Research Article

Boris Khoromskij and Sergey Repin*

Rank Structured Approximation Method for Quasi-Periodic Elliptic Problems

DOI: 10.1515/cmam-2017-0014
Received February 24, 2017; revised April 28, 2017; accepted May 16, 2017

Abstract: We consider an iteration method for solving an elliptic type boundary value problem $\mathcal{A}u = f$, where a positive definite operator $\mathcal{A}$ is generated by a quasi-periodic structure with rapidly changing coefficients (a typical period is characterized by a small parameter ϵ). The method is based on using a simpler operator $\mathcal{A}_0$ (inversion of $\mathcal{A}_0$ is much simpler than inversion of $\mathcal{A}$), which can be viewed as a preconditioner for $\mathcal{A}$. We prove contraction of the iteration method and establish explicit estimates of the contraction factor q. Certainly the value of q depends on the difference between $\mathcal{A}$ and $\mathcal{A}_0$. For typical quasi-periodic structures, we establish simple relations that suggest an optimal $\mathcal{A}_0$ (in a selected set of "simple" structures) and compute the corresponding contraction factor. Further, this allows us to deduce fully computable two-sided a posteriori estimates able to control numerical solutions on any iteration. The method is especially efficient if the coefficients of $\mathcal{A}$ admit low-rank representations and if algebraic operations are performed in tensor structured formats. Under moderate assumptions the storage and solution complexity of our approach depends only weakly (merely linear-logarithmically) on the frequency parameter $\frac{1}{\epsilon}$.

Keywords: Elliptic Problems with Periodic and Quasi-Periodic Coefficients, Precondition Methods, Tensor Type Methods, Guaranteed Error Bounds

MSC 2010: 65F30, 65F50, 65N35, 65F10

1 Introduction

Problems with periodic and quasi-periodic structures arise in various natural sciences models and technical applications. Quantitative analysis of such problems requires special methods oriented towards their specific features. For perfectly periodic structures, efficient methods are developed within the framework of the homogenization theory (see, e.g., [1, 3, 8] and other literature cited therein). However, classical homogenization methods cover only one class of problems (all cells are self-similar and the amount of cells is very large). In this paper, we use a different idea and suggest another *modus operandi* for quantitative analysis of boundary value problems with periodic and quasi-periodic coefficients. It generates approximations converging (in the energy space) to the exact solution and provides guaranteed and computable error estimates. The approach is applicable to (see, e.g., Figures 1, 2)

(i) periodic structures, in which the amount of cell is considerable (e.g., 10^3–10^4) but not large enough to neglect the error generated by the respective homogenized model;
(ii) quasi-periodic structures that contain cells with defects and deformations;

Boris Khoromskij: Max Planck Institute for Mathematics in the Sciences, Inselstr. 22–26, 04103 Leipzig, Germany, e-mail: bokh@mis.mpg.de
***Corresponding author: Sergey Repin:** Russian Academy of Sciences, Saint Petersburg Department of V. A. Steklov Institute of Mathematics, Fontanka 27, 191 011 Saint Petersburg, Russia; and University of Jyväskylä, P.O. Box 35, FI-40014, Jyväskylä, Finland, e-mail: repin@pdmi.ras.ru

(iii) multi-periodic structures where the coefficients reflect the combined effect of several functions with different periodicity.

In general terms, the idea of the method is as follows. We consider the problem $\mathcal{P}$:

$$\mathcal{A}u = f, \quad f \in V^*, \tag{1.1}$$

where V is a reflexive Banach space with the norm $\|\cdot\|_V$, V^* is the space conjugate to V (the respective duality pairing is denoted by $\langle v^*, v\rangle$), and $\mathcal{A} : V \to V^*$ is a bounded linear operator. It is assumed that the operator $\mathcal{A}$ is positive definite and invertible, so that problem (1.1) is well posed. However, $\mathcal{P}$ is viewed as a very difficult problem because $\mathcal{A}$ is generated by a complicated physical structure, which may contain a huge amount of details. Therefore, attempts to solve (1.1) numerically by standard methods may lead to enormous expenditures. Similar difficulties arise if we wish to verify the quality of a numerical solution.

Assume that the operator $\mathcal{A}$ is approximated by a simplified positive definite operator $\mathcal{A}_\circ$ and the inversion of $\mathcal{A}_\circ$ is much simpler than the inversion of $\mathcal{A}$. By means of $\mathcal{A}_\circ$, we construct an iteration method based on solving a "simple" problem $\mathcal{P}_0$: $\mathcal{A}_\circ u_\circ = g$. In other words, the method is based on the operation $g \to \mathcal{A}_\circ^{-1} g$. It also includes the operation $v \to \mathcal{A}v$, which can be performed very efficiently by *tensor-type decomposition methods* provided that physical structures generated $\mathcal{A}$ have low-rank representations. We prove that iterations generate a sequence of functions converging to the exact solution of (1.1) with a geometrical rate. Furthermore, we deduce explicitly computable and guaranteed a posteriori error estimates adapted to this class of problems. They evaluate the accuracy of approximations computed on each step of the iteration algorithm. These estimates also use only inversion of $\mathcal{A}_\circ$ and operations of the type $v \to \mathcal{A}v$. In the iteration methods and error estimates *inversion of the operator $\mathcal{A}$ is avoided*.

In this paper, we consider one class of problems associated with divergent type elliptic equations where $\mathcal{A} = Q^*\Lambda Q$ and $\mathcal{A}_\circ = Q^*\Lambda_\circ Q$. Here $\Lambda : Y \to Y$ is a bounded operator induced by a complicated quasi-periodic structure while $Q : V \to Y$ and $Q^* : Y \to V^*$ are conjugate operators, i.e.,

$$(y, Qw) = \langle Q^* y, w\rangle \quad \text{for all } y \in Y \text{ and } w \in V,$$

where Y is a Hilbert space with the scalar product $(\cdot,\cdot)$ and the norm $\|\cdot\|$. The operators Q and Q^* are induced by differential operators or certain finite-dimensional approximations of them. Henceforth, it is assumed that $f \in \mathcal{V}$, where $\mathcal{V}$ is a Hilbert space with the scalar product $(\cdot,\cdot)_{\mathcal{V}}$. This space is intermediate between V and V^*, i.e., $V \in \mathcal{V} \in V^*$.

The operator $\mathcal{A}_\circ = Q^*\Lambda_\circ Q$ contains the operator $\Lambda_\circ$ generated by a simplified structure. We assume that the operators Λ and $\Lambda_\circ$ are Hermitian (i.e., $(\Lambda y, z) = (y, \Lambda z)$ and $(\Lambda_\circ y, z) = (y, \Lambda_\circ z)$) and satisfy the conditions

$$\lambda_\ominus^\circ \|y\|^2 \le (\Lambda_\circ y, y) \le \lambda_\oplus^\circ \|y\|^2 \quad \text{for all } y \in Y,$$
$$\lambda_\ominus \|y\|^2 \le (\Lambda y, y) \le \lambda_\oplus \|y\|^2, \quad \lambda_\ominus < \lambda_\oplus.$$

Then, the structural operators Λ and $\Lambda_\circ$ are spectrally equivalent:

$$c_1(\Lambda_\circ y, y) \le (\Lambda y, y) \le c_2(\Lambda_\circ y, y), \tag{1.2}$$

where the constants are the minimal and maximal eigenvalues of the generalized spectral problem $\Lambda y - \mu\Lambda_\circ y = 0$. Obviously, they satisfy the estimates $c_1 \ge \lambda_\ominus/\lambda_\oplus^\circ$ and $c_2 \le \lambda_\oplus/\lambda_\ominus^\circ$ (which may be rather coarse).

Concerning the operator Q, we assume that there exists a positive constant c such that

$$\|Qw\| \ge c\|w\|_V \quad \text{for all } w \in V.$$

Generalized solutions of the problems $\mathcal{P}$ and $\mathcal{P}_0$ are defined by the variational identities

$$(\Lambda Qu, Qw) = \langle f, w\rangle \quad \text{for all } w \in V, \tag{1.3}$$

and

$$(\Lambda_\circ Qu_\circ, Qw) = \langle \tilde{f}, w\rangle \quad \text{for all } w \in V. \tag{1.4}$$

In Section 2, we show that a sequence $\{u_k\}$ converging to u in V can be constructed by solving problems (1.4) with specially constructed right-hand sides $\widetilde{f}_k$ generated by the residual of (1.3). In proving convergence, the key issue is analysis of the spectral radius of the operator

$$\mathbb{B}_\rho := \mathbb{I} - \rho\Lambda_\circ^{-1}\Lambda, \tag{1.5}$$

and selection of such relaxation parameter ρ that provides the best convergence rate. Moreover, iteration procedures of such a type become contracting if the iteration parameter is properly selected. This fact is often used in proving analytical results (see, e.g., [20], where classical results on existence and uniqueness of a variational inequality are established by contraction arguments). Also, these ideas were used in the construction of various numerical methods (see, e.g., [7]). However, achieving our goals requires more than the fact of contraction. We need explicit and realistic estimates of the contraction factor (which are used in error analysis) and a practical method of finding $\Lambda_\circ$ with minimal q. The latter task leads to a special optimization problem that defines the most efficient "simplified" operator $\Lambda_\circ$ among a certain class of "admissible" operators. This question is studied in Section 3. In general, Λ and $\Lambda_\circ$ can be induced by scalar, vector, and tensors functions. We show that selection of the optimal structural operator $\Lambda_\circ$ is reduced to a special interpolation type problem, which is purely algebraical and does not require solving a differential problem (therefore a suitable $\Lambda_\circ$ can be found a priori). We discuss several examples and suggest the corresponding optimal (or quasi-optimal) $\Lambda_\circ$, which guarantees convergence of the iteration sequence with *explicitly known contraction factor*.

Now, it is worth discussing the main differences between our approach and the classical homogenization method developed for regular periodic structures. This method operates with a homogenized boundary value problem $Q^*\Lambda_{\mathrm{H}}Qu_{\mathrm{H}} = f$, where Λ_{H} is defined by means of an auxiliary problem with periodical boundary conditions in the cell of periodicity. The respective solution u_{H} contains an irremovable (modeling) error depending on the cell diameter ϵ. Moreover, if ϵ tends to zero, then typically u_{H} converges to u only weakly (e.g., in L^2). Getting a better convergence (e.g., in H^1) requires certain corrections, which lead to other (more complicated) boundary value problems in the cell of periodicity. The respective "corrected" solution u_{H}^c also contains an error. Typically, the error is proportional to $\sqrt{\epsilon}$ and can be neglected only if the amount of cells is very large. If our method is applied to perfectly periodical structures then setting $\Lambda_\circ := \Lambda_{\mathrm{H}}$ is one possible option. In this case, the homogenized operator (defined without correction procedures) is used for a different purpose: *construction of a suitable preconditioning operator*. The latter operator generates numerical solutions converging to the exact solution in the energy norm (i.e., the method is free from irremovable errors) and can be applied for a rather wide range of ϵ. In addition, the theory suggests other simpler ways of selecting suitable $\Lambda_\circ$. In this context, it is interesting to know whether or not the choice $\Lambda_\circ := \Lambda_H$ always yields the minimal value of the contraction factor. In Section 3, we briefly discuss this question and present an example of that the best $\Lambda_\circ$ may differ from Λ_{H}.

In Section 4, we deduce a posteriori estimates that provide fully computable and guaranteed estimates of the distance to the exact solution u for any numerical approximation $u_{k,h}$ computed for an approximation subspace V_h. These estimates are established by combining functional type a posteriori estimates (see [22, 25, 26] and references cited therein) and estimates generated by the contraction property of the iteration method (see [24, 29]).

The second part of the paper is devoted to a fast solution method for the basic iteration problem (2.1). The key idea consists of using tensor-type representations for approximations, what is quite natural if both coefficients of the respective quasi-periodic structure and the right-hand side admit low-rank tensor-type representations. We notice that the amount of structures representable in terms of low-rank formats is much larger than the amount of periodic structures covered by the homogenization method. The idea of tensor-type approximations of partial differential equations traces back to [9]. In computational mechanics this method is known as the Kantorovich–Krylov (or extended Kantorovich) method. However, it is rarely used in modern numerical technologies. In part, this is due to restrictions on the shape of the domain imposed by the Kantorovich method. Henceforth, we assume that the domain Ω satisfies these restrictions, i.e., it is a tensor-type domain (e.g., rectangular) or a union of tensor-type domains. This assumption induces certain geometrical

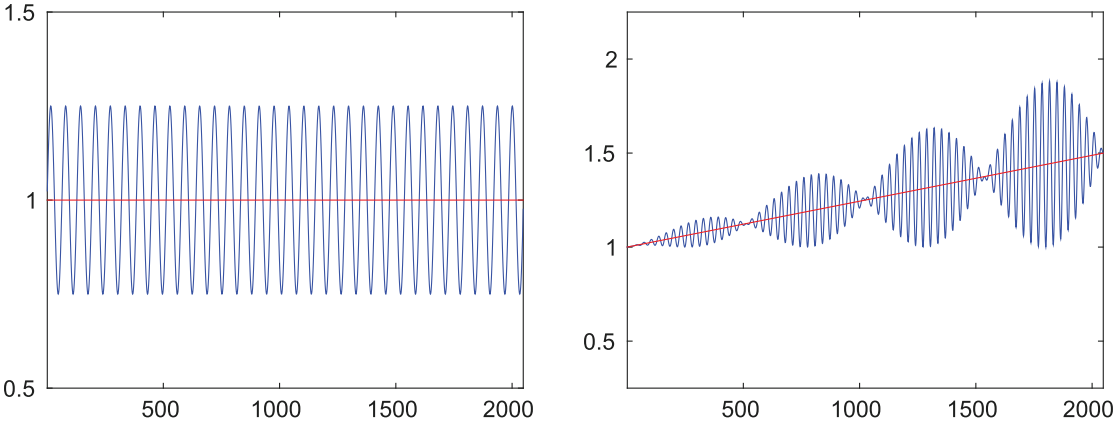

Figure 1. Examples of periodic and modulated periodic coefficients in 1D.

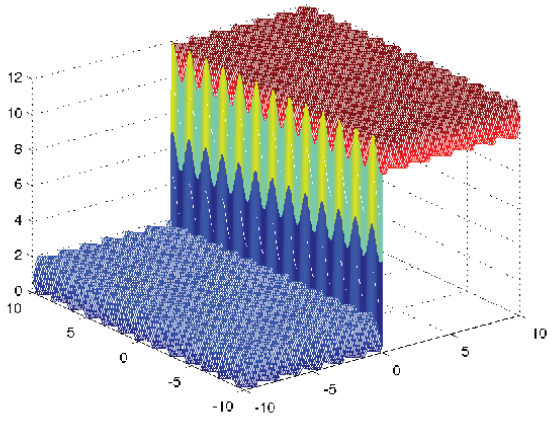

Figure 2. An example of modulated piecewise periodic coefficients in 2D.

limitations. However, they can be bypassed by such methods as coordinate transformation and domain decomposition, which are widely used in modern computational mathematics (e.g., in iso-geometric analysis).

The recent tensor numerical methods (for steady state and dynamical problems) based on the advanced nonlinear tensor approximation algorithms have been developed in the last ten years. Literature survey on the modern tensor numerical methods for multi-dimensional PDEs can be found in [13, 14, 16]. In the context of problems considered in the paper, we are mainly concerned with another specific feature: very complicated material structure. In this case, direct application of standard finite element methods suffers from the necessity to account huge information encompassed in coefficients (especially in multi-dimensional problems). We show that tensor-type methods allow us to reduce computations to a collection of one-dimensional problems, which can be solved very efficiently using low-rank representations with the small storage requests. Similar ideas are applied for computing a posteriori error estimates.

Section 5 discusses numerical aspects of the method and exposes several examples. Typical behavior of quasi-periodic coefficients is described by oscillation around a constant, modulated oscillation around a given smooth function, or oscillation around a piecewise constant function.

Figure 1 (1D case) represents examples of highly oscillating (left) and modulated periodic coefficients (right) functions. Figure 2 (2D case) illustrates the well-separable equation coefficient obtained by a sum of step-type and uniformly oscillating functions, namely,

$$a(x_1, x_2) = C_0 \operatorname{sign}(x_1) + C_1 + \sin\Big(\frac{\pi\omega}{a}(x_1 + x_2)\Big),$$

where $C_0 = 5$, $C_1 = 6$, $a = 10$, and $\omega = 12$.

We show that specially constructed FEM type approximations of PDEs with slightly perturbed or regularly modulated periodic coefficients on d-fold $n \times \cdots \times n$ tensor grids in $\mathbb{R}^d$ may lead to the discretized algebraic equations with the low Kronecker rank stiffness matrix of size $n^d \times n^d$, where $n = O(\frac{1}{\epsilon})$ is proportional to the large frequency parameter $\frac{1}{\epsilon}$. In this case, the rank decomposition with respect to the d spacial variables is applied, such that the discrete solution can be calculated in the low-rank separable form, which requires storage size of only $O(dn)$ instead of $O(n^d) = O(\frac{1}{\epsilon^d})$ complexity representations which are mandatory for the traditional FEM techniques (the latter quickly leads to the bottleneck in case of small parameter $\epsilon > 0$).

The arising linear system of equations can be solved by preconditioned iteration with the simple preconditioner $\Lambda_\circ$, such that the storage and numerical costs scale almost linearly in the univariate discrete problem size n. Numerical examples in Section 5 demonstrate the stable geometric convergence of the preconditioned CG (PCG) iteration with the preconditioner $\Lambda_\circ$ and confirm the low-rank approximate separable representation to the solution with respect to d spacial variables even in the case of complicated quasi-periodic coefficients.

This approach is well suited for applying the quantized-TT (QTT) tensor approximation [15] to functions discretized on large tensor grids of size proportional to the frequency parameter, i.e., $n = O(\frac{1}{\epsilon})$, as it was demonstrated in the previous paper [17] for the case $d = 1$. The use of tensor-structured preconditioned iteration with the adaptive QTT rank truncation may lead to the logarithmic complexity in the grid size, $O(\log^p n)$, see [14, 16, 23] for the rank-truncated iterative methods, [4, 10–12] for various examples of the QTT tensor approximation to lattice structured systems, and [2] for tensor approximation of complicated functions with multiple cusps in $\mathbb{R}^d$.

In Section 6, we conclude with the discussion on further perspectives of the presented approach for 2D and 3D elliptic PDEs with quasi-periodic coefficients.

2 The Iteration Method

Let $v \in V$ and $\rho \in \mathbb{R}_+$. Consider the problem: find u_v such that

$$(\Lambda_\circ Q u_v, Qw) = \ell_v^\circ(w) - \rho \ell_v(w) \quad \text{for all } w \in V, \tag{2.1}$$

where

$$\ell_v(w) := (\Lambda Q v, Qw) - \langle f, w\rangle \quad \text{and} \quad \ell_v^\circ(w) := (\Lambda_\circ Q v, Qw).$$

Obviously, the right-hand side of (2.1) is a bounded linear functional on V, so that this problem has a unique solution u_v. Thus, we have a mapping $T_\rho : V \to V$, which becomes a contraction if the parameter ρ is properly selected. Indeed, for any v_1 and v_2 in V, we obtain

$$(\Lambda_\circ Q\eta, Qw) = (\Lambda_\circ Q\zeta - \rho\Lambda Q\zeta, Qw) \quad \text{for all } w \in V,$$

where $u_1 = T_\rho v_1$, $u_2 = T_\rho v_2$, $\zeta := v_1 - v_2$, and $\eta := u_1 - u_2$. Hence

$$\begin{aligned}\|\eta\|_\circ^2 := (\Lambda_\circ Q\eta, Q\eta) &= (\Lambda_\circ Q\zeta, Q\eta) - \rho(\Lambda Q\zeta, Q\eta)\\ &= (Q\zeta, \Lambda_\circ Q\eta) - \rho(\Lambda_\circ^{-1}\Lambda Q\zeta, \Lambda_\circ Q\eta) = (Q\zeta - \rho\Lambda_\circ^{-1}\Lambda Q\zeta, \Lambda_\circ Q\eta)\\ &\le \|\eta\|_\circ (\Lambda_\circ Q\zeta - \rho\Lambda Q\zeta, Q\zeta - \rho\Lambda_\circ^{-1}\Lambda Q\zeta)^{1/2}.\end{aligned} \tag{2.2}$$

From (2.2) we find that

$$\begin{aligned}\|\eta\|_\circ^2 &\le (\Lambda_\circ Q\zeta, Q\zeta) - 2\rho(\Lambda Q\zeta, Q\zeta) + \rho^2(\Lambda_\circ^{-1}\Lambda Q\zeta, \Lambda Q\zeta)\\ &= (Q\zeta, \Lambda_\circ Q\zeta) - 2\rho(\Lambda_\circ^{-1}\Lambda Q\zeta, \Lambda_\circ Q\zeta) + \rho^2(\Lambda_\circ^{-1}\Lambda\Lambda_\circ^{-1}\Lambda Q\zeta, \Lambda_\circ Q\zeta)\\ &= ((\mathbb{I} - 2\rho\Lambda_\circ^{-1}\Lambda + \rho^2\Lambda_\circ^{-1}\Lambda\Lambda_\circ^{-1}\Lambda)Q\zeta, \Lambda_\circ Q\zeta) = (\Lambda_\circ \mathbb{B}_\rho^2 Q\zeta, Q\zeta)\\ &\le (\Lambda_\circ \mathbb{B}_\rho^2 Q\zeta, \mathbb{B}_\rho^2 Q\zeta)^{1/2}(\Lambda_\circ Q\zeta, Q\zeta)^{1/2},\end{aligned} \tag{2.3}$$

where $\mathbb{B}_\rho$ is defined by (1.5). If ρ is selected such that

$$(\Lambda_\circ \mathbb{B}_\rho^2 Q\zeta, Q\zeta) \le q^2 \|\zeta\|_\circ^2 \quad \text{for some } q < 1, \tag{2.4}$$

then (2.3) shows that T_ρ is a contractive mapping.

It is not difficult to show that ρ satisfying (2.4) can be always found. Indeed, in view of (1.2),

$$\begin{aligned}(\Lambda_\circ \mathbb{B}_\rho^2 Q\zeta, Q\zeta) &= (\Lambda_\circ Q\zeta, Q\zeta) - 2\rho(\Lambda Q\zeta, Q\zeta) + \rho^2(\Lambda\Lambda_\circ^{-1}\Lambda Q\zeta, Q\zeta)\\ &\le (1 - 2\rho c_1)(\Lambda_\circ Q\zeta, Q\zeta) + \rho^2(\Lambda_\circ^{-1}\Lambda Q\zeta, \Lambda Q\zeta).\end{aligned} \tag{2.5}$$

Since Λ and $\Lambda_\circ$ are invertible with trivial kernels, μ and y_μ are an eigenvalue and the respective eigenfunction of $\Lambda y_\mu = \mu\Lambda_\circ y_\mu$ if and only if they are an eigenvalue and the eigenfunction of the problem $\Lambda\Lambda_\circ^{-1}\Lambda y_\mu = \mu\Lambda y_\mu$. This means that

$$c_1(\Lambda y, y) \le (\Lambda\Lambda_\circ^{-1}\Lambda y, y) \le c_2(\Lambda y, y) \le c_2^2(\Lambda_\circ y, y).$$

Hence

$$(\Lambda_\circ^{-1}\Lambda Q\zeta, \Lambda Q\zeta) \le c_2^2\|\zeta\|_\circ^2$$

and (2.5) implies

$$(\Lambda_\circ \mathbb{B}_\rho^2 Q\zeta, Q\zeta) \le (1 - 2\rho c_1 + \rho^2 c_2^2)\|\zeta\|_\circ^2.$$

The minimum of the expression in round brackets is attained if $\rho = \rho_* := c_1/c_2^2$. For $\rho = \rho_*$, we find that

$$q_*^2 := 1 - \frac{c_1^2}{c_2^2} \le \hat{q}^2 := 1 - \frac{\lambda_\ominus^2 \lambda_\ominus^{\circ\,2}}{\lambda_\oplus^2 \lambda_\oplus^{\circ\,2}} \in [0, 1).$$

Hence T_ρ is a contractive mapping with explicitly known contraction factor q_*. Well-known results in the theory of fixed points (see, e.g., [29]) yield the following theorem.

Theorem 2.1. *For any $u_0 \in V$ and $\rho = \rho^*$, the sequence $\{u_k\} \in V$ of functions satisfying the relation*

$$(\Lambda_\circ Q u_k, Qw) = (\Lambda_\circ Q u_{k-1}, Qw) - \rho\big((\Lambda Q u_{k-1}, Qw) - \langle f, w\rangle\big) \quad \textit{for all } w \in V \tag{2.6}$$

converges to u in V and $\|u_k - u\|_\circ \le q_*^k \|u_0 - u\|_\circ$ as $k \to +\infty$.

Remark 2.2. From (2.3) we obtain

$$\|\eta\|_\circ^2 \le \frac{1}{\lambda_{0,\min}} \|\mathbb{B}_\rho^2 Q\zeta\| \|\zeta\|_\circ \le \frac{\|\mathbb{B}_\rho^2\|}{\lambda_{0,\min}^2} \|\zeta\|_\circ^2.$$

This relation yields a simple (but not very sharp) estimate of the contraction factor.

For further analysis, it is convenient to estimate the right-hand side of (2.3) by a different method. Let $\|\mathbb{B}_\rho\|_\circ$ denote the operator norm

$$\|\mathbb{B}_\rho\|_\circ := \sup_{y \in Y} \frac{\|\mathbb{B}_\rho y\|_\circ}{\|y\|_\circ}. \tag{2.7}$$

Then $\|\mathbb{B}_\rho y\|_\circ \le \|\mathbb{B}_\rho\|_\circ \|y\|_\circ$ and

$$(\Lambda_\circ \mathbb{B}_\rho^2 y, y) \le \|\mathbb{B}_\rho\|_\circ^2 \|y\|_\circ^2.$$

Hence (2.3) yields the estimate

$$\|\eta\|_\circ \le \|\mathbb{B}_\rho\|_\circ \|\zeta\|_\circ,$$

which shows that T_ρ is a contraction provided that

$$\|\mathbb{B}_\rho\|_\circ < 1. \tag{2.8}$$

In applications $\mathbb{B}_\rho$ is a self-adjoint bounded operator acting in a finite-dimensional space, so that verification of this condition amounts to finding ρ which yields the respective spectral radius of $\mathbb{B}_\rho$ (see Section 4).

3 Selection of $\Lambda_\circ$

In this section, we discuss how to select $\Lambda_\circ$ in order to minimize q, which is crucial for two major aspects of quantitative analysis: convergence of the iteration method and guaranteed a posteriori estimates. We assume that V, $\mathcal{V}$, and Y are spaces of functions defined in a Lipschitz bounded domain Ω (namely $y(x) \in \mathbb{T}$ for a.e. $x \in \Omega$ where $\mathbb{T}$ may coincide with $\mathbb{R}$, $\mathbb{R}^d$, or $\mathbb{M}^{d\times d}$) and the operators Λ and $\Lambda_\circ$ are generated by bounded scalar functions, matrices or tensors. In this case,

$$(\Lambda y, y) := \int_\Omega \Lambda(x)y \odot y\, dx \quad \text{and} \quad (\Lambda_\circ y, y) := \int_\Omega \Lambda_\circ(x)y \odot y\, dx,$$

where $\odot$ denotes the respective product of scalar, vector, or tensor functions. In view of (2.7) and (2.8), the value of ρ should minimize the quantity $\sup_{y\in Y}(\Lambda_\circ \mathbb{B}_\rho y, \mathbb{B}_\rho y)/(\Lambda_\circ y, y)$. This procedure yields the contraction factor

$$q^2 = \mathcal{Q}(\Lambda, \Lambda_\circ) := \inf_\rho \sup_{y\in Y} \frac{\int_\Omega \Lambda_\circ(x)\mathbb{B}_\rho(x)y \odot \mathbb{B}_\rho(x)y\, dx}{\int_\Omega \Lambda_\circ(x)y \odot y\, dx},$$

whose computation is reduced to solving algebraic problems at a.e. $x \in \Omega$, i.e.,

$$\mathcal{Q}(\Lambda, \Lambda_\circ) := \inf_\rho \sup_{x\in\Omega} \sup_{\tau\in\mathbb{T}} \frac{\Lambda_\circ(x)\mathbb{B}_\rho(x)\tau \odot \mathbb{B}_\rho(x)\tau}{\Lambda_\circ(x)\tau \odot \tau}.$$

Let $\mathbb{S}$ be a certain set of "simple" operators defined a priori (e.g., it can be a finite-dimensional set formed by piecewise constant or polynomial functions). Then, finding the best "simplified" operator amounts to solving the following problem: find $\widehat{\Lambda}_\circ \in \mathbb{S}$ such that $\mathcal{Q}(\Lambda, \widehat{\Lambda}_\circ)$ is minimal. In other words, the optimal $\widehat{\Lambda}_\circ$ is defined by the problem

$$\inf_{\substack{\Lambda_0\in\mathbb{S}\\ \rho\in\mathbb{R}}} \sup_{\substack{x\in\Omega\\ \tau\in\mathbb{T}}} \frac{\Lambda_\circ(x)\mathbb{B}_\rho(x)\tau \odot \mathbb{B}_\rho(x)\tau}{\Lambda_\circ(x)\tau \odot \tau} = q^2. \tag{3.1}$$

Notice that (3.1) is an algebraic problem, which should be solved (analytically or numerically) before computations. The respective solution $\widehat{\Lambda}_\circ$ defines the best operator to be used in the iteration method (2.6) and yields the respective contraction factor. Below we discuss some particular cases, where analysis of this problem generates an optimal (or almost optimal) $\Lambda_\circ$.

Problem (3.1) is explicitly solvable if $\Lambda_\circ$ and Λ have a special structure, namely,

$$\Lambda_\circ = a_\circ(x)\mathbb{I}, \quad \Lambda = a(x)\mathbb{I},$$

where $\mathbb{I}$ is the unit operator and $a_\circ(x)$ and $a(x)$ are positive bounded functions defined in Ω. Then,

$$\mathbb{B}_\rho(x) = (1 - \mathbf{h}(x))\mathbb{I}, \quad \mathbf{h}(x) := \frac{a(x)}{a_\circ(x)}$$

and

$$\sup_{\tau\in\mathbb{T}} \frac{(1-\rho\mathbf{h}(x))^2\tau \odot \tau}{|\tau|^2} = |1 - \rho\mathbf{h}(x)|^2 \quad \text{for all } x \in \Omega.$$

Define $\mathbf{h}_\ominus := \min_{x\in\Omega} \mathbf{h}(x)$ and $\mathbf{h}_\oplus := \max_{x\in\Omega} \mathbf{h}(x)$. It is not difficult to show that

$$\sup_{x\in\Omega} |1 - \rho\mathbf{h}(x)| = \max\{|1 - \rho\mathbf{h}_\ominus|, |1 - \rho\mathbf{h}_\oplus|\}.$$

Minimization with respect to ρ yields the best value $\rho_* = \frac{2}{\mathbf{h}_\ominus + \mathbf{h}_\oplus}$ and the respective value

$$\mathcal{Q}(\Lambda, \Lambda_\circ) = \left(\frac{\mathbf{h}_\oplus - \mathbf{h}_\ominus}{\mathbf{h}_\oplus + \mathbf{h}_\ominus}\right)^2 = \left(\frac{1 - \mathcal{J}(a, a_\circ)}{1 + \mathcal{J}(a, a_\circ)}\right)^2 < 1, \quad \mathcal{J}(a, a_\circ) = \frac{\mathbf{h}_\ominus}{\mathbf{h}_\oplus}. \tag{3.2}$$

In accordance with (3.1), the identification of the optimal simplified problem is reduced to the problem

$$\sup_{a_0\in\mathbb{S}} \mathcal{J}(a, a_\circ), \tag{3.3}$$

where $\mathbb{S}$ is a given set of functions.

We illustrate the above relations by means of several examples.

Example 3.1 (Constant Coefficients). In the simplest case, we set $\mathbb{S} = P^0$, i.e., a_0 is a constant. From (3.3) it follows that $q = (\overline{a} - \underline{a})/(\overline{a} + \underline{a})$, where $\underline{a} := \min_{x\in\Omega} a(x)$ and $\overline{a} := \max_{x\in\Omega} a(x)$. Then $\rho_* = 2a_0/(\overline{a} + \underline{a})$ and the iteration procedure (2.6) with $\rho = \rho_*$ has the form

$$\int_\Omega Qu_k \odot Qw\,dx = \int_\Omega \Big(1 - \frac{2a}{\overline{a} + \underline{a}}\Big) Qu_{k-1} \odot Qw\,dx + \frac{2}{\overline{a} + \underline{a}} \int_\Omega fw\,dx.$$

From Theorem 2.1, it follows that

$$\int_\Omega |Q(u_k - u)|^2\,dx \le C\Big(\frac{\overline{a} - \underline{a}}{\overline{a} + \underline{a}}\Big)^{2k}.$$

Example 3.2 (Oscillation Around a Given Function). Consider a somewhat different example. Let $a(x)$ be a function oscillating around a certain mean function $g(x)$ so that

$$\frac{a(x)}{g(x)} \in [1 - \epsilon, 1 + \epsilon], \quad \epsilon \in (0, 1).$$

If g is a relatively simple function, then it is natural to set $a_\circ(x) = g(x)$. By (3.2), we find that $\mathbf{h}_\oplus = 1 + \epsilon$, $\mathbf{h}_\ominus = 1 - \epsilon$, and $q = \epsilon$. Hence the method is very efficient for small ϵ (i.e., if a oscillates around g with a relatively small amplitude). Figures 1 and 2 illustrate three examples of quasi-periodic coefficients a and respective $a_\circ$ corresponding to the case of oscillation around a constant with smooth modulation, oscillation around a given smooth function, or oscillation around a piecewise constant function.

Example 3.3 (Piecewise Constant Coefficients). Consider a more complicated case, where Ω is divided into N nonoverlapping subdomains Ω_i and $\Lambda_\circ(x) = c_i\mathbb{I}$ if $x \in \Omega_i$. Define the numbers

$$a_\ominus^{(i)} = \min_{x\in\Omega_i} a(x), \qquad a_\oplus^{(i)} = \max_{x\in\Omega_i} a(x),$$
$$\mathbf{h}_\ominus = \min\Big\{\frac{a_\ominus^{(1)}}{c_1}, \frac{a_\ominus^{(2)}}{c_2}, \dots, \frac{a_\ominus^{(N)}}{c_N}\Big\}, \quad \mathbf{h}_\oplus = \max\Big\{\frac{a_\oplus^{(1)}}{c_1}, \frac{a_\oplus^{(2)}}{c_2}, \dots, \frac{a_\oplus^{(N)}}{c_N}\Big\}.$$

Since the constants c_i are defined up to a common multiplier, we can without a loss of generality assume that

$$\sum_{i=1}^{(N)} \lambda_i = 1, \quad \text{where } \lambda_i = \frac{1}{c_i}. \tag{3.4}$$

In accordance with (3.3), the maximum of $\mathfrak{Q}(\Lambda, \Lambda_\circ)$ is attained if

$$\frac{\min\{\lambda_1 a_\ominus^{(1)}, \lambda_2 a_\ominus^{(2)}, \dots, \lambda_N a_\ominus^{(N)}\}}{\max\{\lambda_1 a_\oplus^{(1)}, \lambda_2 a_\oplus^{(2)}, \dots, \lambda_N a_\oplus^{(N)}\}} \to \max, \tag{3.5}$$

where $\lambda_i > 0$ and satisfy (3.4). If $N = 2$, then problem (3.5) has a simple solution, which shows that the ratio λ_1/λ_2 (i.e., c_2/c_1) can be any in the interval $[\xi_1, \xi_2]$, where

$$\xi_1 = \min\Big\{\frac{a_\ominus^{(2)}}{a_\ominus^{(1)}}, \frac{a_\oplus^{(2)}}{a_\oplus^{(1)}}\Big\}, \quad \xi_2 = \max\Big\{\frac{a_\ominus^{(2)}}{a_\ominus^{(1)}}, \frac{a_\oplus^{(2)}}{a_\oplus^{(1)}}\Big\}.$$

It is interesting to compare these results with those generated by homogenized models in the case of perfectly periodic structures. For this purpose, we consider a simple one-dimensional problem

$$(au')' - f = 0 \quad \text{in } (0, 1)$$

with

$$a(x) = a^{(1)}(x) \quad \text{in } \Omega_1 = (0, \beta), \quad \beta \in (0, 1),$$
$$a(x) = a^{(2)}(x) \quad \text{in } \Omega_2 = (\beta, 1),$$

where $a^{(1)}(x)$ is a perfectly periodical function attaining only two values $a^{(1)}_{\ominus}$ and $a^{(1)}_{\oplus}$. The Lebesgue measure of the set where $a(x) = a^{(1)}_{\oplus}$ is $\kappa_1|\Omega_1|$, $\kappa_1 \in (0, 1)$. Analogously, $a^{(2)}(x)$ is a periodical function attaining only two values $a^{(2)}_{\ominus}$ and $a^{(2)}_{\oplus}$. The Lebesgue measure of the set where $a(x) = a^{(2)}_{\oplus}$ is $\kappa_2|\Omega_2|$, $\kappa_2 \in (0, 1)$. We assume that $a^{(i)}_{\ominus} < a^{(2)}_{\oplus}$, $i = 1, 2$ and the amount of periods is very large. Then, the homogenization method can be successfully applied. The corresponding homogenized problem has the coefficients (see, e.g., [8])

$$a^{(1)}_H := \Big(\frac{1}{\beta}\int_0^{\beta}\frac{1}{a^{(1)}(x)}\,dx\Big)^{-1} \quad \text{in } \Omega_1, \qquad a^{(2)}_H := \Big(\frac{1}{1-\beta}\int_{\beta}^{1}\frac{1}{a^{(2)}(x)}\,dx\Big)^{-1} \quad \text{in } \Omega_2.$$

It is easy to see that

$$a^{(1)}_H = \frac{a^{(1)}_{\ominus}a^{(1)}_{\oplus}}{\kappa_1 a^{(1)}_{\ominus} + (1-\kappa_1)a^{(1)}_{\oplus}} \in (a^{(1)}_{\ominus}, a^{(1)}_{\oplus}), \quad a^{(2)}_H = \frac{a^{(2)}_{\ominus}a^{(2)}_{\oplus}}{\kappa_2 a^{(2)}_{\ominus} + (1-\kappa_2)a^{(2)}_{\oplus}} \in (a^{(2)}_{\ominus}, a^{(2)}_{\oplus}).$$

Hence

$$\frac{a^{(2)}_H}{a^{(1)}_H} \in (\xi^H_1, \xi^H_2), \quad \text{where } \xi^H_1 := \frac{a^{(2)}_{\ominus}}{a^{(1)}_{\oplus}} < \min\Big\{\frac{a^{(2)}_{\ominus}}{a^{(1)}_{\ominus}}, \frac{a^{(2)}_{\oplus}}{a^{(1)}_{\oplus}}\Big\} = \xi_1, \ \xi^H_2 := \frac{a^{(2)}_{\oplus}}{a^{(1)}_{\ominus}} > \xi_2$$

and $(\xi_1, \xi_2) \subset (\xi^H_1\xi^H_2)$. Therefore, the coefficients $a^{(1)}_H$ and $a^{(2)}_H$ may not generate the best piecewise constant $a_\circ$, which produces the smallest contraction factor in the iteration procedure (2.6).

4 Error Estimates

4.1 General Estimate

Since T_ρ is a contractive mapping, we can use the Ostrowski estimates (see [24, 26, 29]) of the distance between $v \in V$ and u (the fixed point). The estimates state that

$$\|v-u\|_\circ \in \Big\{\frac{\epsilon}{1+q(\rho)}, \frac{\epsilon}{1-q(\rho)}\Big\}, \quad \text{where } \epsilon := \|T_\rho v - v\|_\circ. \tag{4.1}$$

This estimate cannot be directly applied because $v_\rho := T_\rho v$ is generally unknown (it is the exact solution of a boundary value problem). Instead, we must use a numerical approximation $\widetilde{v}_\rho$ (in our analysis, we impose no restrictions on the method by which the function $\widetilde{v}_\rho \in V$ was constructed). Thus, the difference $\eta_\rho := v - \widetilde{v}_\rho$ is a known function and the quantity $\delta_\rho = \|\eta_\rho\|_\circ$ is directly computable. It is easy to see that

$$\delta_\rho - \|\widetilde{v}_\rho - v_\rho\|_\circ \le \|v_\rho - v\|_\circ \le \delta_\rho + \|\widetilde{v}_\rho - v_\rho\|_\circ. \tag{4.2}$$

To deduce a fully computable majorant of the norm $\|\widetilde{v}_\rho - v_\rho\|_\circ$ we use the method suggested in [25, 26]. First, we rewrite (2.1) in the form

$$(\Lambda_\circ Q v_\rho, Qw) = (\Lambda_\circ Qv, Qw) - \rho\big((\Lambda Qv, Qw) - \langle f, w\rangle\big). \tag{4.3}$$

For any $y \in Y$ and $w \in V_0$, we have

$$\begin{aligned}(\Lambda_\circ Q(v_\rho - \widetilde{v}_\rho), Qw) &= (\Lambda_\circ Q(v-\widetilde{v}_\rho), Qw) - \rho\big((\Lambda Qv, Qw) - \langle f, w\rangle\big) \\ &= (\Lambda_\circ Q(v-\widetilde{v}_\rho) - \rho\Lambda Qv + y, Qw) - \langle Q^*y - \rho f, w\rangle.\end{aligned} \tag{4.4}$$

We estimate the first term on the right-hand side of (4.4) as follows:

$$\begin{aligned}\big(\Lambda_\circ Q(v-\widetilde{v}_\rho) - \rho\Lambda Qv + y, Q(v_\rho - \widetilde{v}_\rho)\big) &= \big(Q(v-\widetilde{v}_\rho) - \rho\Lambda_\circ^{-1}\Lambda Qv + \Lambda_\circ^{-1}y, \Lambda_\circ Q(v_\rho - \widetilde{v}_\rho)\big) \\ &\le \big(\Lambda_\circ Q(v-\widetilde{v}_\rho) + \tau_y, Q(v-\widetilde{v}_\rho) + \Lambda_\circ^{-1}\tau_y\big)^{1/2}\|v_\rho - \widetilde{v}_\rho\|_\circ,\end{aligned}$$

where $\tau_y := y - \rho\Lambda Qv$. The second term meets the estimate

$$\langle Q^*y - \rho f, v_\rho - \widetilde{v}_\rho\rangle \le \|Q^*y - \rho f\|\|v_\rho - \widetilde{v}_\rho\| \le \frac{1}{(\lambda^\circ_\ominus)^{1/2}}\|Q^*y - \rho f\|\|v_\rho - \widetilde{v}_\rho\|_\circ,$$

where $\|w^*\| = \sup_{w\in V}\langle w^*, w\rangle/\|w\|_\circ$ is the dual norm. Hence

$$\|v_\rho - \widetilde{v}_\rho\|_\circ \le \left(\Lambda_\circ Q\eta_\rho + \tau_y, Q\eta_\rho + \Lambda_\circ^{-1}\tau_y\right)^{1/2} + \frac{1}{\sqrt{\lambda^\circ_\ominus}}\|Q^*y - \rho f\| =: M_\oplus(\eta_\rho, \tau_y). \tag{4.5}$$

Notice that

$$\inf_{y\in Y} M_\oplus(\eta_\rho, \tau_y) = \|v_\rho - \widetilde{v}_\rho\|_\circ.$$

Indeed, set $y = \Lambda_\circ Q(v_\rho - v) + \rho\Lambda Qv$ (in this case, $\tau_y = \Lambda_\circ Q(v_\rho - v)$). In view of (4.3), $\langle Q^*y - \rho f, w\rangle = 0$, and the majorant is equal to $\|v_\rho - \widetilde{v}_\rho\|^2_\circ$. Thus, estimate (4.5) has no irremovable gap and a properly selected y yields a sharp upper bound of the error.

Remark 4.1. It is not difficult to show that the last term of $M_\oplus(\eta_\rho, \tau_y)$ can be estimated via an explicitly computable quantity provided that y has the same regularity as the true flux (see [26]). However, in our subsequent analysis $Q^*y - \rho f = 0$ and these advanced forms of the majorant are not required. In this case, the majorant has a simpler form:

$$M^2_\oplus(\eta_\rho, \tau_y) = (\Lambda_\circ Q\eta_\rho, Q\eta_\rho) + (\Lambda_\circ^{-1}\tau_y, \tau_y) + 2(Q\eta_\rho, \tau_y).$$

It is important that the computation of the majorant $M_\oplus$ does not require inversion of the operator Λ associated with a complicated quasi-periodic problem.

Now, (4.1), (4.2), and (4.5) yield the following result.

Theorem 4.2. *The error $e = v - u$ is subject to the estimate*

$$\|e\|_\circ \in \left[\max\left\{0, \frac{\delta_\rho - M_\oplus(\eta_\rho, \tau_y)}{1 + q(\rho)}\right\}, \frac{\delta_\rho + M_\oplus(\eta_\rho, \tau_y)}{1 - q(\rho)}\right], \tag{4.6}$$

where $M_\oplus$ is defined by (4.5), *$\tau_y := y - \rho\Lambda Qv$, and y is a function in Y.*

Remark 4.3. Here η_ρ and δ_ρ are directly computable and $q(\rho)$ is defined in accordance with relations presented in the previous section. Hence the cost of (4.6) is mainly related to the quantity $M_\oplus(\eta_\rho, \tau_y)$, which is an a posteriori error majorant of the functional type. The derivation of such estimates is performed by purely functional methods and does not exploit special features of approximations (e.g., Galerkin orthogonality), numerical method, and exact solution (e.g., extra regularity). Properties of the majorants are well studied (see [25, 26] and the literature cited therein). Numerous tests performed for different boundary value problems have confirmed high practical efficiency of error majorants of the functional type. It was shown that $M_\oplus$ is a guaranteed and efficient majorant of the global error and generates good indicators of local errors if y is replaced by a certain numerical reconstruction of the exact dual solution. There are many different ways to obtain suitable reconstructions with minimal expenditures (concerning this point we refer to [21] where the reader will find a systematic discussion of computational aspects in the context of various boundary value problems). Error majorants of this type can be also used for the evaluation of modeling errors (see [27, 28]).

Usually, the cost of a good estimate (with the efficiency index between 1 and 2) is comparable with the cost of a numerical solution. However, the proportion essentially depends on the numerical method used. For the classical FEM schemes the expenditures are maximal (because this method generates rather coarse approximations of fluxes). For the dual mixed method, finite volume method, isogeometric approximations, and other methods producing locally equilibrated fluxes, the expenditures may be two to three times smaller than for the numerical solution.

4.2 Examples

Now we shortly discuss applications of Theorem 4.2 to problems where Q and Q^* are defined by the operators ∇ and div, respectively, $\Lambda_\circ = a_\circ(x)\mathbb{I}$, $\Lambda = a(x)\mathbb{I}$, $x \in \Omega$, and $V = \mathring{H}^1(\Omega)$.

***d* = 1.** Let $\Omega = (0, 1)$. Equation (1.1) has the form $(a(x)u')' - f = 0$. In this case, $Qw = w'$, $Q^*y = -y'$, and (4.3) is reduced to

$$\int_0^1 a_\circ(v_\rho - v)'w'\,dx + \rho\int_0^1 (av'w' + fw)\,dx = 0. \tag{4.7}$$

In order to apply Theorem 4.2, we set $y = \rho(g(x) + \mu)$, where $g(x) = -\int_0^x f dx$ and μ is a constant. Then $-y' - \rho f = 0$ and $\tau = \rho(g(x) + \mu) - \rho a v' = \rho(\mu + g - av')$. The best constant μ is defined by minimization of $M^2_\oplus(\eta_\rho, \tau)$, which has the form

$$\int_0^1 (a_\circ(\eta'_\rho)^2 + a_\circ^{-1}\rho^2(\mu + g - av')^2 + 2\eta'_\rho\rho(\mu + g - av'))\,dx$$

Since $\int_0^1 \eta'_\rho\,dx = 0$, the problem is reduced to minimization of the second term and the best μ satisfies the equation

$$\int_0^1 a_\circ^{-1}(\mu + g(x) - av')\,dx = 0.$$

Hence

$$\mu = \bar\mu := \frac{\int_0^1 a_\circ^{-1}(av' - g)\,dx}{\int_0^1 a_\circ^{-1}\,dx},$$

and (4.6) yields the estimate

$$\|e\|_\circ \in \Big[\max\Big\{0, \frac{\delta_\rho - I_\oplus(v, \tilde v_\rho)}{1 + q(\rho)}\Big\}, \frac{\delta_\rho + I_\oplus(v, \tilde v_\rho)}{1 - q(\rho)}\Big], \tag{4.8}$$

where

$$I^2_\oplus(v, \tilde v_\rho) = \int_0^1 a_\circ^{-1}\big(a_\circ(v - \tilde v_\rho)' + \rho(\bar\mu + g - av')\big)^2\,dx.$$

Here v and $\tilde v_\rho$ are two consequent numerical approximations (e.g., finite element approximations $v_{k,h}$ and $v_{k+1,h}$ computed on a mesh $\mathfrak{I}_h$). Then

$$\eta_\rho = \eta_{k,h} := v_{k,h} - v_{k+1,h} \quad \text{and} \quad \delta_\rho = \delta_{k,h} := \|v_{k,h} - v_{k+1,h}\|_\circ$$

are directly computable. Since $a_\circ$ is a "simple" function, the integrals

$$F_1 = \int_0^1 a_\circ^{-1}\,dx, \quad F_2 = \int_0^1 a_\circ^{-1} g dx, \quad F_3 = \int_0^1 a_\circ(\eta'_{k,h})^2\,dx,$$
$$F_4 = \int_0^1 a_\circ^{-1}(\bar\mu + g)^2\,dx, \quad F_5 = \int_0^1 (\bar\mu + g)\eta'_{k,h}\,dx$$

are easy to compute. Other integrals

$$G_1 = \int_0^1 a_\circ^{-1} a v'_{k,h}\,dx, \quad G_2 = \int_0^1 a v'_{k,h}\eta'_{k,h}\,dx,$$
$$G_3 = \int_0^1 (\bar\mu + g)a_\circ^{-1} a v'_{k,h}\,dx, \quad G_4 = \int_0^1 a_\circ^{-1} a^2 (v'_{k,h})^2\,dx$$

contain a highly oscillating coefficient a multiplied by piecewise polynomial mesh functions. If a and f have low QTT rank tensor representations [15], then the integrals can be efficiently computed by tensor-type methods discussed in [17] (see also Section 5 below). We have

$$I^2_{\oplus}(v, \bar{v}_\rho) = F_3 + 2\rho G_2 + 2\rho F_5 + \rho^2(F_4 - 2G_3 + G_4) =: \varepsilon^2_{k,h}, \quad \bar{\mu} = \frac{G_1 - F_2}{F_1}.$$

Notice that the quantity $\varepsilon_{k,h}$ is the value of the majorant, where the flux has been selected in the best way. It is not difficult to show that

$$a_\circ(v'_\rho - v') = \rho(g + \bar{\mu}) - av'$$

and, therefore,

$$I^2_{\oplus}(v, \bar{v}_\rho) = \int_0^1 a_\circ(v'_\rho - \bar{v}'_\rho)^2 \, dx.$$

In other words, this term coincides with the error of the Galerkin solution related to the simplified boundary value problem (4.7), where $v = v_{k,h}$. In accordance with Section 3, we set $\rho = 2/(\mathbf{h}_\ominus + \mathbf{h}_\oplus)$ and find that $q = (\mathbf{h}_\oplus - \mathbf{h}_\ominus)/(\mathbf{h}_\ominus + \mathbf{h}_\oplus)$. Now (4.8) yields easily computable lower and upper bounds of the error encompassed in $v_{k,h}$:

$$\frac{\delta_{k,h} - \varepsilon_{k,h}}{1 + q} \le \|v_{k,h} - u\|_\circ \le \frac{\delta_{k,h} + \varepsilon_{k,h}}{1 - q}. \tag{4.9}$$

Remark 4.4. It is worth adding comments on convergence properties of the quantities $\delta_{k,h}$ and $\varepsilon_{k,h}$ entering (4.9). If the mesh $\mathcal{T}_h$ is fixed, then $v_{k,h}$ tends to the Galerkin approximation u_h of problem (1.1) on this mesh. This fact follows from Theorem 2.1 applied to the case where the iterations are performed on the respective finite-dimensional space V_h (considered as the space V). Then, $\|v_{k,h} - u_h\|_\circ \le q^k \|v_{0,h} - u_h\|$ and for any h the term $\delta_{k,h}$ tends to zero with the geometric rate. The quantity $\varepsilon_{k,h}$ is equal to the error of the Galerkin solution to the simplified problem. It has different asymptotic properties. It mainly depends on $\mathcal{T}_h$, and for a given mesh it does not tend to zero when $k \to +\infty$. However, for any given v (which in our example is defined by $v_{k,h}$) this term goes to zero if $h \to 0$ provided that the mesh satisfies the standard regularity conditions. The problem with $a_\circ$ is assumed to be much more regular than the problem with a. Therefore, in terms of h the approximation error $\varepsilon_{k,h}$ (associated with $a_\circ$) will decrease faster than the analogous error in the original problem (e.g., for $a_\circ = \text{const}$, the term $\varepsilon_{k,h}$ is proportional to h).

Since both quantities $\delta_{k,h}$ and $\varepsilon_{k,h}$ are explicitly known, estimate (4.9) (and other analogous estimates) contains a very useful information unavailable in the context of purely asymptotic error analysis. Using this information, we can organize the computational process in the best possible way by comparing iteration and discretization errors. In this process, rapidly converging iterations with respect to k should be continued until $\delta_{k,h} > \varepsilon_{k,h}$. If $\delta_{k,h} \approx \varepsilon_{k,h}$, then further iterations on the mesh $\mathcal{T}_h$ are unable to essentially improve the numerical solution. Instead, we should refine $\mathcal{T}_h$, project $u_{k,h}$ on the refined mesh, and use it as a starting point for a new series of iterations generated by problem (4.7). For each step, we compute the right-hand side of (4.9) and stop the process when it becomes smaller than the desired tolerance.

***d* = 2.** The computation of $M_\oplus$ for 2D problems can be also reduced to the computation of one-dimensional integrals. Certainly, on the multidimensional case the amount of integrals is much larger. However, the basic tensor decomposition methods remain the same. Below we briefly discuss them with the paradigm of a simple case where

$$f = f^{(1)}(x_1)f^{(2)}(x_2) \quad \text{and} \quad a = a^{(1)}(x_1)a^{(2)}(x_2).$$

Assume that approximations are represented in the form of series formed by one-dimensional functions $\phi^{(1)}_i$ and $\phi^{(2)}_j$ (which may be supported locally or globally), so that

$$v = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \gamma_{ij}\phi^{(1)}_i(x_1)\phi^{(2)}_j(x_2), \quad \bar{v}_\rho = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \bar{\gamma}_{ij}\phi^{(1)}_i(x_1)\phi^{(2)}_j(x_2).$$

In this case,

$$\nabla\eta_\rho = \Big(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\varsigma_{ij}\frac{\partial\phi_i^{(1)}}{\partial x_1}\phi_j^{(2)}, \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\varsigma_{ij}\phi_i^{(1)}\frac{\partial\phi_j^{(2)}}{\partial x_2}\Big), \quad \text{where } \varsigma_{ij} = \gamma_{ij} - \widetilde{\gamma}_{ij}.$$

We define another set of one-dimensional functions $W_k^{(1)}(x_1)$ and $W_l^{(2)}(x_2)$, which form the vector function

$$y = \Upsilon_0 + \sum_{k=1}^{m_1}\sum_{l=1}^{m_2}\sigma_{kl}\Upsilon_{kl}, \quad \Upsilon_{kl} = \Big\{W_k^{(1)}\frac{\partial W_l^{(2)}}{\partial x_2}; -\frac{\partial W_k^{(1)}}{\partial x_1}W_l^{(2)}\Big\}. \tag{4.10}$$

Here Υ_0 is a given function, which can be defined in different ways. In particular, we set

$$\Upsilon_0 = \{W_0^{(1)}(x_1)W_0^{(2)}(x_2); 0\}, \quad W_0^{(1)}(x_1) = \int_0^{x_1} f^{(1)}\,dx_1, \quad W_0^{(2)} = -\rho f^{(2)}.$$

The functions Υ_{kl} must satisfy the usual linear independence conditions in order to guarantee unique solvability of the respective approximation problem. For any smooth function w vanishing on $\partial\Omega$, we have

$$\int_\Omega (\Upsilon_0\cdot\nabla w - \rho f w)\,dx_1\,dx_2 = 0 \quad \text{and} \quad \int_\Omega \Upsilon_{kl}\cdot\nabla w\,dx_1\,dx_2 = 0.$$

Thus, $\|Q^*y - \rho f\| = 0$ and we can use the simplified form of $M_\oplus$.

In the simplest case $\Lambda_\circ = a_\circ\mathbb{I}$, where $a_\circ$ is a constant. The best y minimizes the quantity

$$\begin{aligned} M_\oplus^2(\eta_\rho,\tau) = {}& \int_\Omega a_\circ\nabla\eta_\rho\cdot\nabla\eta_\rho\,dx + \int_\Omega a_\circ^{-1}y\cdot y\,dx + \rho^2\int_\Omega a_\circ^{-1}a^2\nabla v\cdot\nabla v\,dx \\ & - 2\int_\Omega(\rho a_\circ^{-1}a\nabla v + \nabla\eta_\rho)\cdot y\,dx + 2\rho\int_\Omega a\nabla\eta_\rho\cdot\nabla\eta_\rho\,dx, \end{aligned} \tag{4.11}$$

which shows that y must satisfy the relation $y = \rho a\nabla v + a_\circ\nabla\eta_\rho$. We select σ_{kl} that defines the Galerkin approximation of this function, and we arrive at the system

$$\begin{aligned} & \sum_{k=1}^{m_1}\sum_{l=1}^{m_2}\sigma_{kl}\int_\Omega \Upsilon_{kl}\cdot\Upsilon_{st}\,dx_1\,dx_2 + \int_\Omega \Upsilon_0\cdot\Upsilon_{st}\,dx_1\,dx_2 \\ & \qquad = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\int_\Omega(\rho a\gamma_{ij} + a_\circ\varsigma_{ij})\Big(\frac{\partial\phi_i^{(1)}}{\partial x_1}\phi_j^{(2)}, \phi_i^{(1)}\frac{\partial\phi_j^{(2)}}{\partial x_2}\Big)\cdot\Upsilon_{st}\,dx_1\,dx_2. \end{aligned} \tag{4.12}$$

Introduce the following matrices:

$$\mathrm{D}^{(1)} = \{D_{kl}^{(1)}\}, \quad D_{kl}^{(1)} = \int_0^a \frac{\partial W_k^{(1)}}{\partial x_1}\frac{\partial W_l^{(1)}}{\partial x_1}\,dx_1, \qquad \mathrm{W}^{(1)} = \{W_{kl}^{(1)}\}, \quad W_{kl}^{(1)} = \int_0^a W_k^{(1)}W_l^{(1)}\,dx_1,$$

$$\mathrm{D}^{(2)} = \{D_{kl}^{(2)}\}, \quad D_{kl}^{(2)} = \int_0^b \frac{\partial W_k^{(2)}}{\partial x_2}\frac{\partial W_l^{(2)}}{\partial x_2}\,dx_2, \qquad \mathrm{W}^{(2)} = \{W_{kl}^{(2)}\}, \quad W_{kl}^{(2)} = \int_0^b W_k^{(2)}W_l^{(2)}\,dx_2,$$

$$\mathrm{F}^{(1)} = \{F_{ik}^{(1)}\}, \quad F_{ik}^{(1)} = \int_0^a \frac{\partial\phi_i^{(1)}}{\partial x_1}W_k^{(1)}\,dx_1, \qquad \mathrm{G}^{(1)} = \{G_{ik}^{(1)}\}, \quad G_{ik}^{(1)} = \int_0^a \phi_i^{(1)}\frac{\partial W_k^{(1)}}{\partial x_1}\,dx_1,$$

$$\mathrm{F}^{(2)} = \{F_{jl}^{(2)}\}, \quad F_{jl}^{(2)} = \int_0^b \phi_j^{(2)}\frac{\partial W_l^{(2)}}{\partial x_2}\,dx_2, \qquad \mathrm{G}^{(2)} = \{G_{jl}^{(2)}\}, \quad G_{jl}^{(2)} = \int_0^b \frac{\partial\phi_j^{(2)}}{\partial x_2}W_l^{(1)}\,dx_2,$$

$$\widehat{\mathrm{F}}^{(1)} = \{\widehat{F}_{ik}^{(1)}\}, \quad \widehat{F}_{ik}^{(1)} = \int_0^a a_1(x_1)\frac{\partial\phi_i^{(1)}}{\partial x_1}W_k^{(1)}\,dx_1, \quad \widehat{\mathrm{G}}^{(1)} = \{\widehat{G}_{ik}^{(1)}\}, \quad \widehat{G}_{ik}^{(1)} = \int_0^a a_1(x_1)\phi_i^{(1)}\frac{\partial W_k^{(1)}}{\partial x_1}\,dx_1,$$

$$\widehat{\mathrm{F}}^{(2)} = \{\widehat{F}_{jl}^{(2)}\}, \quad \widehat{F}_{jl}^{(2)} = \int_0^b a_2(x_2)\phi_j^{(2)}\frac{\partial W_l^{(2)}}{\partial x_2}\,dx_2, \quad \widehat{\mathrm{G}}^{(2)} = \{\widehat{G}_{jl}^{(2)}\}, \quad \widehat{G}_{jl}^{(2)} = \int_0^b a_2(x_2)\frac{\partial\phi_j^{(2)}}{\partial x_2}W_l^{(1)}\,dx_2,$$

and vectors

$$\mathbf{g}^{(1)} = \{g_k^{(1)}\}, \quad g_k^{(1)} = \int_0^a W_0^{(1)} W_k^{(1)}\, dx_1, \quad \mathbf{g}^{(2)} = \{g_l^{(2)}\}, \quad g_l^{(2)} = \int_0^b W_0^{(2)} \frac{\partial W_l^{(2)}}{\partial x_2}\, dx_2.$$

Notice that all coefficients are presented by one-dimensional integrals, which can be efficiently computed with the help of special (tensor-type) methods (see, e.g., [15–19]).

It is not difficult to see that

$$Y_{klst} := \int_\Omega \Upsilon_{kl} \cdot \Upsilon_{st}\, dx = W_{ks}^{(1)} D_{lt}^{(2)} + D_{ks}^{(1)} W_{lt}^{(2)}$$

and

$$\int_\Omega \Upsilon_0 \cdot \Upsilon_{st}\, dx_1\, dx_2 = \int_\Omega W_0^{(1)} W_s^{(1)} W_0^{(2)} \frac{\partial W_t^{(2)}}{\partial x_2}\, dx_1\, dx_2 = g_s^{(1)} g_t^{(2)},$$

where $\mathrm{Y} = \{Y_{klst}\}$ is the fourth-order tensor. Hence the left-hand side of the system (4.12) has the form $\mathrm{Y}\boldsymbol{\sigma} + \mathbf{g}^{(1)} \otimes \mathbf{g}^{(2)}$. In the right-hand side we have the term

$$\int_\Omega a_\circ \varsigma_{ij}\Big(\frac{\partial \phi_i^{(1)}}{\partial x_1}\phi_j^{(2)}, \phi_i^{(1)}\frac{\partial \phi_j^{(2)}}{\partial x_2}\Big) \cdot \Upsilon_{st}\, dx_1\, dx_2 = a_\circ \mathrm{H}\boldsymbol{\varsigma},$$

where $\mathrm{H} = \{H_{ijst}\}$, $H_{stij} = F_{is}^{(1)} F_{jt}^{(2)} - G_{is}^{(1)} G_{jt}^{(2)}$. Another term is

$$\int_\Omega \rho a \gamma_{ij}\Big(\frac{\partial \phi_i^{(1)}}{\partial x_1}\phi_j^{(2)}, \phi_i^{(1)}\frac{\partial \phi_j^{(2)}}{\partial x_2}\Big) \cdot \Upsilon_{st}\, dx_1\, dx_2 = \widehat{\mathrm{H}}\boldsymbol{\gamma},$$

where $\widehat{\mathrm{H}} = \{\widehat{H}_{ijst}\}$, $\widehat{H}_{stij} = \widehat{F}_{is}^{(1)} \widehat{F}_{jt}^{(2)} - \widehat{G}_{is}^{(1)} \widehat{G}_{jt}^{(2)}$.

Now (4.12) implies

$$\boldsymbol{\sigma} = \mathrm{Y}^{-1}(\widehat{\mathrm{H}}\boldsymbol{\gamma} + a_\circ \mathrm{H}\boldsymbol{\varsigma} - \mathbf{g}^{(1)} \otimes \mathbf{g}^{(2)})$$

and the value of $M_\oplus$ is obtained by (4.6), (4.10), and (4.11).

5 Low-Rank Solution of the Discrete Equation

We consider the following elliptic diffusion equation with quasi-periodic coefficient $a(x) > 0$ (whose oscillations are characterized by the parameter ϵ):

$$\mathcal{A}u = -\operatorname{div}(a(x)\nabla u) = f(x), \quad x = (x_1, \dots, x_d) \in \Omega = (0,1)^2, \quad u_{|\Gamma} = 0, \tag{5.1}$$

where the function f corresponds to the modified right-hand side in problem (4.3). In this case $\Gamma = \partial\Omega$, $\Lambda = a\mathbb{I}$, $Qw = \nabla w$, and $Q^* y = -\operatorname{div} y$.

In what follows we assume that f and a admit low-rank representation i.e.,

$$f = \sum_{i=1}^{R_f} f_1^i(x_1) f_2^i(x_2), \quad a = \sum_{j=1}^{R_a} a_1^j(x_1) a_2^j(x_2),$$

where the parameters R_f and R_a are called the separation rank. Then one may assume that the exact FEM solution can be well approximated by $u^K(x) = \sum_{j=1}^K u_1^j(x_1) u_2^j(x_2)$, where K depends on the separation rank of f and a. In some cases this important property can be rigorously proven (e.g., for the Laplacian and other closely related operators). Similar low-rank approximations can be observed for the QTT tensor approximations (see [17]). Existence of a low-rank solution means that for some moderate K we have $u_K \approx u$ up to the rank truncation threshold.

First, we sketch the rank-structured computational scheme. In our set of examples the original problem is to find u such that

$$\int_\Omega a(x)\nabla u \cdot \nabla w\, dx = \int_\Omega fw\, dx \quad \text{for all } w \in V_0 := H_0^1(\Omega).$$

It is approximated by the following Galerkin problem for the low-rank representation u^K:

$$\int_\Omega a(x)\nabla u^K \cdot \nabla w^K\, dx = \int_\Omega fw^K\, dx \quad \text{for all } w^K \in V_0^K, \tag{5.2}$$

where V_0^K is a subset of V_0 formed by functions of the type

$$w^K(x) = \sum_{j=1}^{K} \phi_1^j(x_1)\phi_2^j(x_2).$$

Therefore, in terms of the general scheme exposed in the introduction, the problem $\mathcal{P}$ is now problem (5.2) and we solve it by iterations with the help of the simplified (preconditioned) problem

$$\int_\Omega a_\circ(x)\nabla u_k^K \cdot \nabla w^K\, dx = \int_\Omega f_{k-1}w^K\, dx \quad \text{for all } w^K \in V_0^K, \tag{5.3}$$

where f_{k-1} depends on u_{k-1}^K and $a_\circ$ is a simple function (i.e., it is representable by a sum of terms $a_1^\circ(x_1)\dots a_2^\circ(x_2)$ with very simple multipliers).

Given the right-hand side, problem (5.3) is much simpler than the initial problem and the stiffness matrix associated with (5.3) is computed much easier and has a simpler (low Kronecker rank) form that allows a rank-structured representation of its inverse.

5.1 Kronecker Product Representation of the Stiffness Matrix

Figure 3 illustrates a 2D example of the $L \times L$ periodic coefficient with $L = 6$ corresponding to the choice $\epsilon = \frac{1}{L}$. In this example, the scalar coefficient is represented by the separable function

$$a(x) = C + a_1(x_1)a_1(x_2), \quad C > 0,$$

where the generating univariate function $a_1(x_1)$ has the shape of six uniformly distributed bumps of height 1 as shown in Figure 3, right. Figure 3, left, presents the oscillating part of a 2D coefficients function, which is $a_1(x_1)a_1(x_2)$. Here, the coefficient bumps are displaced on the coarse grid of size $8L \times 8L$ in such a way that bumps occupy the 4×4 central box in each of the 8×8 cells, which compose the whole $L \times L$ lattice-type decomposition of Ω (we have $L = 6$ in Figure 3, i.e., the size of the coarse grid is 48×48, while $L = 12$ in Figures 4 and 5, i.e., the size of the coarse grid is 96×96). Hence the axis scale 20, 40, 60, 80 denotes the coarse grid in both x_1 and x_2 that describes the construction of coefficient in detail.

The examples of other possible shapes of the equation coefficient corresponding to the cases (i), (ii) and (iii) specified in Section 1 are presented in Figures 1 and 2.

We apply the FEM Galerkin discretization of equation (5.1) by means of tensor-product piecewise affine basis functions (instead of "linear finite elements")

$$\{\varphi_{\mathbf{i}}(x) := \varphi_{i_1}(x_1)\cdots\varphi_{i_d}(x_d)\}, \quad \mathbf{i} = (i_1, \dots, i_d),\ i_\ell \in \mathcal{I}_\ell = \{1, \dots, n_\ell\},\ \ell = 1, \dots, d,$$

where φ_{i_k} are 1D finite element basis functions (say, piecewise linear hat functions).

We associate the univariate basis functions with the uniform grid $\{v_j\}$, $j = 1, \dots, n_\ell$, on $[0, 1]$ with the mesh size $h = 1/(n_\ell + 1)$. In this construction we have $N = n_1 n_2 \dots n_d$ basis functions $\varphi_{\mathbf{i}}$. Notice that the univariate grid size n_ℓ is of the order of $n_\ell = O(\frac{1}{\epsilon})$ designating the total problem size $N = O(\frac{1}{\epsilon^d})$.

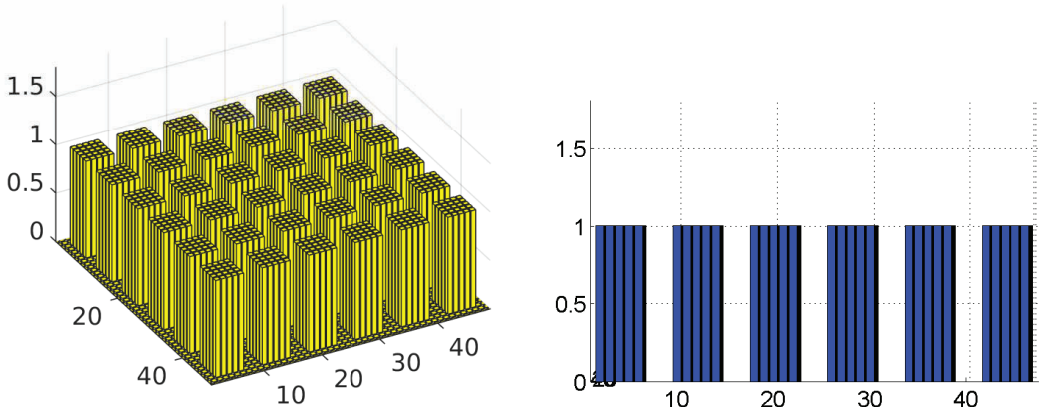

Figure 3. Example of the 2D periodic oscillating coefficients (left) and the respective 1D factor $a_1(x_1)$.

For ease of exposition we first consider the case $d = 2$, and further assume that the scalar diffusion coefficient $a(x_1, x_2)$ can be represented in the form

$$a(x_1, x_2) = \sum_{k=1}^{R} a_k^{(1)}(x_1) a_k^{(2)}(x_2) > 0$$

with a small rank parameter R.

The $N \times N$ stiffness matrix is constructed by the standard mapping of the multi-index $\mathbf{i}$ into the N-long univariate index i representing all degrees of freedom. For instance, we use the so-called big-endian convention for $d = 3$ and $d = 2$:

$$\mathbf{i} \mapsto i := i_3 + (i_2 - 1)n_3 + (i_1 - 1)n_2 n_3, \quad \mathbf{i} \mapsto i := i_2 + (i_1 - 1)n_2,$$

respectively. Hence all matrices and vectors are defined on the long index i as usual, however, the special Kronecker structure allows the low-storage and low-complexity matrix vector multiplications when appropriate, i.e., when a vector also admits the low-rank Kronecker form representation. In particular, the basis function $\varphi_{\mathbf{i}}$ is designated via the long index, i.e., $\varphi_i = \varphi_{\mathbf{i}}$.

First, we consider the simplest case $R = 1$ and let $d = 2$. We construct the Galerkin stiffness matrix $A = [a_{ij}] \in \mathbb{R}^{N \times N}$ in the form of a sum of Kronecker products of small "univariate" matrices. Recall that given $p_1 \times q_1$ matrix A and $p_2 \times q_2$ matrix B, their Kronecker product is defined as a $p_1 p_2 \times q_1 q_2$ matrix C via the block representation

$$C = A \otimes B = [a_{ij} B], \quad i = 1, \ldots, p_1, \ j = 1, \ldots, q_1.$$

We say that the Kronecker rank of the matrix A in the representation above equals 1. Now the elements of the Galerkin stiffness matrix take the form

$$\begin{aligned} a_{ij} = \langle \mathcal{A}\varphi_i, \varphi_j \rangle &= \int_\Omega a^{(1)}(x_1) a^{(2)}(x_2) \nabla \varphi_i(x) \dot{\nabla} \varphi_j(x) dx \\ &= \int_0^1 a^{(1)}(x_1) \frac{\partial \varphi_{i_1}(x_1)}{\partial x_1} \frac{\partial \varphi_{j_1}(x_1)}{\partial x_1} dx_1 \int_0^1 a^{(2)}(x_2) \varphi_{i_2}(x_2) \varphi_{j_2}(x_2) dx_2 \\ &\quad + \int_0^1 a^{(1)}(x_1) \varphi_{i_1}(x_1) \varphi_{j_1}(x_1) dx_1 \int_0^1 a^{(2)}(x_2) \frac{\partial \varphi_{i_2}(x_2)}{\partial x_2} \frac{\partial \varphi_{j_2}(x_2)}{\partial x_2} dx_2, \end{aligned}$$

which leads to the rank-2 Kronecker product representation

$$A = [a_{ij}] = A_1 \otimes M_2 + M_1 \otimes A_2,$$

where $\otimes$ denotes the conventional Kronecker product of matrices. Here $A_1 = [a_{i_1 j_1}] \in \mathbb{R}^{n_1 \times n_1}$ and $A_2 = [a_{i_2 j_2}] \in \mathbb{R}^{n_2 \times n_2}$ denote the univariate stiffness matrices and $M_1 = [m_{i_1 j_1}] \in \mathbb{R}^{n_1 \times n_1}$ and $M_2 = [m_{i_2 j_2}] \in \mathbb{R}^{n_2 \times n_2}$ define the corresponding weighted mass matrices, e.g.,

$$a_{i_1 j_1} = \int_0^1 a^{(1)}(x_1) \frac{\partial \varphi_{i_1}(x_1)}{\partial x_1} \frac{\partial \varphi_{j_1}(x_1)}{\partial x_1} dx_1, \quad m_{i_1 j_1} = \int_0^1 a^{(1)}(x_1) \varphi_{i_1}(x_1) \varphi_{j_1}(x_1) dx_1.$$

By simple algebraic transformations (e.g., by lamping of the tri-diagonal mass matrices, which does not effect the approximation order of the FEM discretization) the matrix A can be simplified to the form

$$A \mapsto A = A_1 \otimes D_2 + D_1 \otimes A_2, \tag{5.4}$$

where D_1, D_2 are the diagonal matrices. The matrix A corresponds to the FEM discretization of the initial elliptic PDE with complicated highly oscillating coefficients.

The simple choice of the spectrally equivalent preconditioner $A_\circ$ corresponds to the operator Laplacian. In this case the representation in (5.4) is simplified to the discrete Laplacian matrix in the form of rank-2 Kronecker sum

$$A \mapsto A_\circ = A_1 \otimes I_2 + I_1 \otimes A_2, \tag{5.5}$$

where I_1 and I_2 denote the identity matrices of the corresponding size. Here the simple tree-diagonal matrices A_1 and A_2 represent the FEM/FDM Laplacian in 1D. This matrix will be used in what follows as a prototype preconditioner for solving the linear system of equations

$$A\mathbf{u} = \mathbf{f}. \tag{5.6}$$

The matrix A is constructed in general for the R-term separable coefficient $a(x_1, x_2)$ with $R \geq 1$ which leads to the rank-$2R$ Kronecker sum representation

$$A = \sum_{k=1}^{R} [A_{1,k} \otimes D_{2,k} + D_{1,k} \otimes A_{2,k}],$$

with matrices of the respective size.

5.2 On Existence of the Low-Rank Solution

In this paper we discuss the approach based on the low-rank separable ϵ-approximation of the solution to equation (5.6) that is considered as the d-dimensional real-valued array $\mathbf{u} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$. In general, for the case $R > 1$ this favorable property is not guaranteed by the low Kronecker rank representation to the Galerkin system matrix A, discussed in Section 5.1.

Let $R = 1$ and $d = 2$. The existence of the low-rank approximation to the solution of equation (5.6) with the low-rank right-hand side

$$\mathbf{f} = \sum_{k=1}^{R_f} \mathbf{f}_k^{(1)} \otimes \mathbf{f}_k^{(2)}, \quad \mathbf{f}_k^{(\ell)} \in \mathbb{R}^{n_\ell},$$

and with the system matrix in the form (5.5) can be justified by plugging the representation (5.5) in the sinc-quadrature approximation to the Laplace integral transform [5]

$$\Lambda_\circ^{-1} = \int_{\mathbb{R}_+} e^{-t\Lambda_\circ} dt \approx B_M := \sum_{k=-M}^{M} c_k e^{-t_k \Lambda_\circ} = \sum_{k=-M}^{M} c_k e^{-t_k A_1} \otimes e^{-t_k A_2}, \tag{5.7}$$

taking into account that the matrices A_1 and A_2 commute with I_1 and I_2, respectively. Hence equation (5.7) represents the accurate rank-$(2M+1)$ Kronecker product approximation to the preconditioner $\Lambda_\circ^{-1}$ which can be applied directly to the right-hand side to obtain

$$\mathbf{u} = \Lambda_\circ^{-1} \mathbf{f} \approx B_M \mathbf{f} = \sum_{k=-M}^{M} c_k \sum_{m=1}^{R_f} e^{-t_k A_1} \mathbf{f}_m^{(1)} \otimes e^{-t_k A_2} \mathbf{f}_m^{(2)}.$$

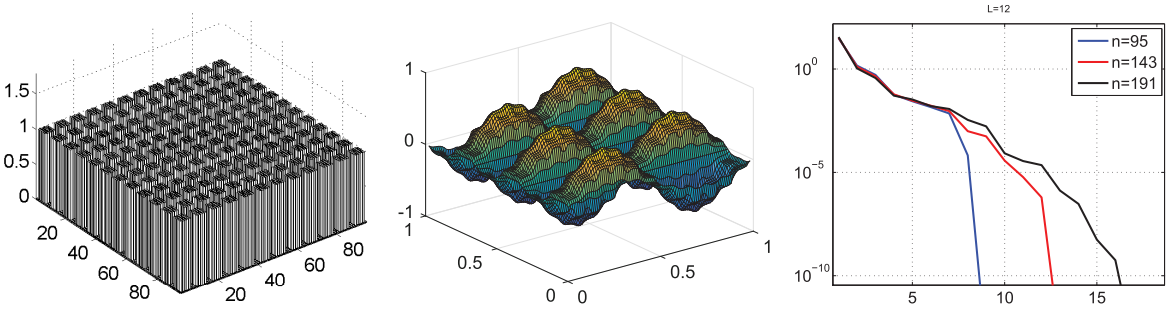

Figure 4. Rank decomposition of the solution for the 12×12 periodic coefficient.

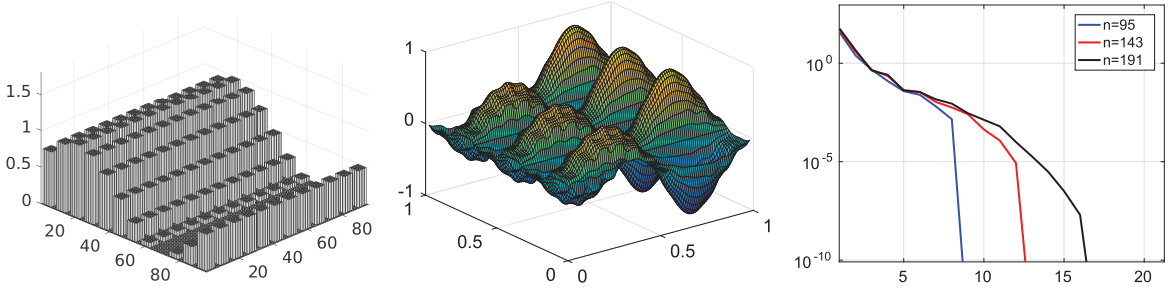

Figure 5. Rank decomposition of the solution for the 12×12 modulated periodic coefficient.

The numerical efficiency of the representation (5.7) can be explained by the fact that the quadrature parameters t_k, c_k can be chosen in such a way that the low Kronecker rank approximation B_M converges to $\Lambda_\circ^{-1}$ exponentially fast in M. For example, under the choice $t_k = e^{kh}$, $c_k = ht_k$ with $h = \pi/\sqrt{M}$ there holds [5]

$$\|\Lambda_\circ^{-1} - B_M\| \le Ce^{-\beta\sqrt{M}}\|\Lambda_\circ^{-1}\|,$$

in the Frobenius norm, which means that the approximation error $\epsilon > 0$ can be achieved with the number of terms $R_B = 2M + 1$ of the order of $R_B = O(|\log \epsilon|^2)$.

Figures 4 and 5 demonstrate the singular values of the discrete solution on the $n \times n$ grid for $n = 95, 143$ and 191, indicating very moderate dependence of the ϵ-rank on the grid size n. As in the case of Figure 3, in Figures 4 and 5 we only represent the oscillating part of the coefficients and omit the small constant $C > 0$.

Further enhancement of the tensor approximation can be based on the application of the quantized-TT (QTT) tensor approximation which has been already applied in [17] to the 1D equations with quasi-periodic coefficients. The power of the QTT approximation method is due to the perfect low-rank decompositions applied to the wide class of function-related tensors [15]. See [17] for a more detailed discussion and a number of numerical examples.

One can apply QTT approximations to problems with quasi-periodic coefficients, which can be described by oscillation with smooth modulation around a constant value, oscillation around a given smooth function, or oscillation around a piecewise constant function, see Figure 1 and examples in [17].

Let the vector $\mathbf{x} \in \mathbb{C}^N$, $N = 2^L$, be obtained by sampling a continuous function $f \in C[0, 1]$ (or even piecewise smooth functions), on the uniform grid of size N. For the following examples of univariate functions the explicit QTT-rank estimates of the corresponding QTT tensor representations are valid uniformly in the vector size N, see [15]:

(A) $r = 1$ for complex exponentials, $f(x) = e^{i\omega x}$, $\omega \in \mathbb{R}$.

(B) $r = 2$ for trigonometric functions, $f(x) = \sin \omega x$, $f(x) = \cos \omega x$, $\omega \in \mathbb{R}$.

(C) $r \le m + 1$ for polynomials of degree m.

(D) For a function f with the QTT-rank r_0 modulated by another function g with the QTT-rank r (say, step-type function, plain wave, polynomial) the QTT rank of a product fg is bounded by a multiple of r and r_0,

$$\mathrm{rank}_{\mathrm{QTT}}(fg) \le \mathrm{rank}_{\mathrm{QTT}}(f)\,\mathrm{rank}_{\mathrm{QTT}}(g).$$

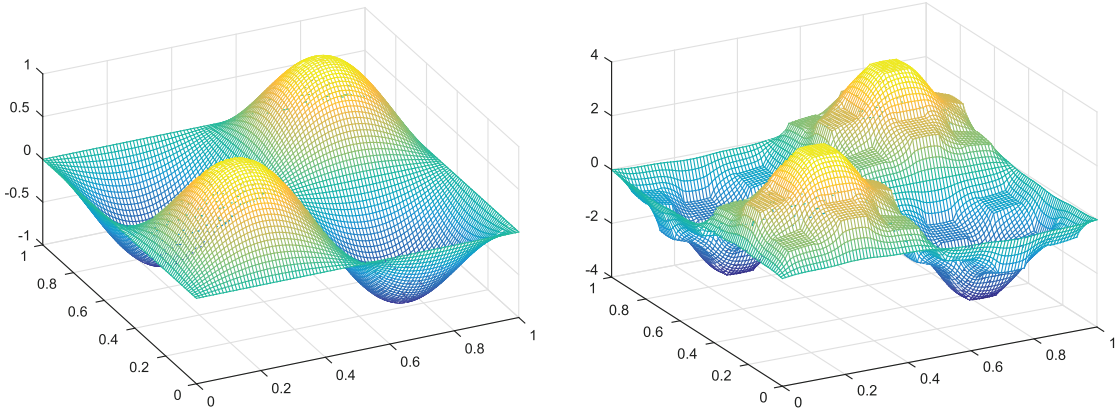

Figure 6. The right-hand side and solution for periodic oscillating coefficients shown in Figure 3.

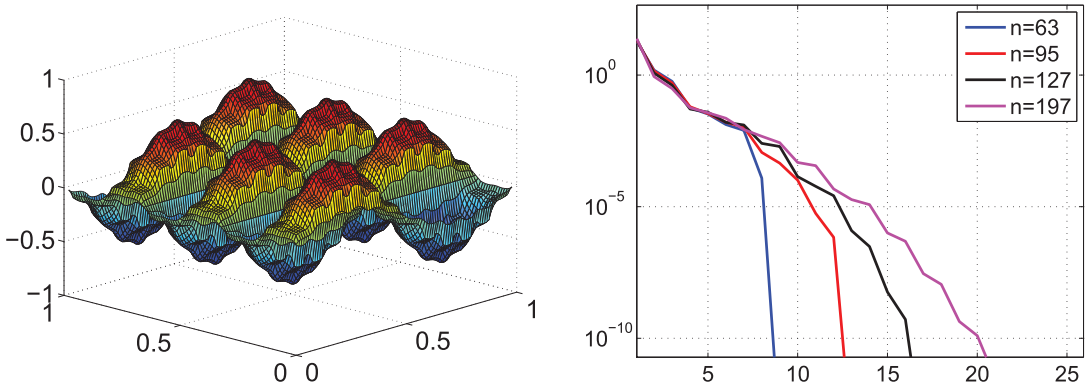

Figure 7. Accuracy of the rank decomposition of the solution vs. rank parameter for the 8×8 periodic coefficient and grid size $n \times n$.

(E) Furthermore, the following result holds [11]: the QTT rank for the periodic amplification of a reference function on a unit cell to a rectangular lattice is of the same order as that for the reference function.

The rank of the QTT tensor representation to the 1D Galerkin FEM matrix in the case of oscillating coefficients was discussed in [4, 17].

5.3 Numerical Test on the Rank Decomposition of u

Figure 6 represents the right-hand side $f_1(x_1, x_2)$ and the respective solution for the discretization to equation (5.1) (with the coefficient depicted in Figure 3) on a 400×400-grid, where

$$f_1(x_1, x_2) = \sin(2x_1)\sin(2x_2).$$

The PCG solver for the system of equations (5.6) with the discrete Laplacian inverse as the preconditioner demonstrates robust convergence with the rate $q \ll 1$. The next example demonstrates the rank behavior in the singular value decomposition (SVD) of a matrix representing the solution vector $\mathbf{u} \in \mathbb{R}^{n_1 \times n_2}$ to equation (5.6) with the 12×12 periodic coefficient shown in Figure 4, left. Figure 7 represents the rank behavior in the SVD decomposition of the solution in the case of the 8×8 periodic coefficient.

Comparing Figures 4 and 7 indicates that the exponential decay of the approximation error in the rank parameter is stable with respect to the size of the $L \times L$ lattice structure of the coefficient, i.e., the behavior of the singular values remains almost the same for different parameters $\epsilon = \frac{1}{L}$.

Our iterative scheme includes only the matrix-vector multiplication with the stiffness matrix A that has the small Kronecker rank $2R$, and the action of the preconditioner defined by the approximate inverse to the Laplacian type matrix. The latter has low Kronecker rank of order $R_B = O(|\log \varepsilon|^2)$ as shown above.

Given rank-1 vector $\mathbf{u} = \mathbf{u}_1 \otimes \mathbf{u}_2$, the standard property of the Kronecker product matrices

$$A\mathbf{u} = A_1\mathbf{u}_1 \otimes M_2\mathbf{u}_2 + M_1\mathbf{u}_1 \otimes A_2\mathbf{u}_2$$

indicates that the matrix-vector multiplication enlarges the initial rank by the factor of 2, and similar with the action of the preconditioner. Hence each iterative step should be supplemented with certain rank truncation procedure which can be implemented adaptively to the chosen approximation threshold or fixed bound on the rank parameter.

Remark 5.1. Notice that for $d = 3$ the transformed matrix $A_\circ$ takes the form

$$A_\circ = A_1 \otimes I_2 \otimes I_3 + I_1 \otimes A_2 \otimes I_3 + I_1 \otimes I_2 \otimes A_3,$$

and it obeys the d-term Kronecker sum representation. Hence in the general case of $d \geq 2$ and $R \geq 1$ the Kronecker rank of the matrix $A_\circ$ is given by

$$\operatorname{rank}_{\mathrm{Kron}}(A_\circ) = dR.$$

6 Conclusions

We present a preconditioned iteration method for solving an elliptic type boundary value problem in $\mathbb{R}^d$ with the operator generated by a quasi-periodic structure with rapidly changing coefficients characterized by a small length parameter ϵ. We use tensor product FEM discretization that allows to approximate the stiffness matrix A in the form of a low-rank Kronecker sum. The preconditioner $A_\circ$ is constructed based on certain averaging (homogenization) procedure of the initial equation coefficients such that the inversion of $A_\circ$ is much simpler than the inversion of A. We prove contraction of the iteration method and establish explicit estimates of the contraction factor $q < 1$. For typical quasi-periodic structures we deduce fully computable two-sided a posteriori estimates which are able to control numerical solutions on any iteration.

We apply the tensor-structured approximation which is especially efficient if the equation coefficients admit low-rank representations and algebraic operations are performed in tensor structured formats. Under moderate assumptions the storage and solution complexity of our approach depends only weakly (merely linear-logarithmically) on the frequency parameter $\frac{1}{\epsilon}$. Numerical tests demonstrate that the FEM solution allows the accurate low-rank separable approximation which is the basic prerequisite for application of the tensor numerical methods to the problems of geometric homogenization.

The approach allows further enhancement based on the quantized-TT (QTT) tensor approximation which is the topic for future research work. Another direction is related to fully tensor structured implementation of the computable two-sided a posteriori error estimates. The interesting question arises how far the presented approach can be extended to the numerical analysis of elliptic equations with rather unstructured jumping coefficients arising in stochastic homogenization, see, e.g., [6].

Acknowledgment: SR appreciates the support provided by the Max-Planck Institute for Mathematics in the Sciences (Leipzig, Germany) during his scientific visit in 2016. The authors are thankful to Dr. V. Khoromskaia (MPI MIS, Leipzig) for the help with numerical experiments.

References

[1] N. S. Bakhvalov and G. Panasenko, *Homogenisation: Averaging Processes in Periodic Media. Mathematical Problems in the Mechanics of Composite Materials*, Springer, Berlin, 1989.

[2] P. Benner, V. Khoromskaia and B. N. Khoromskij, Range-separated tensor formats for numerical modeling of many-particle interaction potentials, preprint (2016), http://arxiv.org/abs/1606.09218.

[3] A. Bensoussan, J.-L. Lions and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, 1978.

[4] S. Dolgov, V. Kazeev and B. N. Khoromskij, The tensor-structured solution of one-dimensional elliptic differential equations with high-dimensional parameters, Preprint 51/2012, MPI MiS, Leipzig, 2012.

[5] I. P. Gavrilyuk, W. Hackbusch and B. N. Khoromskij, Hierarchical tensor-product approximation to the inverse and related operators in high-dimensional elliptic problems, *Computing* **74** (2005), 131–157.

[6] A. Gloria and F. Otto, Quantitative estimates on the periodic approximation of the corrector in stochastic homogenization, *ESAIM Proc.* **48** (2015), 80–97.

[7] R. Glowinski, J.-L. Lions and R. Trémolierés, *Analyse Numérique des Inéquations Variationnelles*, Dunod, Paris, 1976.

[8] V. V. Jikov, S. M. Kozlov and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer, Berlin, 1994.

[9] L. V. Kantorovich and V. L. Krylov, *Approximate Methods of Higher Analysis*, Interscience, New York, 1958.

[10] V. Kazeev, O. Reichmann and C. Schwab, Low-rank tensor structure of linear diffusion operators in the TT and QTT formats, *Linear Algebra Appl.* **438** (2013), no. 11, 4204–4221.

[11] V. Khoromskaia and B. N. Khoromskij, Grid-based lattice summation of electrostatic potentials by assembled rank-structured tensor approximation, *Comp. Phys. Commun.* **185** (2014), no. 12, 3162–3174.

[12] V. Khoromskaia and B. N. Khoromskij, Tensor approach to linearized Hartree–Fock equation for lattice-type and periodic systems, preprint (2014), https://arxiv.org/abs/1408.3839.

[13] V. Khoromskaia and B. N. Khoromskij, Tensor numerical methods in quantum chemistry: From Hartree–Fock to excitation energies, *Phys. Chem. Chem. Phys.* **17** (2015), 31491–31509.

[14] B. N. Khoromskij, Tensor-structured preconditioners and approximate inverse of elliptic operators in $\mathbb{R}^d$, *J. Constr. Approx.* **30** (2009), 599–620.

[15] B. N. Khoromskij, $O(d \log N)$-quantics approximation of N-d tensors in high-dimensional numerical modeling, *Constr. Approx.* **34** (2011), 257–280.

[16] B. N. Khoromskij, Tensors-structured numerical methods in scientific computing: Survey on recent advances, *Chemometr. Intell. Lab. Syst.* **110** (2012), 1–19.

[17] B. N. Khoromskij and S. Repin, A fast iteration method for solving elliptic problems with quasiperiodic coefficients, *Russian J. Numer. Anal. Math. Modelling* **30** (2015), no. 6, 329–344.

[18] B. N. Khoromskij, S. Sauter and A. Veit, Fast quadrature techniques for retarded potentials based on TT/QTT tensor approximation, *Comput. Methods Appl. Math.* **11** (2011), no. 3, 342–362.

[19] B. N. Khoromskij and G. Wittum, *Numerical Solution of Elliptic Differential Equations by Reduction to the Interface*, Lect. Notes Comput. Sci. Eng. 36, Springer, Berlin, 2004.

[20] J.-L. Lions and G. Stampacchia, Variational inequalities, *Comm. Pure Appl. Math.* **20** (1967), 493–519.

[21] O. Mali, P. Neittaanmaki and S. Repin, *Accuracy Verification Methods. Theory and Algorithms*, Springer, New York, 2014.

[22] P. Neittaanmaki and S. Repin, *Reliable Methods for Computer Simulation. Error Control and a Posteriori Estimates*, Elsevier, Amsterdam, 2004.

[23] I. V. Oseledets and S. V. Dolgov, Solution of linear systems and matrix inversion in the TT-format, *SIAM J. Sci. Comput.* **34** (2012), no. 5, A2718–A2739.

[24] A. Ostrowski, Les estimations des erreurs a posteriori dans les procédés itératifs, *C. R. Acad. Sci Paris Sér. A–B* **275** (1972), A275–A278.

[25] S. Repin, A posteriori error estimation for variational problems with uniformly convex functionals, *Math. Comp.* **69** (2000), no. 230, 481–500.

[26] S. Repin, *A Posteriori Estimates for Partial Differential Equations*, Walter de Gruyter, Berlin, 2008.

[27] S. Repin, T. Samrowski and S. Sauter, Combined a posteriori modeling-discretization error estimate for elliptic problems with complicated interfaces, *ESAIM Math. Model. Numer. Anal.* **46** (2012), no. 6, 1389–1405.

[28] S. Repin, S. Sauter and A. Smolianski, A posteriori estimation of dimension reduction errors for elliptic problems on thin domains, *SIAM J. Numer. Anal.* **42** (2004), no. 4, 1435–1451.

[29] E. Zeidler, *Nonlinear Functional Analysis and Its Applications. I: Fixed-Point Theorems*, Springer, New York, 1986.