widehat{L}^p(D) \,|\, \|u\|_{L^p(D)} = 0\}\\ &= \{u \in \widehat{L}^p(D) \,|\, u = 0 \text{ a.e. in } D\},\end{aligned}$$

and define the space of Lebesgue functions as equivalence classes almost everywhere,

$$L^p(D) := \widehat{L}^p(D)/\ker(\|\cdot\|_{L^p(D)})\,.$$

Two Lebesgue functions coincide (i.e., they belong to the same class of Lebesgue functions) if they are equal almost everywhere.

**Theorem A.9 (Hölder inequality).** *Assume* $1 \le p, q \le \infty$*,* $1/p + 1/q = 1$*. Then for* $u \in L^p(D), v \in L^q(D)$*, it holds*

$$\|uv\|_{L^1(D)} \le \|u\|_{L^p(D)}\,\|v\|_{L^q(D)}\,.$$

If $p = q = 2$, the Hölder inequality is known as *Schwarz inequality* or *Cauchy-Schwarz inequality*.

A very important fact is the following theorem.

**Theorem A.10.** *For* $D \subseteq \mathbb{R}^d$ *open and* $1 \le p \le \infty$*,* $L^p(D)$ *is a Banach space.*

*Proof.* A proof employs the dominated convergence theorem and is left as an exercise.

Given any non-empty open set $D \subset \mathbb{R}^d$, recall

$$C^\infty(D) := \bigcap_{k \in \mathbb{N}o} C^k(D)$$

and let

$$\mathrm{d}(D) := C_c^\infty(D) := \{f \in C^\infty(\mathbb{R}^d) : \mathrm{supp} f \subset\subset D\}$$

denote the space of test functions. The support of $f$ is

$$\mathrm{supp} f = \overline{\{x \in \mathbb{R}^d \,|\, f(x) \neq 0\}} \tag{A.3}$$

and $\subset\subset$ denotes a compact subset. This means that $f \in \mathrm{d}(D)$ vanishes outside a big ball and also in a neighbourhood of the boundary.

**Theorem A.11.** *For* $1 \le p < \infty$ *it holds that*

$$\mathcal{D}(D) \text{ is dense in } L^p(D).$$

*Proof.* For a proof we refer to [60, Ch. 3, Thm. 3.14, p. 69].

*Remark A.6.* 1. $\mathcal{D}(D)\backslash\{0\}$ does not include any complex differentiable functions as they would be bounded and entire. The theorem of Liouville implies that this function is constant which leads to a contradiction.
2. For every domain $D \subset \mathbb{R}^d$ it holds $\mathcal{D}(D) \subset \mathcal{D}(\mathbb{R}^d)$.
3. For every domain $D \subset \mathbb{R}^d$, Theorem A.11 states that $\mathcal{D}(D)$ is dense in $L^2(D)$. Hence, $L^2$-functions have no boundary data.

*Example A.3.* Define functions $f$ and $g$ by

$$f(x) = \begin{cases} \exp(-1/x) & \text{for } x > 0, \\ 0 & \text{for } x \leq 0 \end{cases}$$

and

$$g(x) = f(x)\, f(1-x).$$

Then it holds $\operatorname{supp}(g) = [0,1]$ and hence $g \in \mathcal{D}(-1,2)$.

**Definition A.18.** For each $\varepsilon > 0$, define the standard mollifier $\eta_\varepsilon$ by

$$\eta_\varepsilon := C(d)/\varepsilon^d \begin{cases} \exp\left(\varepsilon^2/(|x|^2 - \varepsilon^2)\right) & \text{if } |x| < \varepsilon, \\ 0 & \text{if } |x| \geq \varepsilon. \end{cases} \tag{A.4}$$

The functions $\eta_\varepsilon$ are $C^\infty$ and satisfy

$$\int_{\mathbb{R}^d} \eta_\varepsilon \,\mathrm{d}x = 1 \quad \text{and} \quad \operatorname{supp}\eta_\varepsilon = \overline{B(0,\varepsilon)}.$$

We set

$$L^p_{\mathrm{loc}}(D) = \{f : D \to \mathbb{R} \text{ measurable, such that } f|_\omega \in L^p(\omega) \text{ for all } K \subset\subset D\}.$$

**Definition A.19.** If $f : D \to \mathbb{R}$ is locally integrable, i.e., $f \in L^1_{\mathrm{loc}}(D)$, define its mollification

$$f^\varepsilon := \eta_\varepsilon * f \quad \text{in} \quad D_\varepsilon := \{x \in D \,|\, \operatorname{dist}(x, \partial D) > \varepsilon\} \tag{A.5}$$

for any $x \in D_\varepsilon$, hence $B(x,\varepsilon) \subseteq D$, by

$$f^\varepsilon(x) = \int_D \eta_\varepsilon(x-y) f(y) \,\mathrm{d}y = \int_{B(0,\varepsilon)} \eta_\varepsilon(y) f(x-y) \,\mathrm{d}y.$$

**Theorem A.12 (Properties of mollifiers).** *It holds*

*1. If $f$ is locally integrable then $f^\varepsilon \in C^\infty(D_\varepsilon)$*
*2. If $f$ is in $L^1(D)$ then $f^\varepsilon \to f$ a.e. as $\varepsilon \to 0$*

3. *If $f \in C(D)$, then $f^\varepsilon \to f$ uniformly on compact subsets of $D$.*
4. *If $1 \le p < \infty$ and $f \in L^p_{loc}(D)$, then $f^\varepsilon \to f$ in $L^p_{loc}(D)$.*
5. *If $1 \le p < \infty, k \in \mathbb{N}_0$ and $f \in W^{k,p}_{loc}(D)$, then $f^\varepsilon \to f$ in $W^{k,p}_{loc}(D)$.*
6. *If $D^\alpha f$ is locally integrable for some derivative $D^\alpha$ of $f$ then*

$$\eta_\varepsilon * (D^\alpha f) = D^\alpha(\eta_\varepsilon * f).$$

7. *If $f : (a,b) \to \mathbb{R}$ is monoton, so is $\eta_\varepsilon * f$.*
8. *If $f : D \to \mathbb{R}$ is convex, so is $\eta_\varepsilon * f$.*
9. *If $f : B(0,1) \to \mathbb{R}$ is asymmetric, namely $f(x) = -f(-x)$, then $\eta_\varepsilon * f$ is asymmetric in $B(0, 1-\varepsilon)$.*

## A.3 Sobolev spaces

### *A.3.1 Weak derivatives and Sobolev functions*

**Definition A.20 (Weak derivative).** Suppose that $D \subseteq \mathbb{R}^d$ is open and $f \in L^1_{\ell oc}(D)$. A function $g_j \in L^1_{\text{loc}}(D)$ is called *weak derivative* of $f$ with respect to $x_j$ for $j \in \{1, ..., d\}$, if for all $\varphi \in \mathcal{D}(D)$ there holds

$$\int_D f \frac{\partial \varphi}{\partial x_j} dx = -\int_D g_j \varphi dx. \tag{A.6}$$

In this case we say that $f$ is *weakly differentiable* with respect to $x_j$ and set

$$\frac{\partial f}{\partial x_j} = g_j.$$

If all weak derivatives $\frac{\partial f}{\partial x_1}, ..., \frac{\partial f}{\partial x_d}$ exist, then we say that $f$ is weakly differentiable and define

$$\nabla f = \Big(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ..., \frac{\partial f}{\partial x_d}\Big),$$

the weak derivative of $f$.

**Example A.1** *Let $D = (-1, 1)$ and $f(x) := |x|$ for $x \in D$. Then $f$ is weakly differentiable with*

$$\nabla f = \begin{cases} +1 \text{ for } x > 0, \\ -1 \text{ for } x < 0. \end{cases}$$

*Proof.* Let $\varphi \in D(D)$. There holds by integration by parts

$$\begin{aligned}\int_D |x|\varphi'(x)\,dx &= -\int_{(-1,0)} x\varphi'(x)\,dx + \int_{(0,1)} x\varphi'(x)\,dx \\ &= \int_{(-1,0)} \varphi(x)\,dx - [x\varphi(x)]_{-1}^{0} - \int_{(0,1)} \varphi(x)\,dx + [x\varphi(x)]_{0}^{1} \\ &= \int_{(-1,0)} \varphi(x)\,dx - \int_{(0,1)} \varphi(x)\,dx \\ &= -\int_{(-1,1)} \operatorname{sgn}(x)\varphi(x)\,dx,\end{aligned}$$

where we used $\varphi(-1) = \varphi(1) = 0$.

**Lemma A.1 (Uniquness of the weak derivative).** *The weak derivative is (up to sets of measure zero) uniquely defined.*

*Proof.* If $g_j$ and $h_j$ are weak partial derivatives of $f \in L^1_{\mathrm{loc}}(D)$ with respect to $x_j$ then by (A.6) we get

$$\int_D (g_j - h_j)\varphi\,dx = 0 \quad \text{for all } \varphi \in D(D).$$

Owing to Theorem A.11 there holds $g_j = h_j$ almost everywhere in $D$.

**Lemma A.2.** *For $f \in C^1(\overline{D})$, the classical (strong) derivative and the weak derivative coincide (almost everywhere).*

*Proof.* Let us assume that $\partial D$ is sufficient regular so that Gauss' theorem holds, i.e.,

$$\int_D F\,dx = \int_{\partial D} F \cdot n\,ds \quad \text{for all } F \in C^1(\overline{D})^d.$$

Set $F = \varphi f e_j$, where $(e_j : j = 1,2,...,d)$ is the canonical basis of $\mathbb{R}^d$, i.e., the $j$-th component of $e_j$ equals 1 and and all other components are equal to 0. Since $\varphi|_{\partial D} = 0$, we have $\mathrm{div} F = \frac{\partial}{\partial x_j}(f\varphi) = \frac{\partial f}{\partial x_j}\varphi + f\frac{\partial \varphi}{\partial x_j}$ and $F|_{\partial D} = 0$. Therefore, we get

$$\int_D \frac{\partial f}{\partial x_j}\varphi + f\frac{\partial \varphi}{\partial x_j}\,dx = 0 \quad \text{for all } \varphi \in \mathcal{D}(D). \tag{A.7}$$

Hence $\frac{\partial f}{\partial x_j}$ is the weak partial derivative of $f$ with respect to $x_j$.

*Remark A.7.* Suppose that $D \subseteq \mathbb{R}^d$ is open and connected, and $f \in L^1_{\mathrm{loc}}(D)$ is weakly differentiable with $\nabla f = 0$. Then $f$ is constant.

*Remark A.8.* Lipschitz continuous functions are weakly differentiable.

**Example A.2** *The function $\log\big|\log|x|\big|$, $x \in B_{1/2}(0) \subseteq \mathbb{R}^d$, $d = 2,3$, has a singularity at $x = 0$, but is weakly differentiable.*

**Definition A.21.** Let $D \subseteq \mathbb{R}^d$ be open. A function $f : D \to \mathbb{R}$ is called *Sobolev function* if $f$ is weakly differentiable and if there exists $p$ with $1 \le p \le \infty$ such that $f \in L^p(D)$ and $\nabla f \in L^p(D)^d$.

We will use multiindeces to describe partial derivatives of any order. In more details, $\alpha = (\alpha_1, \ldots, \alpha_d), \alpha \in \mathbb{N}_0^d$ for $\alpha_1, \ldots, \alpha_d \in \mathbb{N}_0$, and

$$|\alpha| = \alpha_1 + \cdots + \alpha_d\,, \quad \alpha! = \alpha_1! \cdots \alpha_d!\,, \quad D^\alpha := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}\,.$$

The summation $\sum_{|\alpha|=0}^m$ means the sum over all such mutliindices with $|\alpha| = 0, 1, 2, \ldots, m$. For $m = 0$ this is only one, $\alpha = (0,0)$; for $m = 1$ this is $\alpha = (0,0), (1,0), (0,1)$ and for $m = 2$ this is $\alpha = (0,0), (1,0), (0,1), (2,0), (1,1), (0,2)$. Compare the related notation for the functional matrix $D$ and the Hessian $D^2$.

**Definition A.22 (Higher weak derivative).** Given a multiindex $\alpha = (\alpha_1, ..., \alpha_d) \in \mathbb{N}_0^d$, we define

$$D^\alpha \varphi = \frac{\partial^{|\alpha|} \varphi}{\partial x^\alpha} := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} ... \partial x_d^{\alpha_d}}$$

for $\varphi \in C^{|\alpha|}(D)$. We say that $f \in L^1_{\ell oc}$ possesses the *weak partial derivative* $\frac{\partial^\alpha}{\partial x^\alpha} f$, if there exists a function $g \in L^1_{\text{loc}}(D)$ such that

$$\int_D g\,\varphi\, dx = (-1)^{|\alpha|} \int_D f\, \frac{\partial^{|\alpha|} \varphi}{\partial x^\alpha}\, dx \quad \text{for all } \varphi \in \mathcal{D}(D).$$

In this case we set $\frac{\partial^\alpha f}{\partial x^\alpha} = g_\alpha$.

### *A.3.2 Sobolev spaces*

**Definition A.23.** Let $D \subseteq \mathbb{R}^d$ be open, $k$ a non-negative integer, and $f \in L^1_{\text{loc}}(D)$. Suppose that $f$ possesses weak partial derivatives $\frac{\partial^\alpha f}{\partial x^\alpha}$ for all $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \le k$. We define the *Sobolev norm*

$$\|f\|_{W^{k,p}(D)} = \Big( \sum_{\alpha \in \mathbb{N}_0^d, |\alpha| \le k} \Big\| \frac{\partial^{|\alpha|} f}{\partial x^\alpha} \Big\|^p_{L^p(D)} \Big)^{1/p} \quad \text{for } 1 \le p < \infty$$

and

$$\|f\|_{W^{k,p}(D)} = \max_{\alpha \in \mathbb{N}_0^d, |\alpha| \le k} \Big\| \frac{\partial^{|\alpha|} f}{\partial x^\alpha} \Big\|_{L^p(D)} \quad \text{for } p = \infty.$$

In either case we define the *Sobolev space* $W^{k,p}(D)$ as

$$W^{k,p}(D) = \{ f \in L^1_{\text{loc}}(D) : \|f\|_{W^{k,p}(D)} < \infty \}$$

and the *periodic Sobolev space*

$$W^{k,p}_{\#}(D) = \{f \in L^1_{\mathrm{loc}}(\mathbb{R}^d) : \ f \in W^{k,p}(D),\ f \text{ periodic w.r.t } D\}$$

**Theorem A.13.** *For $D \subseteq \mathbb{R}^d$ open, $k \in \mathbb{N}_0$, and $1 \le p \le \infty$ the Sobolev space $W^{k,p}(D)$ is a Banach space.*

*Proof.* Exercise.

*Remark A.9 (Notation).* For $k \in \mathbb{N}$, it is customary to write

$$H^k(D) = W^{k,2}(D).$$

**Theorem A.14.** *For $k \in \mathbb{N}$, the bilinear form $\langle \cdot, \cdot \rangle_{H^k(D)} : H^k(D) \times H^k(D) \to \mathbb{R}$ given by*

$$\langle u, v \rangle_{H^k(D)} := \sum_{\alpha \in \mathbb{N}_0^d, |\alpha| \le k} \left\langle \frac{\partial^{|\alpha|} u}{\partial x^\alpha}, \frac{\partial^{|\alpha|} v}{\partial x^\alpha} \right\rangle$$

*for $u, v \in H^k(D)$ defines a scalar product. The spaces $H^k(D)$ are a Hilbert spaces.*

*Proof.* Exercise.

The following theorem is due to Meyers and allows for an alternative definition of $W^{k,p}(D)$ in case $1 \le p < \infty$ (cf. Remark A.15 below).

**Theorem A.15.** *Let $D \subseteq \mathbb{R}^d$ be an open set and $1 \le p < \infty$. Then $C^\infty(D) \cap W^{k,p}(D)$ is dense in $W^{k,p}(D)$, i.e., given any $f \in W^{k,p}(D)$ and $\varepsilon > 0$, there exists $g \in C^\infty(D)$ such that*

$$\|f - g\|_{W^{k,p}(D)} < \varepsilon.$$

*Proof.* See [24].

We will frequently make use of the following *semi-norms*: For $f \in W^{k,p}(D)$ set

$$|f|_{W^{k,p}(D)} = \Big( \sum_{\alpha \in \mathbb{N}_0^d, |\alpha| = k} \Big\| \frac{\partial^{|\alpha|} f}{\partial x^\alpha} \Big\|^p_{L^p(D)} \Big)^{1/p} \quad \text{for } 1 \le p < \infty$$

and

$$|f|_{W^{k,p}(D)} = \max_{\alpha \in \mathbb{N}_0^d, |\alpha| = k} \Big\| \frac{\partial^{|\alpha|} f}{\partial x^\alpha} \Big\|_{L^p(D)} \quad \text{for } p = \infty.$$

### *A.3.3 Lipschitz domains and integration by parts*

**Definition A.24.** A set $D \subseteq \mathbb{R}^d$ is called *Lipschitz domain*, if it is open and connected and if for each $x \in \partial D$ there exists a coordinate transformation $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ (i.e.,

$\Phi(y) = Ay + z$ with $A \in \mathbb{R}^{d\times d}$ orthogonal and $z \in \mathbb{R}^d$), some parameter $\delta > 0$, and a Lipschitz continuous function $\eta : [-\delta,\delta]^{d-1} \to \mathbb{R}$ such that

$$\begin{aligned} D \cap B_\delta(x) &= \Phi(\{(y_1,...,y_d) \in \mathbb{R}^d : \eta(y_1,...,y_{d-1}) > y_d\}) \cap B_\delta(x), \\ \partial D \cap B_\delta(x) &= \Phi(\{(y_1,...,y_d) \in \mathbb{R}^d : \eta(y_1,...,y_{d-1}) = y_d\}) \cap B_\delta(x), \\ B_\delta \backslash \overline{D}(x) &= \Phi(\{(y_1,...,y_d) \in \mathbb{R}^d : \eta(y_1,...,y_{d-1}) < y_d\}) \cap B_\delta(x). \end{aligned}$$

Roughly speaking, Lipschitz domains are open and connected sets, whose boundary is locally parameterized by a Lipschitz continuous function and which lies on one side of its boundary.
Lipschitz domains allow for the following integration by parts formula with boundary terms:

**Theorem A.16 (Integration by parts).** *For a bounded Lipschitz domain $D \subseteq \mathbb{R}^d$ and functions $f, g \in C^1(D) \cap C(\overline{D})$ there holds*

$$\int_D \Big(\frac{\partial f}{\partial x_j} g + f \frac{\partial g}{\partial x_j}\Big) dx = \int_{\partial D} f g n_j \, ds \quad \text{for } 1 \le j \le d. \tag{A.8}$$

*Here $n_j$ is the $j$-th component of the outer unit normal $n$ to $\partial D$.*

*Proof.* We refer to standard literature, e.g. [48, Theorem3.34], for a proof of Theorem A.16 and only mention that it is based on the one-dimensional integration by parts formula.

*Remark A.10.* The outer unit normal $n$ on $\partial D$ is only defined almost everywhere on $\partial D$ (with respect to the surface measure). This however is sufficient to define the integral on the right-hand side of (A.8).

*Remark A.11.* Suppose that $D \subseteq \mathbb{R}^d$ is a bounded Lipschitz domain. Then $W^{1,\infty}(D) = Lip(D) = \{f \in C(D) : \ f \text{ is Lipschitz continuous in } D\}$ in the sense that $f \in W^{1,\infty}(D)$ if and only if there exists $f^* \in Lip(D)$ such that $f = f^*$ almost everywhere.

### *A.3.4 Traces of Sobolev functions*

We want to extend the integration by parts formula (A.8) to functions $f \in W^{1,p}(D)$. Therefore, we need to understand in which sense Sobolev functions have well defined boundary values. Recall that it does not make sense to talk about boundary values for functions in $L^p(D)$.

Throughout this subsection, we assume that $D$ is a bounded Lipschitz domain in $\mathbb{R}^d$.

**Definition A.25 (Trace operator).** Suppose $1 \le p \le \infty$. The *trace operator* $\gamma$ is defined for $f \in W^{1,p}(D)$ and $x \in \partial D$ as

$$(\gamma f)(x) := \begin{cases} \lim_{r\to 0} |D \cap B_r(x)|^{-1} \int_{D\cap B_r(x)} f(y)\,dy & \text{if this limit exists,} \\ 0 & \text{otherwise.} \end{cases} \tag{A.9}$$

*Remark A.12.*

1. We interpret the trace $\gamma f$ of $f$ as the boundary values of $f$. For $f \in C(\overline{D}) \cap W^{1,p}(D)$ and $x \in \partial D$, there holds $(\gamma f)(x) = f(x)$.
2. The trace operator extends the restriction $\cdot|_{\partial D} : f \mapsto f|_{\partial D}$ from $C(\overline{D}) \cap W^{1,p}(D)$ to $W^{1,p}(D)$.

**Theorem A.17 (Bounded traces).** *For $1 \le p \le \infty$ the trace operator $\gamma$ defines a bounded, linear mapping $\gamma : W^{1,p}(D) \to L^p(D)$, i.e., $\gamma$ is linear and $\|\gamma f\|_{L^p(\partial D)} \le C_\gamma \|f\|_{W^{1,p}(D)}$ for $C_\gamma > 0$ and all $f \in W^{1,p}(D)$.*

*Remark A.13.* The traces of $H^1(D)$-functions are dense in $L^2(\partial D)$.

**Theorem A.18 (Generalized integration by parts).** *Suppose $f \in C^1(D) \cap C(\overline{D})$ and $u \in W^{1,p}(D)$ for $1 \le p \le \infty$. Then*

$$\int_D u \frac{\partial f}{\partial x_j}\,dx + \int_D \frac{\partial u}{\partial x_j} f\,dx = \int_{\partial D} f n_j \gamma(u)\,ds. \tag{A.10}$$

*If $F \in C^1(D)^d \cap C(\overline{D})^d$ and $v \in W^{1,p}(D)$*

$$\int_D v \operatorname{div} F\,dx + \int_D \nabla v \cdot F\,dx = \int_{\partial D} F \dot{n} \gamma(v)\,ds. \tag{A.11}$$

In order to deal with Dirichlet type boundary conditions on some (closed) part $\Gamma_0$ of $\partial D$ in the Poisson problem, it is useful to define subspaces of Sobolev spaces for which functions vanish on $\Gamma_0$.

**Definition A.26.** Let $\Gamma_0$ be a closed subset of $\partial D$. Define

$$W^{1,p}_D(D) = \{f \in W^{1,p}(D) : \gamma f|_{\Gamma_0} = 0\}.$$

If $p = 2$ we may write $H^1_D(D) = W^{1,2}_D(D)$.

*Remark A.14 (Notation).*

1. If $\Gamma_0 = \partial D$ we also write $W^{1,p}_0(D)$ instead of $W^{1,p}_D(D)$.
2. We sometimes only write $f|_{\Gamma_0}$ instead of $(\gamma f)|_{\Gamma_0}$.

*Remark A.15 (Sobolev Spaces by Density).* An alternative way of defining the Sobolev spaces reads as follows. Let

$$W^{m,p}(\mathbb{R}^d) := \overline{\mathcal{D}(\mathbb{R}^d)}^{\|\cdot\|_{W^{m,p}(\mathbb{R}^d)}}$$

denote the completion of the normed linear space $\mathcal{D}(\mathbb{R}^d)$ endowed with the Sobolev norm. The restriction to $D$ leads to an equivalent definition of Sobolev spaces

$$W^{m,p}(D) = \{f|_D \,|\, f \in W^{m,p}(\mathbb{R}^d)\}.$$

The completion of $\mathcal{D}(D)$ with respect to Sobolex norms yields spaces

$$W_0^{m,p}(D) := \overline{\mathcal{D}(D)}^{\|\cdot\|_{W^{m,p}(D)}}. \tag{A.12}$$

Clearly, $W_0^{m,p}(D) \subseteq W^{m,p}(D)$. Note that the definitions of $W_0^{1,p}(D)$ in (A.12) and Definition A.26 for $\Gamma_0 = \partial D$ are equivalent.

### *A.3.5 Important theorems*

We have seen that Sobolev functions are special Lebesgue functions, at least, they are not more general measures or other strange objects. They can be identified by a function, namely the precise representation. This lecture series will make use of a few properties of Sobolev functions only and leaves the technical proofs to the PDE literature.

Throughout this section, we assume that $D$ is a bounded Lipschitz domain in $\mathbb{R}^d$.

If a Sobolev function is sufficiently often weakly differentiable then it equals a continuous function almost everywhere.

**Theorem A.19 (Sobolev embeddings).** *Let $D \subset \mathbb{R}^d$ be a Lipschitz domain and $k, p, d \in \mathbb{N}$. If $kp > d$, then there exists a continuous embedding $W^{k,p}(D) \hookrightarrow C(\overline{D})$. If $kp < d$, then there exists a continuous embedding $W^{k,p}(D) \hookrightarrow L^{dp/(d-p)}(\overline{D})$.*

For detailled proofs of these embeddings, we refer to [26, Sec. 7.7.].

**Theorem A.20 (Stein Extension Theorem).** *Given a bounded Lipschitz domain $D \subset\subset \widehat{D}$ compactly included in a bounded open set $\widehat{D}$, there exists a bounded linear extension operator*

$$E : W^{1,p}(D) \to W^{1,p}(\widehat{D})$$

*such that, for each $u \in W^{1,p}(D)$,*

1. *$Eu = u$ a.e. in $D$;*
2. *$\operatorname{supp}(Eu) \subset \widehat{D}$;*
3. *$\|Eu\|_{W^{1,p}(\mathbb{R}^d)} \le C(p, D, \widehat{D}) \|u\|_{W^{1,p}(D)}$.*

*Proof.* We refer to [63] (or [24]) for a proof.

**Theorem A.21 (Rellich-Kondrachov Compactness Theorem).** *Assume $D$ is a bounded open subset of $\mathbb{R}^d$, with a Lipschitz boundary $\partial D$. Suppose $1 \le p < n$. Then*

$$W^{1,p}(D) \overset{c}{\hookrightarrow} L^q(D)$$

*for each $1 \le q < p^* := pd/(d-p)$.*

*Proof.* For a proof we refer to [24].

This compactness result gives rise to the following inequality.

**Theorem A.22 (Poincaré inequality).** *Given a bounded, connected and open subset of $D \subset \mathbb{R}^d$ with a Lipschitz boundary $\partial D$, and $1 \le p \le \infty$. Then there exists a constant $C(d,p,D) < \infty$ with*

$$\|f - |D|^{-1} \int_D f \,\mathrm{d}x\|_{L^p(D)} \le C_P(d,p,D) \|Df\|_{L^p(D)}$$

*for every $f \in W^{1,p}(D)$.*

*Proof.* For a proof we refer to [24].

A similar result can be established for function that vanishes at some part of the boundary; the $L^p$-norm of functions in $W^{1,p}_D(D)$ can be bounded by the $L^p$-norm of their weak gradients.

**Theorem A.23 (Friedrichs inequality).** *Given a bounded, connected and open subset of $D \subset \mathbb{R}^d$ with a Lipschitz boundary $\partial D$, and $\Gamma_0 \subset \partial D$ with positive surface measure $|\Gamma_0|$ and $1 \le p \le \infty$. Then there exists a constant $C(d,p,\Gamma_0,D) < \infty$ with*

$$\|f\|_{L^p(D)} \le C_F(d,p,\Gamma_0,D) \|Df\|_{L^p(D)}$$

*for every $f \in W^{1,p}_{\Gamma_0}(D) := \{u \in W^{1,p}(D) \,|\, u = 0 \text{ on } \Gamma_0\}$.*

*Proof.* For a proof we refer to [11] where this inequality is called Poincaré inequality.

In two particular situations for $p = 2$, the constants $C_P$ and $C_F$ can be estimated exampleicitly.

**Theorem A.24 (Payne-Weinberger).** *Given a convex bounded open set $D \subseteq \mathbb{R}^d$ of diameter* $\operatorname{diam}(D) := \sup\{|x-y| \,|\, x,y \in D\}$ *it holds*

$$C_P(n,2,D) \le \operatorname{diam}(D)/\pi.$$

*In other words,*

$$\left\| f - |D|^{-1} \int_D f \,\mathrm{d}x \right\|_{L^2(D)} \le \operatorname{diam}(D)/\pi \|Df\|_{L^2(D)}$$

*for any $f \in H^1(D)$.*

*Proof.* The original proof in [57] relies on a weighted one-dimensional estimate plus a nice intersection argument. The given application for $d \ge 3$ contains some mistake which can be removed [8]. The assumption holds for all $d \ge 1$ and is sharp in the sense that the constant cannot be better under the assumption to have $D$ arbitrary convex.

**Theorem A.25 (Friedrichs inequality in $H_0^1(D)$).** *For $\Gamma_0 = \partial D$ it holds $C(d,2,\partial D, D) \leq \operatorname{width}(D)/\pi$ for the size*

$$\operatorname{width}(D) := L := \beta - \alpha$$

*defined as the smallest length $L := \beta - \alpha$ such that the open set $D$ lies between two parallel hyperplanes $\{x \cdot \nu = \alpha\}$ and $\{x \cdot \nu = \beta\}$ of distance $L$ for some unit vector $\nu \in \mathbb{R}^d$. In other words,*

$$\|f\|_{L^2(D)} \leq \operatorname{width}(D)/\pi \|Df\|_{L^2(D)}$$

*for all $f \in H_0^1(D)$.*

*Proof.* Without loss of generality, let the coordinate system be with $\nu = (0,\dots,0,1)$ and

$$D \subseteq \widehat{D} := \{(\xi, x_n) \in \mathbb{R}^d \,|\, 0 < x_n < L\}.$$

Any test function $f \in \mathcal{D}(D)$ is extended by zero to $\widehat{D}$. For any $\xi \in \mathbb{R}^{d-1}$, the partial function

$$f(\xi,\cdot) : (0,L) \to \mathbb{R}$$

belongs to $\mathcal{D}(0,L) \subseteq H_0^1(0,L)$ and the one-dimensional Friedrichs inequality results in

$$\int_0^L |f(\xi,x_n)|^2 \,\mathrm{d}x_n \leq (L/\pi)^2 \int_0^L |\frac{\partial f}{\partial x_n}(\xi,x_n)|^2 \,\mathrm{d}x_d.$$

An integration with respect to $\xi \in \mathbb{R}^d$ and $\left|\frac{\partial f}{\partial x_n}\right| \leq |Df|$ lead to

$$\int_{\widehat{D}} |f(x)|^2 \,\mathrm{d}x \leq (L/\pi)^2 \int_{\widehat{D}} |Df(x)|^2 \,\mathrm{d}x$$

for any $f \in \mathcal{D}(D)$. A density argument with $\overline{\mathcal{D}(D)}^{\|\cdot\|_{H^1(D)}}$ proves the assumption.

## A.4 Well-posedness of linear problems

The analysis and the computation of PDEs is based on a weak (or ultra-weak) form which involves some bilinear form

$$b : X \times Y \to \mathbb{R} \tag{A.13}$$

on some real-valued Sobolev spaces $X$ and $Y$ which are reflexive Banach spaces or even Hilbert spaces. Recall that a Banach space is reflexive if it can be identified with its bidual $X^{**} := (X^*)^*$ (Hilbert spaces are reflexive by Riesz' representation theorem A.5).

While $X = Y$ for many simple elliptic second order PDEs (e.g. the Poission problem), the Banach spaces $X$ and $Y$ may be very different in other circumstances (e.g. for ultra weak formulations).

This section discusses the general well-posedness of linear problems of the primal form

$$b(x,\cdot) = F \tag{A.14}$$

or the dual form

$$b(\cdot,y) = G. \tag{A.15}$$

The point is that, given $F \in Y^*$ in the dual $Y^*$ of $Y$ (resp. given $G \in X^*$ in the dual of $X$) there exists some unique solution $x \in X$ (resp. $y \in Y$) of the primal problem (A.14) (resp. the dual problem (A.15)).

Besides the unique solvability of the two problems (A.14) and (A.15), the perturbation analysis is relevant. Well-posedness means that the solution $x$ of (A.14) (resp. $y$ of (A.15)) depends continuously on the right-hand side $F \in Y^*$ (resp. $G \in X^*$). It will be a consequence of the fundamental properties of linear operators between Banach spaces that unique solvability of (A.14) (resp. (A.15)) readily implies the well-posedness, the unique solvability of the primal problem is equivalent to the unique sovability of the dual problem, and all this is equivalent to the inf-sup conditions

$$0 < \alpha := \inf_{x\in X\setminus\{0\}} \frac{\|b(x,\cdot)\|_{Y^*}}{\|x\|_X} = \inf_{y\in Y\setminus\{0\}} \frac{\|b(\cdot,y)\|_{X^*}}{\|y\|_Y}. \tag{A.16}$$

To illustrate this important condition, suppose for the moment that the bounded linear operator

$$B_1 : X \to Y^*, \quad x \mapsto b(x,\cdot) \tag{A.17}$$

is continously invertable. In other words, the linear operator

$$B_1^{-1} : Y^* \to X$$

is bounded. The operator norm of $B_1^{-1}$ reads

$$\|B_1^{-1}\| = \sup_{F\in Y^*\setminus\{0\}} \frac{\|B_1^{-1}F\|_X}{\|F\|_{Y^*}}.$$

Given the solution $x = B_1^{-1}F$ of (A.14),

$$\|B_1^{-1}F\|_X = \|x\|_X \ \text{ and } \ \|F\|_{Y^*} = \|B_1 x\|_{Y^*}.$$

Since all $F \in Y^*$ can be written in this form, it follows

$$\frac{1}{\|B_1^{-1}\|} = \inf_{F\in Y^*\setminus\{0\}} \frac{\|F\|_{Y^*}}{\|B_1 F\|_X} = \inf_{x\in X\setminus\{0\}} \frac{\|B_1 x\|_{Y^*}}{\|x\|_X} = \alpha.$$

In other words, the inf-sup constant (A.16) equals the reciprocal of the norm of $B_1^{-1}$. The linear operator

$$B_2 : Y \to X^*, \quad y \mapsto b(\cdot,y) \tag{A.18}$$

is the dual of $B_1$ for reflexive Banach spaces where $X$ and $Y$ are identified with their respective bidual spaces $X^{**}$ and $Y^{**}$.

In fact, the dual operator $B_1^* : Y^{**} \to X^*$ of $B_1$ is defined by

$$B_1^* : Y^{**} \to X^*, \quad \Lambda \mapsto \Lambda \circ B_1$$

via the composition $\Lambda \circ B_1 : X \to \mathbb{R}$, $x \mapsto \Lambda(B_1 x)$ which maps any $\Lambda \in Y^{**}$ (this is a bounded linear functional $\Lambda : Y^* \to \mathbb{R}$) onto its value at $B_1 x = b(x, \cdot) \in Y^*$. The identification of $Y$ with $Y^{**}$ can be written as the evaluation functional $\delta_y$ at some $y \in Y$, i.e.,

$$\Lambda(F) \equiv \delta_y(F) := F(y) \quad \text{for any } F \in Y^{**}.$$

For any $y \in Y$, $\delta_y$ belongs to $Y^{**}$. For a reflexive Banach space $Y$, those evaluation functionals describe all elements in $Y^{**}$, i.e., the mapping

$$\delta : Y \to Y^{**}, \quad y \mapsto \delta_y.$$

is surjective. This implies, for all $x \in X$, that

$$(B_1^*(\delta_y))(x) = \delta_y(B_1 x) = \delta_y(b(x, \cdot)) = b(x, y) = (B_2(y))(x).$$

Since $x \in X$ is arbitrary, this reads $B_1^* \delta_y = B_2 y$. The identification $Y = Y^{**}$ and the aforementioned calculations allow the notation

$$B_1^* : Y \to X^*, \quad y \mapsto b(y, \cdot)$$

and hence $B_1^* = B_2$. The same argument for $X = X^{**}$ shows $B_2^* = B_1$. This is behind the equality in (A.16), namely

$$\|B_1^{-1}\| = \|(B_1^*)^{-1}\| = \|B_2^{-1}\| = \alpha^{-1}.$$

We summarize the previous discussion in the subsequent theorem.

**Theorem A.26 ().** *Let $X$ and $Y$ reflexive Banach spaces and let $b : X \times Y \to \mathbb{R}$ be a bounded bilinear form with $X$ and $Y$ as above. Then the following conditions are pairwise equivalent:*

*(a)* $\forall F \in Y^* \exists! x \in X,\ b(x, \cdot) = F$;
*(b)* $\forall G \in X^* \exists! y \in Y,\ b(\cdot, y) = F$;
*(c) The infsup conditions* (A.16) *are satisfied.*

For a complete proof of the theorem, we refer to classical textbooks in Functional Analysis. An important special case for the present lecture is when $X = Y$ and for some Hilbert space $X$.

**Definition A.27 (Ellipticity).** Some bilinear form $a : X \times X \to \mathbb{R}$ is called $X$-elliptic if there exists $\alpha > 0$ such that, for all $v \in X$, it holds

$$\alpha \|v\|_X^2 \leq a(x, x).$$

For an $X$-elliptic bilinear form $a$, Theorem A.26 readily yields a famous result due to Lax and Milgram which plays a dominant role in the existence theory of elliptic PDEs.

**Corollary A.3 (Lax-Milgram theorem).** *Suppose $X$ is a Hilbert space, $F \in X^*$ and $a : X \times X \to \mathbb{R}$ is an elliptic continuous bilinear form. Then there is a unique $u \in X$ such that*

$$a(u, v) = F(v) \quad \text{for all } v \in X. \tag{A.19}$$

*Moreover,*

$$\|u\|_X \leq \frac{1}{\alpha} \|F\|_{X^*}.$$

The difference between the Lax-Milgram theorem and Riesz' representation theorem is that $a$ does not need to be symmetric in the Lax-Milgram theorem. For the sake of completeness, we present a proof below.

*Proof.* For all $v \in X$ we know that $a(v, \cdot) \in X^*$. Hence, by Riesz' representation theorem, there exists $Av = \mathcal{R}^{-1} a(v, \cdot) \in X$ for all $v \in X$ such that

$$\langle Av, w \rangle_X = a(v, w) \quad \text{for all } w \in X.$$

Moreover, there exists $f = \mathcal{R}^{-1} F \in X$ such that

$$F(w) = \langle f, w \rangle_X \quad \text{for all } w \in X.$$

The mapping $A : v \to Av$ is linear and continuous with

$$\|Av\|_X = \|Ra(v, \cdot)\|_X = \|a(v, \cdot)\|_{X^*} = \sup_{0 \neq w \in X} \frac{a(v, w)}{\|w\|_X} \leq \sup_{0 \neq w \in X} \frac{\beta \|v\|_X \|w\|_X}{\|w\|_X} = \beta \|v\|_X \tag{A.20}$$

where we used that the operator $R$ defined in Theorem A.5 is an isometry. With this notation (A.19) is equivalent to finding $u \in X$ such that

$$Au = f.$$

We want to show that the mapping

$$T_\delta : X \longrightarrow X, \quad v \mapsto v - \delta (Av - f)$$

is a contraction for an appropriate $\delta > 0$, i.e., $T_\delta$ satisfies $\|T_\delta v - T_\delta w\|_X \leq q \|v - w\|_X$ with some $0 < q < 1$ for all $v, w \in X$. Given $v, w \in X$, set $e = v - w$. Then

$$
\begin{aligned}
\|T_\delta v - T_\delta w\|_X^2 &= \|v - \delta(Av - f) - w + \delta(Aw - f)\|_X^2 \\
&= \|v - w - \delta(Av - Aw)\|_X^2 \\
&= \|v - w - \delta A(v - w)\|_X^2 \\
&= \|e - \delta Ae\|_X^2 \\
&= \|e\|_X^2 - 2\delta \underbrace{\langle e, Ae\rangle_X}_{=\langle Ae,e\rangle_X = a(e,e)} + \delta^2 \|Ae\|_X^2 \\
&= \|e\|_X^2 - 2\delta \underbrace{a(e,e)}_{\geq \alpha\|e\|_X^2} + \delta^2 \underbrace{\|Ae\|_X^2}_{\leq \beta^2\|e\|_X^2 \text{ by (A.20)}} \\
&\leq \|e\|_X^2 - 2\delta\alpha\|e\|_X^2 + \delta^2\beta^2\|e\|_X^2 \\
&= (1 - \delta\alpha + \delta^2\beta^2)\|v - w\|_X^2.
\end{aligned}
$$

For $\delta$ such that

$$0 < q^2 = 1 - 2\alpha\delta + \delta^2\beta^2 < 1,$$

e.g. $\delta = \frac{3\alpha}{2\beta^2}$ (recall that $\alpha \leq \beta$), the operator $T_\delta$ is a contraction. By Banach's fixed point theorem, there exists a unique fixed point $u \in X$. For this particular $u$, we have

$$u = T_\delta u = u - \delta(Au - f)$$

and, hence, $Au = f$ (or equivalently $a(u, v) = F(v)$ for all $v \in X$). Choosing $v = u$, we verify that

$$\alpha\|u\|_X^2 \leq a(u,u) = F(u) \leq \|F\|_{X^*}\|u\|_X,$$

which finishes the proof. □

# References


1. A. Abdulle. The finite element heterogeneous multiscale method: a computational strategy for multiscale PDEs. In *Multiple scales problems in biomathematics, mechanics, physics and numerics*, volume 31 of *GAKUTO Internat. Ser. Math. Sci. Appl.*, pages 133–181. Gakkōtosho, Tokyo, 2009. (Not cited)
2. Grégoire Allaire. Homogenization and two-scale convergence. *SIAM J. Math. Anal.*, 23(6):1482–1518, 1992. (Not cited)
3. S. Armstrong, T. Kuusi, and J.-C. Mourrat. The additive structure of elliptic homogenization. *Invent. Math.*, 2017. To appear. (Not cited)
4. S. N. Armstrong and C. K. Smart. Quantitative stochastic homogenization of convex integral functionals. *Ann. Sci. Éc. Norm. Supér. (4)*, 49(2):423–481, 2016. (Not cited)
5. Jean-Pierre Aubin. *Analyse fonctionnelle appliquée. Tome 1 & 2*. Mathématiques. [Mathematics]. Presses Universitaires de France, Paris, 1987. (Not cited)
6. I. Babuska and R. Lipton. Optimal local approximation spaces for generalized finite element methods with application to multiscale problems. *Multiscale Model. Simul.*, 9(1):373–406, 2011. (Not cited)
7. I. Babuška and J. E. Osborn. Generalized finite element methods: their performance and their relation to mixed methods. *SIAM J. Numer. Anal.*, 20(3):510–536, 1983. (Not cited)
8. M. Bebendorf. A note on the Poincaré inequality for convex domains. *J. Anal. Appl.*, 22:751–756, 2003. (Not cited)
9. A. Bensoussan, J.-L. Lions, and G. Papanicolaou. *Asymptotic Analysis for Periodic Structures*. North-Holland Publ., 1978. (Not cited)
10. A. Bourgeat and A. Piatnitski. Approximations of effective coefficients in stochastic homogenization. *Ann. Inst. H. Poincaré Probab. Statist.*, 40(2):153–165, 2004. (Not cited)
11. Susanne C. Brenner and L. Ridgway Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008. (Not cited)
12. Franco Brezzi and Alessandro Russo. Choosing bubbles for advection-diffusion problems. *Math. Models Methods Appl. Sci.*, 4(4):571–587, 1994. (Not cited)
13. P. G. Ciarlet. *The finite element method for elliptic problems*. North-Holland, 1987. (Not cited)
14. D. Cioranescu, A. Damlamian, and G. Griso. The periodic unfolding method in homogenization. *SIAM Journal on Mathematical Analysis*, 40(4):1585–1620, 2008. (Not cited)
15. Doina Cioranescu, Alain Damlamian, and Georges Griso. Periodic unfolding and homogenization. *Comptes Rendus Mathematique*, 335(1):99 – 104, 2002. (Not cited)
16. 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. (Not cited)
17. 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. (Not cited)
18. Ennio De Giorgi. Sulla convergenza di alcune successioni d'integrali del tipo dell'area. *Rend. Mat. (6)*, 8:277–294, 1975. Collection of articles dedicated to Mauro Picone on the occasion of his ninetieth birthday. (Not cited)
19. M. Duerinckx, A. Gloria, and F. Otto. The structure of fluctuations in stochastic homogenization. *arXiv e-prints*, 1602.01717 [math.AP], 2016. (Not cited)
20. W. E and B. Engquist. The heterogeneous multiscale methods. *Commun. Math. Sci.*, 1(1):87–132, 2003. (Not cited)
21. 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. (Not cited)

22. Yalchin Efendiev and Thomas Y. Hou. *Multiscale finite element methods*, volume 4 of *Surveys and Tutorials in the Applied Mathematical Sciences*. Springer, New York, 2009. (Not cited)
23. Alexandre Ern and Jean-Luc Guermond. Finite element quasi-interpolation and best approximation. *arXiv e-prints*, 1505.06931, 2016. Preprint. (Not cited)
24. Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010. (Not cited)
25. D. Gallistl and D. Peterseim. Computation of quasi-local effective diffusion tensors and connections to the mathematical theory of homogenization. *SIAM Multiscale Model. Simul.*, 15(4), 2017. (Not cited)
26. D. Gilbarg and N.S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer-Verlag, 1983. (Not cited)
27. A. Gloria, S. Neukamm, and F. Otto. A regularity theory for random elliptic operators. *arXiv e-prints*, 1409.2678, 2014. (Not cited)
28. A. Gloria, S. Neukamm, and F. Otto. Quantification of ergodicity in stochastic homogenization: optimal bounds via spectral gap on Glauber dynamics. *Invent. Math.*, 199(2):455–515, 2015. (Not cited)
29. A. Gloria and F. Otto. An optimal variance estimate in stochastic homogenization of discrete elliptic equations. *Ann. Probab.*, 39(3):779–856, 2011. (Not cited)
30. A. Gloria and F. Otto. An optimal error estimate in stochastic homogenization of discrete elliptic equations. *Ann. Appl. Probab.*, 22(1):1–28, 2012. (Not cited)
31. A. Gloria and F. Otto. The corrector in stochastic homogenization: optimal rates, stochastic integrability, and fluctuations. *arXiv e-prints*, 1510.08290 [math.AP], 2015. (Not cited)
32. A. Gloria and F. Otto. Quantitative results on the corrector equation in stochastic homogenization. *J. Eur. Math. Soc. (JEMS)*, 19(11):3489–3548, 2017. (Not cited)
33. L. Grasedyck, I. Greff, and S. Sauter. The al basis for the solution of elliptic problems in heterogeneous media. *Multiscale Model. Simul.*, 10(1):245–258, 2012. (Not cited)
34. P. Henning, A. Målqvist, and D. Peterseim. A localized orthogonal decomposition method for semi-linear elliptic problems. *ESAIM: Mathematical Modelling and Numerical Analysis*, eFirst, 2013. (Not cited)
35. P. Henning and D. Peterseim. Oversampling for the multiscale finite element method. *Multiscale Modeling & Simulation*, 11(4):1149–1175, 2013. (Not cited)
36. Patrick Henning, Philipp Morgenstern, and Daniel Peterseim. Multiscale partition of unity. In Michael Griebel and Marc Alexander Schweitzer, editors, *Meshfree Methods for Partial Differential Equations VII*, volume 100 of *Lecture Notes in Computational Science and Engineering*, pages 185–204. Springer International Publishing, 2015. (Not cited)
37. T. Y. Hou and P. Liu. Optimal Local Multi-scale Basis Functions for Linear Elliptic Equations with Rough Coefficient. *ArXiv e-prints*, August 2015. (Not cited)
38. Thomas Y. Hou and Xiao-Hui Wu. A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comput. Phys.*, 134(1):169–189, 1997. (Not cited)
39. T. Hughes and G. Sangalli. Variational multiscale analysis: the fine-scale Green's function, projection, optimization, localization, and stabilized methods. *SIAM J. Numer. Anal.*, 45(2):539–557, 2007. (Not cited)
40. T. J. R. Hughes. Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods. *Comput. Methods Appl. Mech. Engrg.*, 127(1-4):387–401, 1995. (Not cited)
41. R. Kornhuber, D. Peterseim, and H. Yserentant. An analysis of a class of variational multiscale methods based on subspace decomposition. *Mathematics of Computation*, 2018 (online). (Not cited)
42. Ralf Kornhuber and Harry Yserentant. Numerical homogenization of elliptic multiscale problems by subspace decomposition. *Multiscale Modeling & Simulation*, 14(3):1017–1036, 2016. (Not cited)
43. S. M. Kozlov. The averaging of random operators. *Mat. Sb. (N.S.)*, 109(151)(2):188–202, 327, 1979. (Not cited)

44. M. G. Larson and A. Målqvist. Adaptive variational multiscale methods based on a posteriori error estimation: energy norm estimates for elliptic problems. *Comput. Methods Appl. Mech. Engrg.*, 196(21-24):2313–2324, 2007. (Not cited)
45. Axel Målqvist. *Adaptive variational multiscale methods*. 2005. (Not cited)
46. Axel Målqvist and Daniel Peterseim. Computation of eigenvalues by numerical upscaling. *Numer. Math.*, 130(2):337–361, 2014. (Not cited)
47. Axel Målqvist and Daniel Peterseim. Localization of elliptic multiscale problems. *Math. Comp.*, 83(290):2583–2603, 2014. (Not cited)
48. William McLean. *Strongly elliptic systems and boundary integral equations*. Cambridge University Press, Cambridge, 2000. (Not cited)
49. F. Murat and L. Tartar. H-convergence. *Séminaire d'Analyse Fonctionnelle et Numérique de l'Université d'Alger*, 1978. (Not cited)
50. 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. (Not cited)
51. Gabriel Nguetseng. A general convergence result for a functional related to the theory of homogenization. *SIAM J. Math. Anal.*, 20(3):608–623, 1989. (Not cited)
52. J. Nolen, G. Papanicolaou, and O. Pironneau. A framework for adaptive multiscale methods for elliptic problems. *Multiscale Model. Simul.*, 7(1):171–196, 2008. (Not cited)
53. H. Owhadi. Multigrid with rough coefficients and Multiresolution operator decomposition from Hierarchical Information Games. *ArXiv e-prints*, March 2015. (Not cited)
54. H. Owhadi, L. Zhang, and L. Berlyand. Polyharmonic homogenization, rough polyharmonic splines and sparse super-localization. *ESAIM: Math. Model. Numer. Anal.*, eFirst, 2013. (Not cited)
55. Houman Owhadi. Bayesian numerical homogenization. *Multiscale Modeling & Simulation*, 13(3):812–828, 2015. (Not cited)
56. G. C. Papanicolaou and S. R. S. Varadhan. Boundary value problems with rapidly oscillating random coefficients. In *Random fields, Vol. I, II (Esztergom, 1979)*, volume 27 of *Colloq. Math. Soc. János Bolyai*, pages 835–873. North-Holland, Amsterdam-New York, 1981. (Not cited)
57. L. E. Payne and H. F. Weinberger. An optimal Poincaré inequality for convex domains. *Arch. Rat. Mech. Anal.*, 5:286–292, 1960. (Not cited)
58. Daniel Peterseim. Variational multiscale stabilization and the exponential decay of fine-scale correctors. In G.R. Barrenechea, F. Brezzi, A. Cangiani, and E.H. Georgoulis, editors, *Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations*, volume 114 of *Lecture Notes in Computational Science and Engineering*, pages 343–369. Springer International Publishing, 2016. (Not cited)
59. L. A. Richards. Capillary conduction of liquids through porous mediums. *Journal of Applied Physics*, 1(5):318–333, 1931. (Not cited)
60. Walter Rudin. *Real and complex analysis, 3rd ed.* McGraw-Hill, Inc., New York, NY, USA, 1987. (Not cited)
61. S. Sauter. *hp*-finite elements for elliptic eigenvalue problems: error estimates which are explicit with respect to $\lambda$, $h$, and $p$. *SIAM J. Numer. Anal.*, 48(1):95–108, 2010. (Not cited)
62. 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. (Not cited)
63. E. M. Stein. *Singular integrals and differentiability properties of functions*. Princeton Mathematical Series, No. 30. Princeton University Press, Princeton, N.J., 1970. (Not cited)
64. H.W. Wilhelm and S. Luckhaus. Quasilinear elliptic-parabolic differential equations. *Mathematische Zeitschrift*, 183(3):311–341, 1983. (Not cited)
65. V. V. Yurinskiĭ. Averaging of symmetric diffusion in a random medium. *Sibirsk. Mat. Zh.*, 27(4):167–180, 215, 1986. (Not cited)