Received: 23 August 2023 | Revised: 29 April 2024 | Accepted: 1 July 2024

DOI: 10.1002/zamm.202300650

ORIGINAL PAPER

A high-order numerical method for solving non-periodic scattering problems in three-dimensional bi-periodic structures

Tilo Arens[1] | **Nasim Shafieeabyaneh**[1] | **Ruming Zhang**[2]

[1]Institut für Angewandte und Numerische Mathematik, Karlsruher Institut für Technologie, Karlsruhe, Germany

[2]Institut für Mathematik, Technische Universität Berlin, Berlin, Germany

Correspondence
Nasim Shafieeabyaneh, Institut für Angewandte und Numerische Mathematik, Karlsruher Institut für Technologie, Karlsruhe, Germany.
Email: nasim.shafieeabyaneh@kit.edu

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Numbers: 433126998–SFB1173, 258734477

In this paper, we focus on scattering of non-periodic incident fields in three-dimensional bi-periodic structures, as they can not be solved by the classical methods used for the quasi-periodic scattering problems. To solve such non-periodic scattering problems, the Floquet–Bloch transform, which decomposes the unbounded problem into a family of periodic problems in a bounded unit cell, has been applied together with a numerical method by Lechleiter and Zhang (2017). However, its theoretical result indicates that the computational order is too low. Hence, our aim is to propose a high-order numerical approach by using the Floquet–Bloch transform. To this end, the first crucial part is to analyze the regularity of the transformed solution with respect to the Floquet parameter. The second challenging part is to propose a high-order tailor-made quadrature method adapted to singularities of the transformed solution formed by a finite number of circular arcs. Afterwards, we obtain the error estimation of the proposed numerical approach. Eventually, the accuracy and efficiency of the mentioned approach are revealed by several numerical examples.

1 | INTRODUCTION

Scattering problems in periodic structures play a substantial role in modern mathematical physics. They are particularly important in thin solar cell design, photonic crystal band gap engineering, and surface structure optimization for organic light-emitting diodes [1, 2]. The classical periodic scattering problem that a periodic or quasi-periodic incident field such as a plane wave is scattered by a periodic structure, can be directly reduced to a problem posed on a unit cell of the periodic domain [3, 4]. Afterwards, the reduced problem can be solved numerically for example by the finite element [5, 6] or integral equation methods [7]. However, when the incident field is non-periodic, this approach no longer works. Therefore, novel numerical schemes are necessary to efficiently solve these challenging problems.

One way to tackle such problems is the use of the Floquet–Bloch transform. It has been applied most often in two-dimensional scattering problems (e.g., see numerical results in refs. [8–11] and theoretical results in refs. [12, 13]). This is also the approach that will be used in this paper, where we consider a three-dimensional geometry, extending similar approaches from refs. [14–16]. Another possibility is a numerical approach based on the extension of the Robin-to-Robin map by using a recursive doubling procedure, as described in refs. [17, 18]. Moreover, in refs. [19–21], operator equations are solved to construct the Dirichlet-to-Neumann (DtN) map. It is worth noting that the non-periodic scattering problems in

three dimension have been studied by the Floquet–Bloch transform only in refs. [14, 16]. However, the convergence rate of the techniques in these references is rather low.

In this paper, we propose a high-order numerical scheme to solve scattering problems in three-dimensional bi-periodic structures. We first apply the Floquet–Bloch transform to decompose the original scattering problem, involving non-periodic fields and posed in an unbounded but periodic domain, into a family of problems involving only periodic fields and posed on just a single bounded cell of periodicity. In this procedure, the numerical error is the combination of two components: the error in the approximation of the transformed total field due to the employed numerical method and the error due to the approximation of the inverse Floquet–Bloch transform. To approximate the transformed field, we use the finite element method whose error estimation is given by classical results. Hence, the main goal of this paper is to derive a highly accurate and efficient scheme for the inversion of the Floquet–Bloch transform. This, in particular, requires to prove regularity properties of the transformed field with respect to the Floquet parameter: the inverse Floquet–Bloch transform essentially consists of a double integral of the transformed field over a bounded domain, but the integrand has got a particular structure of singularities. Based on the regularity results we establish, we propose a tailor-made quadrature rule to numerically obtain the total field of the original non-periodic scattering problem.

It should be pointed out that the regularity of the transformed field for the two-dimensional scattering problem in ref. [22] is not similar to the three-dimensional case. In ref. [22], it is proved that the transformed field is analytic except for at most two singular points. However, in the three-dimensional case, the singularities of the transformed field no longer consist of a finite number of points; they form a set that is the union of a finite number of circular arcs. Hence, the extension of the high-order numerical methods used for the two-dimensional case in refs. [22, 23] is not appropriate for the three-dimensional case.

The framework of this paper is as follows: Section 2 is devoted to a review of the mathematical formulation of scattering problems in unbounded bi-periodic domains. In Section 3, we introduce the Floquet–Bloch transform and state some of its properties that we require in the later analysis. Furthermore, we apply the Floquet–Bloch transform to the variational formulation of the original scattering problem to derive a family of periodic problems that may be reduced to just a single bounded cell of periodicity. Afterwards, we analyze the regularity of the transformed field with respect to the Floquet parameter. Our first main result in Theorem 5 is a local representation of the transformed field exactly mirroring the expected structure of singularities. This significantly extends similar representations found in refs. [14, 15]. Moreover, we obtain a globally valid representation in Theorem 7. In Section 4, we construct a quadrature rule exactly adapted to the singularity structure of the transformed field. This allows a rigorous analysis of the quadrature error bases on the regularity results established earlier in Corollary 17 and of the overall numerical method in Theorem 20. Some numerical examples illustrating the performance of the proposed scheme are presented in Section 5.

2 | MATHEMATICAL MODEL OF SCATTERING PROBLEMS

We consider acoustic wave propagation in an unbounded domain Ω which is bounded from below by a bi-periodic surface Γ given as the graph of a bounded function ξ, that is,

$$\Omega = \left\{(\widetilde{\boldsymbol{x}}, x_3) : \widetilde{\boldsymbol{x}} \in \mathbb{R}^2,\, x_3 > \xi(\widetilde{\boldsymbol{x}})\right\}, \qquad \Gamma = \left\{(\widetilde{\boldsymbol{x}}, \xi(\widetilde{\boldsymbol{x}})) : \widetilde{\boldsymbol{x}} \in \mathbb{R}^2\right\}.$$

The function ξ is assumed to be 2π-periodic with respect to both variables. A given, non-periodic incident field u^i propagating in Ω is scattered by Γ and generates a scattered field u^s that is to be determined (see Figure 1a). The total field $u = u^i + u^s$ satisfies the Helmholtz equation with the wave number $\kappa > 0$

$$\Delta u + \kappa^2 u = 0 \qquad \text{in } \Omega \subset \mathbb{R}^3, \tag{1}$$

and it is assumed to satisfy a Dirichlet boundary condition

$$u = 0 \qquad \text{on } \Gamma\,. \tag{2}$$

The formulation of the scattering problem is not complete without an appropriate radiation condition. This condition physically guarantees that the scattered field u^s is propagating upwards from Γ and, mathematically, it makes the scattering problem well-posed. Before stating this condition, we need to introduce some definitions.

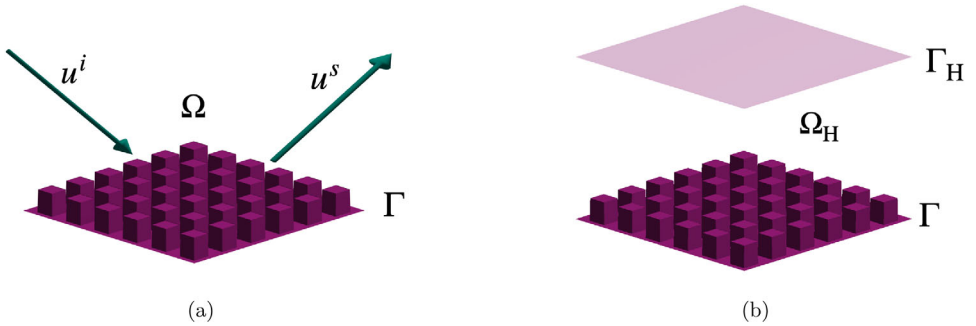


FIGURE 1 A sketch of the presented unbounded domains.

We assume $H > \|\xi\|_\infty$ and let $\Gamma_H := \mathbb{R}^2 \times \{H\}$. By Ω_H, we denote the unbounded domain between Γ and Γ_H. Figure 1b shows a sketch of these domains.

We will use the standard Sobolev spaces $H^s(\Omega_H)$ and $H^s_{loc}(\Omega_H)$ for any $s \in \mathbb{R}$ defined in ref. [24]. For $s, r \in \mathbb{R}$, the weighted Sobolev spaces $H^s_r(\Omega_H)$ and $H^s_r(\Gamma_H)$ are defined by

$$H^s_r(\Omega_H) = \left\{\phi \in H^s_{loc}(\Omega_H) : \quad (1+|\tilde{x}|^2)^{r/2}\phi \in H^s(\Omega_H)\right\},$$
$$H^s_r(\Gamma_H) = \left\{\phi \in H^s_{loc}(\Gamma_H) : \quad (1+|\tilde{x}|^2)^{r/2}\phi \in H^s(\Gamma_H)\right\}.$$

The corresponding spaces of functions satisfying a homogeneous Dirichlet boundary condition on Γ in the trace sense will be denoted with a tilde, that is,

$$\widetilde{H}^s_r(\Omega_H) = \left\{\phi \in H^s_r(\Omega_H) : \quad \phi|_\Gamma = 0\right\}.$$

We will assume that the incident field satisfies $u^i \in H^1_r(\Omega_H)$ and look for the total field in $\widetilde{H}^s_r(\Omega_H)$. In addition, to make the scattering problem physically meaningful, the scattered field u^s will be assumed to satisfy the radiation condition [24, 25],

$$u^s(\tilde{x}, x_3) = \frac{1}{2\pi}\int_{\mathbb{R}^2} e^{i\tilde{x}\cdot\zeta + i\sqrt{\kappa^2-|\zeta|^2}(x_3-H)}\hat{u}^s(\zeta, H)d\zeta, \quad x_3 > H, \tag{3}$$

where $\hat{u}^s(\zeta, H)$ denotes the Fourier transform of u^s restricted to Γ_H, that is,

$$\hat{u}^s(\zeta, H) := \frac{1}{2\pi}\int_{\mathbb{R}^2} e^{-i\tilde{x}\cdot\zeta} u^s(\tilde{x}, H)\, d\tilde{x}.$$

In ref. [24], it is shown that $\hat{u}^s(\cdot, H) \in H^{1/2}_r(\mathbb{R}^2)$ and it is elaborated why the integral on the right-hand side of (3) exists for $u^s \in H^1_r(\Omega_H)$, $|r| < 1$.

The radiation condition (3) can be equivalently formulated as the transparent boundary condition

$$\frac{\partial u}{\partial x_3}(\tilde{x}, H) - \left(T^+ u|_{\Gamma_H}\right) = \frac{\partial u^i}{\partial x_3}(\tilde{x}, H) - \left(T^+ u^i|_{\Gamma_H}\right) =: f, \qquad \text{on } \Gamma_H, \tag{4}$$

where the DtN map T^+ is defined by

$$\left(T^+\psi|_{\Gamma_H}\right)(\tilde{x}, H) = \frac{i}{2\pi}\int_{\mathbb{R}^2}\sqrt{\kappa^2-|\zeta|^2}e^{i\tilde{x}\cdot\zeta}\hat{\psi}(\zeta, H)d\zeta.$$

In refs. [24, 25], it is proved that $T^+ : H^{1/2}_r(\Gamma_H) \to H^{-1/2}_r(\Gamma_H)$ is well-defined and continuous for $|r| < 1$.

Now, the problem (1)–(2) can be reduced to a boundary value problem in Ω_H together with the transparent boundary condition (4) on Γ_H. The variational formulation of this boundary value problem with $f \in H^{-1/2}_r(\Gamma_H)$ for $|r| < 1$ is to find

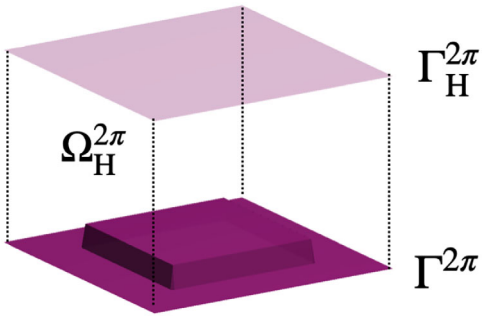


FIGURE 2 The three-dimensional bounded unit cell $\Omega_{\mathrm{H}}^{2\pi}$.

$u \in \widetilde{H}_r^1(\Omega_{\mathrm{H}})$ such that it satisfies

$$\int_{\Omega_{\mathrm{H}}} \left(\nabla u \cdot \nabla \overline{v} - \kappa^2 u \overline{v}\right) \mathrm{d}\boldsymbol{x} - \int_{\Gamma_{\mathrm{H}}} \left(T^+ u|_{\Gamma_{\mathrm{H}}}\right)\overline{v}\, \mathrm{d}s = \int_{\Gamma_{\mathrm{H}}} f\, \overline{v}\, \mathrm{d}s, \quad \text{for all } v \in \widetilde{H}^1(\Omega_{\mathrm{H}}). \tag{5}$$

Existence and uniqueness of solution for the variational problem (5) has been proved in ref. [24].

From a numerical point of view, the variational problem (5) is not yet adequate as it is still posed on an unbounded domain. In the next section, by applying the Floquet–Bloch transform, we hence present a decomposed formulation of (5) consisting of a family of periodic problems posed on a single bounded unit cell of the periodic domain.

3 | THE FLOQUET–BLOCH TRANSFORM

Consider the square lattice $\{2\pi \boldsymbol{j} : \boldsymbol{j} \in \mathbb{Z}^2\}$ with the primitive cell $\mathrm{V} = \{2\pi\boldsymbol{\eta} : \boldsymbol{\eta} \in \mathbb{R}^2, -1/2 < \boldsymbol{\eta}_{1,2} \le 1/2\}$. We define the three-dimensional bounded unit cell by restricting the domain Ω_{H} to $\widetilde{\boldsymbol{x}}$ in the considered primitive cell, that is, $\Omega_{\mathrm{H}}^{2\pi} = \{\boldsymbol{x} \in \Omega_{\mathrm{H}} : \widetilde{\boldsymbol{x}} \in \mathrm{V}\}$ as depicted in Figure 2.

Definition 1 [12, 16]. For $\varphi \in C_0^\infty(\Omega_{\mathrm{H}})$, the Floquet–Bloch transform $\mathcal{J}\varphi$ is defined by

$$(\mathcal{J}\varphi)(\boldsymbol{\alpha}, \boldsymbol{x}) = \sum_{\boldsymbol{j} \in \mathbb{Z}^2} \varphi(\widetilde{\boldsymbol{x}} + 2\pi\boldsymbol{j}, x_3) \mathrm{e}^{-\mathrm{i}\boldsymbol{\alpha}\cdot(\widetilde{\boldsymbol{x}} + 2\pi\boldsymbol{j})}, \tag{6}$$

where $\boldsymbol{x} = (x_1, x_2, x_3)^\top$, $\widetilde{\boldsymbol{x}} = (x_1, x_2)^\top \in \mathbb{R}^2$ and $\boldsymbol{\alpha} \in \mathbb{R}^2$ denotes the Floquet parameter.

Note that $\mathcal{J}\varphi$ is bi-periodic with respect to $\widetilde{\boldsymbol{x}}$ with period 2π in each coordinate direction. Moreover, for every $\boldsymbol{x}$ the function $\mathrm{e}^{\mathrm{i}\boldsymbol{\alpha}\cdot\widetilde{\boldsymbol{x}}}(\mathcal{J}\varphi)(\boldsymbol{\alpha}, \boldsymbol{x})$ is bi-periodic with respect to $\boldsymbol{\alpha}$ with period 1 in each coordinate direction. The fundamental cell of periodicity of $\mathcal{J}\varphi$ thus is $\mathrm{V}^* \times \Omega_{\mathrm{H}}^{2\pi}$ where $\mathrm{V}^* := [-1/2, 1/2]^2$. To investigate more properties of the Floquet–Bloch transform, we introduce the Sobolev space of bi-periodic functions

$$\widetilde{H}^s_{\mathrm{per}}(\Omega_{\mathrm{H}}^{2\pi}) = \left\{\phi \in H^s(\Omega_{\mathrm{H}}^{2\pi}) : \phi(\widetilde{\boldsymbol{x}} + 2\pi\boldsymbol{j}, x_3) = \phi(\boldsymbol{x}),\ \boldsymbol{j} \in \mathbb{Z}^2,\ \phi|_{\Gamma^{2\pi}} = 0\right\},$$

and the space $H^r(\mathrm{V}^*; \widetilde{H}^s_{\mathrm{per}}(\Omega_{\mathrm{H}}^{2\pi}))$ with norm

$$\|\phi\|_{H^r(\mathrm{V}^*; \widetilde{H}^s_{\mathrm{per}}(\Omega_{\mathrm{H}}^{2\pi}))} = \left(\sum_{m \in \mathbb{N}^2 : |m| \le r} \int_{\mathrm{V}^*} \|\partial_{\boldsymbol{\alpha}}^m \phi(\boldsymbol{\alpha})\|^2_{\widetilde{H}^s_{\mathrm{per}}(\Omega_{\mathrm{H}}^{2\pi})}\, \mathrm{d}\boldsymbol{\alpha}\right)^{1/2}.$$

For $\psi \in H^r(\mathrm{V}^*; \widetilde{H}^s_{\mathrm{per}}(\Omega_{\mathrm{H}}^{2\pi}))$ we will write $\psi(\boldsymbol{\alpha}) \in \widetilde{H}^s_{\mathrm{per}}(\Omega_{\mathrm{H}}^{2\pi})$, but continue using $\psi(\boldsymbol{\alpha}, \boldsymbol{x})$ instead of $\psi(\boldsymbol{\alpha})(\boldsymbol{x})$. The next theorem states the mapping properties of the Floquet–Bloch transform in the framework of these spaces.

Theorem 2 Theorem 1, [16]. *The Floquet–Bloch transform $\mathcal{J}$ extends to an isomorphism between $\widetilde{H}^s_r(\Omega_{\mathrm{H}})$ and $H^r(\mathrm{V}^*;\widetilde{H}^s_{\mathrm{per}}(\Omega^{2\pi}_{\mathrm{H}}))$ for all $s,r\in\mathbb{R}$. Moreover, the inverse Floquet–Bloch transform is obtained by*

$$(\mathcal{J}^{-1}\psi)(\tilde{\boldsymbol{x}}+2\pi\boldsymbol{j},x_3)=\int_{\mathrm{V}^*}\psi(\boldsymbol{\alpha},\boldsymbol{x})\mathrm{e}^{\mathrm{i}\boldsymbol{\alpha}\cdot(\tilde{\boldsymbol{x}}+2\pi\boldsymbol{j})}\mathrm{d}\boldsymbol{\alpha},\boldsymbol{x}\in\Omega^{2\pi}_{\mathrm{H}},\boldsymbol{j}\in\mathbb{Z}^2. \tag{7}$$

According to ref. [12], the mapping properties of the Floquet–Bloch transform when operating on functions defined on Γ_{H} or Γ are analogous.

We now use the Floquet–Bloch transform to decompose the scattering problem in the unbounded domain Ω_{H} to a family of periodic problems in the unit cell $\Omega^{2\pi}_{\mathrm{H}}$. Let the Floquet–Bloch transform of the total field u be denoted by $w:=\mathcal{J}u$. By applying the Floquet–Bloch transform to the Helmholtz equation in Ω_{H} and to the boundary conditions (2) and (4), it turns out that for every $\boldsymbol{\alpha}\in\mathrm{V}^*$, $w(\boldsymbol{\alpha})\in\widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_{\mathrm{H}})$ is a weak solution to the problem

$$\begin{cases}\Delta_{\boldsymbol{x}}w(\boldsymbol{\alpha})+2\mathrm{i}\,\boldsymbol{\alpha}\cdot\widetilde{\nabla}_{\boldsymbol{x}}w(\boldsymbol{\alpha})+\left(\kappa^2-|\boldsymbol{\alpha}|^2\right)w(\boldsymbol{\alpha})=0 & \text{in } \Omega^{2\pi}_{\mathrm{H}}, \quad (8)\\ w(\boldsymbol{\alpha})=0 & \text{on } \Gamma^{2\pi}, \quad (9)\\ \dfrac{\partial w(\boldsymbol{\alpha})}{\partial x_3}-T'_{\alpha}w(\boldsymbol{\alpha})=\mathcal{F}(\boldsymbol{\alpha}) & \text{on } \Gamma^{2\pi}_{\mathrm{H}}, \quad (10)\end{cases}$$

where $\mathcal{F}$ denotes the Floquet–Bloch transform of f defined in (4). This means (8) is understood in the distributional sense and (9), (10) in the trace sense and $\widetilde{\nabla}_{\boldsymbol{x}}w(\boldsymbol{\alpha})=(\partial w(\boldsymbol{\alpha})/\partial x_1,\partial w(\boldsymbol{\alpha})/\partial x_2)^{\top}$. The periodic DtN map $T'_{\alpha}:H^{1/2}_{\mathrm{per}}(\Gamma^{2\pi}_{\mathrm{H}})\to H^{-1/2}_{\mathrm{per}}(\Gamma^{2\pi}_{\mathrm{H}})$ is defined by

$$(T'_{\alpha}\Psi)(\tilde{\boldsymbol{x}})=\mathrm{i}\sum_{\boldsymbol{j}\in\mathbb{Z}^2}\sqrt{\kappa^2-|\boldsymbol{\alpha}-\boldsymbol{j}|^2}\hat{\Psi}(\boldsymbol{j})\mathrm{e}^{\mathrm{i}\tilde{\boldsymbol{x}}\cdot\boldsymbol{j}}\text{for}\Psi=\sum_{\boldsymbol{j}\in\mathbb{Z}^2}\hat{\Psi}(\boldsymbol{j})\mathrm{e}^{\mathrm{i}\tilde{\boldsymbol{x}}\cdot\boldsymbol{j}}.$$

Theorem 3 Theorem 2, [16]. *Let $|r|<1$ and $u^i\in H^1_r(\Omega_{\mathrm{H}})$. A function $u\in\widetilde{H}^1_r(\Omega_{\mathrm{H}})$ satisfies (5) if and only if $w\in H^r(\mathrm{V}^*;\widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_{\mathrm{H}}))$ is a solution to the variational problem*

$$\int_{\mathrm{V}^*}a_{\alpha}(w(\boldsymbol{\alpha}),\psi(\boldsymbol{\alpha}))\,\mathrm{d}\boldsymbol{\alpha}-\sum_{\boldsymbol{j}\in\mathbb{Z}^2}\int_{\mathrm{V}^*}\sqrt{\kappa^2-|\boldsymbol{\alpha}-\boldsymbol{j}|^2}\;b_{\boldsymbol{j}}(w(\boldsymbol{\alpha}),\psi(\boldsymbol{\alpha}))\,\mathrm{d}\boldsymbol{\alpha}$$
$$=\int_{\mathrm{V}^*}\int_{\Gamma^{2\pi}_{\mathrm{H}}}\mathcal{F}(\boldsymbol{\alpha},\boldsymbol{x})\,\overline{\psi(\boldsymbol{\alpha},\boldsymbol{x})}\,\mathrm{d}s\,\mathrm{d}\boldsymbol{\alpha}\qquad\text{for all }\psi\in H^{-r}(\mathrm{V}^*;\widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_{\mathrm{H}})). \tag{11}$$

Moreover, if $\boldsymbol{\alpha}\mapsto\mathcal{F}(\boldsymbol{\alpha})$ is continuous, then $\boldsymbol{\alpha}\mapsto w(\boldsymbol{\alpha})$ is also continuous, and for every $\boldsymbol{\alpha}\in\mathrm{V}^*$,

$$a_{\alpha}(w(\boldsymbol{\alpha}),\zeta)-\sum_{\boldsymbol{j}\in\mathbb{Z}^2}\sqrt{\kappa^2-|\boldsymbol{\alpha}-\boldsymbol{j}|^2}\,b_{\boldsymbol{j}}(w(\boldsymbol{\alpha}),\zeta)=\int_{\Gamma^{2\pi}_{\mathrm{H}}}\mathcal{F}(\boldsymbol{\alpha})\,\overline{\zeta}\,\mathrm{d}s\,, \tag{12}$$

for all $\zeta\in\widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_{\mathrm{H}})$. Here,

$$a_{\alpha}(v,\zeta)=\int_{\Omega^{2\pi}_{\mathrm{H}}}\left(\nabla v\cdot\nabla\overline{\zeta}-2\mathrm{i}\,\boldsymbol{\alpha}\cdot\widetilde{\nabla}v\,\overline{\zeta}-(\kappa^2-|\boldsymbol{\alpha}|^2)\,v\,\overline{\zeta}\right)\mathrm{d}\boldsymbol{x}\,,$$
$$b_{\boldsymbol{j}}(v,\zeta)=\mathrm{i}\int_{\Gamma^{2\pi}_{\mathrm{H}}}\overline{\zeta(\boldsymbol{x})}\,\hat{v}(\boldsymbol{j})\,\mathrm{e}^{\mathrm{i}\tilde{\boldsymbol{x}}\cdot\boldsymbol{j}}\,\mathrm{d}s\;=4\mathrm{i}\pi^2\,\hat{v}(\boldsymbol{j})\,\overline{\hat{\zeta}(\boldsymbol{j})}\,.$$

Unique solvability of the variational problem (12) in $\widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_{\mathrm{H}})$ has been proved in ref. [26] for any arbitrary, but fixed $\boldsymbol{\alpha}\in\mathrm{V}^*$. We can thus compute numerical approximations to the transformed field $w(\boldsymbol{\alpha})$ for every $\boldsymbol{\alpha}\in\mathrm{V}^*$ by using some

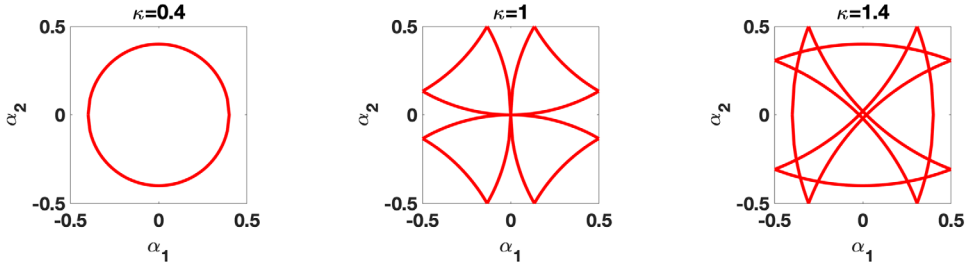


FIGURE 3 Structure of $\mathcal{S}$ for different values of κ on $\mathrm{V}^* = [-0.5, 0.5]^2$.

numerical method of choice. In the second step, the inverse Floquet–Bloch transform (7) must be computed to obtain an approximation to the solution of (11). This essentially amounts to the evaluation of a double integral of w over the domain V^*. The accuracy of the numerical solution of (11) depends not only on the selected numerical method for solving (12), but also on the accuracy of the numerical integration method employed for this double integral. In order to construct a high-order numerical scheme, requiring few quadrature points for high accuracy, it is necessary to precisely know the regularity of the transformed field with respect to the Floquet parameter.

Let us heuristically motivate the results that we shall make rigorous in Theorem 5. In the variational formulation (12), all terms depend analytically on $\boldsymbol{\alpha}$ except for the square root functions. Hence, we may expect the transformed field w to depend analytically on $\boldsymbol{\alpha}$, except for points where (the derivatives of) these functions have singularities, that is, except for points located in the set

$$\mathcal{S} = \left\{ \boldsymbol{\alpha} \in \mathrm{V}^* \ : \ |\boldsymbol{\alpha} - \boldsymbol{j}| = \kappa \text{ for some } \boldsymbol{j} \in \mathbb{Z}^2 \right\}.$$

The set $\mathcal{S}$ is a union of circular arcs formed by the intersection of V^* and circles with center $\boldsymbol{j}$ and radius κ, and we will also refer to this set as the *curves of singular points*. Figure 3 illustrates possible structures of $\mathcal{S}$ for different wave numbers κ on V^*. Any high-order method for approximately inverting the Floquet–Bloch transform will need to take into account the structure of $\mathcal{S}$ that becomes more and more complex as κ increases.

For any $\boldsymbol{\alpha} \in \mathcal{S}$, we also define

$$\mathbf{J}(\boldsymbol{\alpha}) = \{ \boldsymbol{j} \in \mathbb{Z}^2 \ : \ |\boldsymbol{\alpha} - \boldsymbol{j}| = \kappa \}, \tag{13}$$

a finite set with cardinality $\#\mathbf{J}(\boldsymbol{\alpha})$.

Remark 4. When $\kappa < 0.5$, for all $\boldsymbol{\alpha} \in \mathcal{S}$, $\#\mathbf{J}(\boldsymbol{\alpha}) = 1$. When $\kappa \geq 0.5$, there exist finite number of $\boldsymbol{\alpha} \in \mathcal{S}$ with $\#\mathbf{J}(\boldsymbol{\alpha}) > 1$.

For the later analysis of the numerical inversion of the Floquet–Bloch transform, we require a particular regularity of both the transformed incident and the transformed total fields. To formulate these requirements, we make the following definitions: For some open set $U \subseteq \mathbb{R}^2$ and Hilbert space Y, we denote by $C^\omega(U; Y)$ the space of functions from U to Y that depend analytically on $\boldsymbol{\alpha} \in U$. For a Hilbert space Y, let

$$\mathcal{X}(Y) = \{ g : \mathrm{V}^* \to Y \ : \ g \text{ satisfies (C1) and (C2)} \}, \tag{14}$$

where

(C1) for every open subdomain $U \subseteq \mathrm{V}^* \setminus \mathcal{S}$, $g \in C^\omega(U; Y)$,
(C2) for any $\boldsymbol{\alpha}_0 \in \mathcal{S}$, there exists a neighborhood U_0 of $\boldsymbol{\alpha}_0$ such that

$$g(\boldsymbol{\alpha}) = \sum_{\mathcal{I} \subseteq \mathbf{J}(\boldsymbol{\alpha}_0)} \prod_{\boldsymbol{j} \in \mathcal{I}} \sqrt{\kappa^2 - |\boldsymbol{\alpha} - \boldsymbol{j}|^2}\, g_{\mathcal{I}}(\boldsymbol{\alpha}), \tag{15}$$

where $g_{\mathcal{I}} \in C^\omega(U_0; Y)$ for every $\mathcal{I} \subseteq \mathbf{J}(\boldsymbol{\alpha}_0)$.

Theorem 5. *Let $u^i \in H^1_r(\Omega_H)$ for some $|r| < 1$ and additionally $\mathcal{F} \in \mathcal{X}(H^{-1/2}_{per}(\Gamma_H^{2\pi}))$. Then, the transformed total field w that solves (11) satisfies $w \in \mathcal{X}(\widetilde{H}^1_{per}(\Omega_H^{2\pi}))$.*

Proof. Let $\boldsymbol{\alpha}_0 \in \mathrm{V}^*$. Using the Riesz representation theorem, we may define the operators $\mathcal{A}(\boldsymbol{\alpha})$ and $\mathcal{B}(\boldsymbol{j})$ by

$$\langle \mathcal{B}(\boldsymbol{j})v, \zeta\rangle_{\Gamma_H^{2\pi}} = b_j(v,\zeta)\,,$$

$$\langle \mathcal{A}(\boldsymbol{\alpha})v, \zeta\rangle_{\Omega_H^{2\pi}} = a_{\boldsymbol{\alpha}}(v,\zeta) \; - \sum_{\boldsymbol{j}\in\mathbb{Z}^2\setminus\mathbf{J}(\boldsymbol{\alpha}_0)} \sqrt{\kappa^2 - |\boldsymbol{\alpha}-\boldsymbol{j}|^2}\langle \mathcal{B}(\boldsymbol{j})v, \zeta\rangle_{\Gamma_H^{2\pi}}.$$

Note that in a neighborhood of $\boldsymbol{\alpha}_0$, $\mathcal{A}(\boldsymbol{\alpha})$ depends analytically on $\boldsymbol{\alpha}$. Using these operators, and also the antilinear form $\widehat{\mathcal{F}}(\boldsymbol{\alpha})$ induced by the right-hand side of (12), (12) can be reformulated as

$$\left[\mathcal{A}(\boldsymbol{\alpha}) \; - \sum_{\boldsymbol{j}\in\mathbf{J}(\boldsymbol{\alpha}_0)} \sqrt{\kappa^2 - |\boldsymbol{\alpha}-\boldsymbol{j}|^2}\,\mathcal{B}(\boldsymbol{j})\right] w(\boldsymbol{\alpha}) = \widehat{\mathcal{F}}(\boldsymbol{\alpha})\,. \tag{16}$$

If $\boldsymbol{\alpha}_0 \notin \mathcal{S}$, then $\mathbf{J}(\boldsymbol{\alpha}_0) = \emptyset$ and as $\mathcal{F}$ satisfies (C1) with $Y = H^{-1/2}_{per}(\Gamma_H^{2\pi})$, so does w with $Y = H^1_{per}(\Omega_H^{2\pi})$.

We now assume $\boldsymbol{\alpha}_0 \in \mathcal{S}$. Moreover, let $B(\boldsymbol{\alpha}_0, \delta)$ denote an open ball centred at $\boldsymbol{\alpha}_0$ with radius δ. Note that for any $\boldsymbol{j} \in \mathbf{J}(\boldsymbol{\alpha}_0)$, $\|\sqrt{\kappa^2 - |\boldsymbol{\alpha}-\boldsymbol{j}|^2}\,\mathcal{B}(\boldsymbol{j})\| \to 0$ as $|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0| \to 0$. In ref. [1], it has been shown that the operator on the left-hand side of (16) is boundedly invertible. Hence, for small enough δ, the operator $\mathcal{A}(\boldsymbol{\alpha})$ is boundedly invertible for all $\boldsymbol{\alpha} \in B(\boldsymbol{\alpha}_0, \delta)$. Setting $\widetilde{\mathcal{B}}(\boldsymbol{j}) = (\mathcal{A}(\boldsymbol{\alpha}))^{-1}\,\mathcal{B}(\boldsymbol{j})$, we can write the solution w as the Neumann series

$$w(\boldsymbol{\alpha}) = \sum_{n=0}^{\infty}\left(\sum_{\boldsymbol{j}\in\mathbf{J}(\boldsymbol{\alpha}_0)} \sqrt{\kappa^2 - |\boldsymbol{\alpha}-\boldsymbol{j}|^2}\;\widetilde{\mathcal{B}}(\boldsymbol{j})\right)^n (\mathcal{A}(\boldsymbol{\alpha}))^{-1}\widehat{\mathcal{F}}\,.$$

Let $\mathbf{J}(\boldsymbol{\alpha}_0) = \{\boldsymbol{j}_1, \dots, \boldsymbol{j}_m\}$. Applying the multinomial theorem leads to

$$w(\boldsymbol{\alpha}) = \sum_{n=0}^{\infty}\left(\sum_{\substack{\mathrm{K}_1+\mathrm{K}_2+\cdots+\mathrm{K}_m=n,\\ \mathrm{K}_1,\dots,\mathrm{K}_m\geq 0}} \frac{n}{\mathrm{K}_1!\mathrm{K}_2!\dots\,\mathrm{K}_m!}\prod_{\mu=1}^{m}\left(\sqrt{\kappa^2 - |\boldsymbol{\alpha}-\boldsymbol{j}_\mu|^2}\,\widetilde{\mathcal{B}}(\boldsymbol{j}_\mu)\right)^{\mathrm{K}_\mu}\right)(\mathcal{A}(\boldsymbol{\alpha}))^{-1}\widehat{\mathcal{F}}\,.$$

Note that all even powers of the square root functions are analytic. Insertion (15) for $\widehat{\mathcal{F}}$ and combining all analytic terms appropriately into functions $w_{\mathcal{I}}$ for $\mathcal{I} \subseteq \mathbf{J}(\boldsymbol{\alpha}_0)$, gives that w satisfies (C2) with $Y = H^1_{per}(\Omega_H^{2\pi})$. □

Following up on the previous result, the next theorem guarantees that we can make use of (15) for w with the same center of expansion in small balls contained in a neighborhood of $\mathcal{S}$.

Theorem 6. *There exist open balls $B_\ell = B(\boldsymbol{\alpha}_\ell, \rho_\ell)$ with center points $\boldsymbol{\alpha}_\ell \in \mathcal{S}$ and radii ρ_ℓ, $\ell = 1, \dots, L$, such that $\mathcal{S} \subseteq \bigcup_{\ell=1}^{L} B_\ell$ and the representation (15) holds for w on B_ℓ with $\boldsymbol{\alpha}_0 = \boldsymbol{\alpha}_\ell$. Moreover, there exist $r, \delta > 0$ such that*

$$\widetilde{\mathcal{S}} := \{\boldsymbol{\alpha}' \in \mathrm{V}^* \,:\, \mathrm{dist}(\boldsymbol{\alpha}', \mathcal{S}) < r\} \subseteq \bigcup_{\ell=1}^{L} B_\ell\,,$$

and that for every $\boldsymbol{\alpha} \in \widetilde{\mathcal{S}}$ there exists ℓ with $B(\boldsymbol{\alpha}, \delta) \subseteq B_\ell$.

Proof. For every $\boldsymbol{\alpha}_0 \in \mathcal{S}$, we choose $\rho(\boldsymbol{\alpha}_0) > 0$ such that the representation (15) holds for w on $B(\boldsymbol{\alpha}_0, \rho(\boldsymbol{\alpha}_0))$. Then, $\mathcal{S} \subseteq \bigcup_{\boldsymbol{\alpha}_0\in\mathcal{S}} B(\boldsymbol{\alpha}_0, \rho(\boldsymbol{\alpha}_0))$. Since $\mathcal{S}$ is a compact set, we select a finite number of points $\boldsymbol{\alpha}_\ell$ and radii $\rho_\ell = \rho(\boldsymbol{\alpha}_\ell)$, $\ell = 1, \dots, L$, such that $\mathcal{S} \subseteq \bigcup_{\ell=1}^{L} B(\boldsymbol{\alpha}_\ell, \rho_\ell)$. This yields the first part of the theorem.

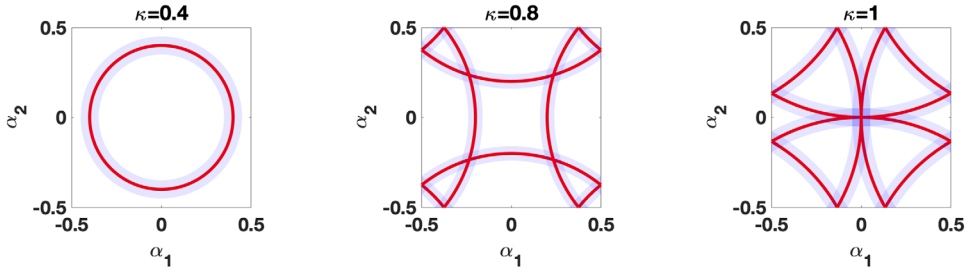


FIGURE 4 Structure of $\widetilde{S}$ for different values of κ on $\mathrm{V}^* = [-0.5, 0.5]^2$.

Choose $q \in (0,1)$ such that still $S \subseteq \bigcup_{\ell=1}^{L} B(\boldsymbol{\alpha}_\ell, q\rho_\ell)$. Choose r such that $\widetilde{S} \subseteq \bigcup_{\ell=1}^{L} B(\boldsymbol{\alpha}_\ell, q\rho_\ell)$ and set $\delta = (1-q) \min_{\ell=1,\dots,L} \rho_\ell$. Now, let $\boldsymbol{\alpha} \in \widetilde{S}$ and $\hat{\ell}$ such that $|\boldsymbol{\alpha} - \boldsymbol{\alpha}_{\hat{\ell}}| < q\,\rho_{\hat{\ell}}$. Then, for any $\boldsymbol{\alpha}' \in B(\boldsymbol{\alpha}, \delta)$, we have

$$|\boldsymbol{\alpha}' - \boldsymbol{\alpha}_{\hat{\ell}}| < q\,\rho_{\hat{\ell}} + \delta = q\,\rho_{\hat{\ell}} + (1-q) \min_{\ell=1,\dots,L} \rho_\ell \le \rho_{\hat{\ell}}\,.$$

This completes the proof. □

The structure of $\widetilde{S}$ for different values of the wave number κ is depicted in Figure 4. For any point $\boldsymbol{\alpha}$ in $\widetilde{S}$, we may use the local representation (15) for the transformed field also on a small neighborhood of that point. In our later analysis, we also require a globally valid representation of w which is provided by the next theorem.

Theorem 7. *Let $\boldsymbol{\alpha}_\ell$, $\ell = 1, \dots, L$, denote the points in Theorem 6 and set $\mathbf{J} = \bigcup_{\ell=1}^{L} \mathbf{J}(\boldsymbol{\alpha}_\ell)$. Then there exist $v_{\mathcal{I}} \in C^\infty(\mathrm{V}^*; \widetilde{H}^1_{\mathrm{per}}(\Omega_{\mathrm{H}}^{2\pi}))$ such that*

$$w(\boldsymbol{\alpha}) = \sum_{\mathcal{I} \subset \mathbf{J}} \prod_{\boldsymbol{j} \in \mathcal{I}} \sqrt{\kappa^2 - |\boldsymbol{\alpha} - \boldsymbol{j}|^2}\; v_{\mathcal{I}}(\boldsymbol{\alpha})\,, \qquad \boldsymbol{\alpha} \in \mathrm{V}^*\,. \tag{17}$$

Moreover, for any $\mu \in \mathbb{N}_0$, there exist a constant C_μ such that

$$\left\| \frac{\partial^\mu v_{\mathcal{I}}(\boldsymbol{\alpha})}{\partial \alpha_\nu^\mu} \right\|_{\widetilde{H}^1_{\mathrm{per}}(\Omega_{\mathrm{H}}^{2\pi})} \le \frac{C_\mu}{\mathrm{dist}(\boldsymbol{\alpha}, \widetilde{S})^\mu}\,, \qquad \mathcal{I} \subseteq \mathbf{J}\,, \quad \nu = 1, 2\,, \quad \boldsymbol{\alpha} \in \mathrm{V}^*\,. \tag{18}$$

Proof. Recall the covering of S by the open balls $B(\boldsymbol{\alpha}_\ell, \delta_\ell)$, $\ell = 1, \dots, L$, from the proof of Theorem 6. Furthermore, let B_0 denote an open subset of $\mathrm{V}^* \setminus S$ such that $\mathrm{V}^* \subseteq B_0 \cup \bigcup_{\ell=1}^{L} B(\boldsymbol{\alpha}_\ell, \delta_\ell)$. Let $\varphi_0, \dots, \varphi_L \subseteq C^\infty(\overline{\mathrm{V}^*})$ denote a partition of unity subject to this open covering. By Theorem 6, in each ball we have

$$w(\boldsymbol{\alpha}) = \sum_{\mathcal{I} \subset \mathbf{J}(\boldsymbol{\alpha}_\ell)} \prod_{\boldsymbol{j} \in \mathcal{I}} \sqrt{\kappa^2 - |\boldsymbol{\alpha} - \boldsymbol{j}|^2}\; w_{\ell,\mathcal{I}}(\boldsymbol{\alpha})\,, \qquad \boldsymbol{\alpha} \in B(\boldsymbol{\alpha}_\ell, \delta_\ell), \quad \ell = 1, \dots, L\,,$$

with $w_{\ell,\mathcal{I}}$ analytic in $B(\boldsymbol{\alpha}_\ell, \delta_\ell)$. Let $\mathbf{J} = \bigcup_{\ell=1}^{L} \mathbf{J}(\boldsymbol{\alpha}_\ell)$ and define $w_{\ell,\mathcal{I}} = 0$ for $\mathcal{I} \subseteq \mathbf{J}$, but $\mathcal{I} \nsubseteq \mathbf{J}(\boldsymbol{\alpha}_\ell)$, $\ell = 1, \dots, L$. Since the function w on B_0 is itself analytic according to the first part of Theorem 5, we set $w_{0,\emptyset} = w$ and $w_{0,\mathcal{I}} = 0$ for all other $\mathcal{I} \subseteq \mathbf{J}$. Finally, on V^* we define

$$v_{\mathcal{I}} = \sum_{\ell=0}^{L} \varphi_\ell\, w_{\ell,\mathcal{I}}, \qquad \mathcal{I} \subseteq \mathbf{J}\,,$$

where we extend each product on the right-hand side by 0 outside its domain of definition. Then

$$w(\boldsymbol{\alpha}) = \sum_{\mathcal{I} \subseteq \mathbf{J}} \prod_{\boldsymbol{j} \in \mathcal{I}} \sqrt{\kappa^2 - |\boldsymbol{\alpha} - \boldsymbol{j}|^2}\; v_{\mathcal{I}}(\boldsymbol{\alpha})\,, \qquad \boldsymbol{\alpha} \in \mathrm{V}^*\,.$$

By definition, $v_{\mathcal{I}} \in C^\infty(\mathrm{V}^*; \widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_{\mathrm{H}}))$. A standard estimate for analytic functions (see Theorem 2.2.7, [27]) gives that for some constant C

$$\max_{\boldsymbol{\alpha}\in B(\boldsymbol{\alpha}_\ell,\delta_\ell)} \left\| \frac{\partial^\mu w_{\ell,\mathcal{I}}}{\partial \alpha_\nu^\mu} \right\|_{\widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_{\mathrm{H}})} \leq \frac{C\,\mu!}{\delta_\ell^\mu}, \qquad \nu = 1,2, \quad \mu \in \mathbb{N}_0, \quad \ell = 1,\dots,L. \tag{19}$$

Finally, we uniformly bound each derivative of w on $\overline{B_0}$ and for some $\widetilde{\delta} > 0$, $\mathrm{dist}(B_0, S) \geq \widetilde{\delta} > 0$. Thus, together with bounds on the derivative of the function φ_ℓ, we obtain the assertion. □

4 | A NUMERICAL INVERSION OF THE FLOQUET–BLOCH TRANSFORM

We propose a numerical scheme to obtain the total field in a scattering problem by combining a numerical method, such as the finite element method, to compute the transformed field $w(\boldsymbol{\alpha})$ for fixed $\boldsymbol{\alpha}$ with a tailor-made quadrature rule to approximate the inverse Floquet–Bloch transform to high order. The regularity properties of the transformed field reported in the previous section are an essential prerequisite for the derivation of such a rule. According to (7), the total field is calculated by the inverse Floquet–Bloch transform as

$$u(\widetilde{\boldsymbol{x}} + 2\pi\boldsymbol{j}, x_3) = \int_{\mathrm{V}^*} w(\boldsymbol{\alpha}, \boldsymbol{x})\, e^{\mathrm{i}\boldsymbol{\alpha}\cdot(\widetilde{\boldsymbol{x}}+2\pi\boldsymbol{j})}\, \mathrm{d}\boldsymbol{\alpha}, \quad \boldsymbol{x} \in \Omega^{2\pi}_{\mathrm{H}}, \quad \boldsymbol{j} \in \mathbb{Z}^2. \tag{20}$$

For an analysis of the approximation of this integral, it obviously suffices to consider the case $\boldsymbol{j} = 0$ as the analytic phase factor $\exp(\mathrm{i}\boldsymbol{\alpha}\cdot 2\pi\boldsymbol{j})$ does not affect the regularity of the integrand.

A naive way to approximately compute the integral in (20) is to generate an equidistant uniform square mesh in V^* and then use the set of vertices in this mesh to define a composite trapezoidal rule [14–16]. However, convergence of such an approach is typically slow: due to the square root singularities present in the representation of $w(\boldsymbol{\alpha})$ in (15), one can not even attain second order convergence in the mesh width.

We instead propose to generate a specific quadrature rule matching the a priori known structure of singularities in w to achieve high order of convergence. A recursively refined square mesh, dependent only on the wave number, is generated, with elements getting smaller with decreasing distance to the curves of singularities. On each square, except for the finest level, a tensor-product Gauss–Legendre rule is applied to approximate the integral in (20). On the finest level, a tensor-product trapezoidal rule is employed.

4.1 | Mesh generation adapted to the set of singular curves

First, note that although $\mathrm{V}^* = [-0.5, 0.5]^2$, it suffices to generate a mesh on $[0, 0.5]^2$ due to the symmetry of the curves of singular points S (see Figure 3 for an illustration). We start by subdividing $[0, 0.5]^2$ into squares of lateral length $h_0 = 1/(2n_0)$ for some $n_0 \in \mathbb{N}_{\geq 2}$. Then N refinement steps are taken, further subdividing those squares close to the curves of singular points, which are circular arcs of radius κ centred at $\boldsymbol{j} \in \widetilde{\mathbf{J}} := \cup_{\boldsymbol{\alpha}\in[0,0.5]^2}\mathbf{J}(\boldsymbol{\alpha})$. The complete procedure is presented as Algorithm 1 whose the output is illustrated in Figure 5 for $N = 6$ and different values of the wave numbers κ.

In the proposition below, we list properties of the adapted mesh $\mathcal{G}_N$ generated by Algorithm 1. To concisely formulate these results, we introduce the sets of squares of lateral length $h_n = h_0/2^n$ in the mesh,

$$\mathcal{M}_n = \{K : K \in \mathcal{G}_N \text{ and } K \text{ has lateral length } h_n\}, \qquad n = 0,\dots,N, \tag{21}$$

as well as the union of all squares of lateral length h_n,

$$\mathcal{R}_n = \bigcup_{K\in\mathcal{M}_n} \overline{K}, \qquad n = 0,\dots,N. \tag{22}$$

ALGORITHM 1 Generate adapted mesh

Input: $\kappa, N, n_0, \tilde{\mathbf{J}}$
$h_0 \leftarrow 1/(2n_0)$
$\mathcal{G}_0 \leftarrow \left\{[\mu_1 h_0, (\mu_1+1)h_0] \times [\mu_2 h_0, (\mu_2+1)h_0] : \mu_1, \mu_2 = 0, \dots, n_0 - 1\right\}$
for $n = 1, \dots, N$ **do**
  $\mathcal{G}_n \leftarrow \{\}$
  $h_n \leftarrow h_{n-1}/2$
  for $K \in \mathcal{G}_{n-1}$ **do**
    let $\boldsymbol{\xi}_K$ denote the center of K
    $\text{dist}(\boldsymbol{\xi}_K, \mathcal{S}) \leftarrow \min_{\boldsymbol{j}\in\tilde{\mathbf{J}}} |\kappa - |\boldsymbol{\xi}_K - \boldsymbol{j}||$
    if $\text{dist}(\boldsymbol{\xi}_K, \mathcal{S}) \le 1/2^n$ **then**
      Refine K into $K_1, \dots, K_4$ of lateral length h_n
      $\mathcal{G}_n \leftarrow \mathcal{G}_n \cup \{K_1, \dots, K_4\}$
    else
      $\mathcal{G}_n \leftarrow \mathcal{G}_n \cup \{K\}$
return $\mathcal{G}_N$

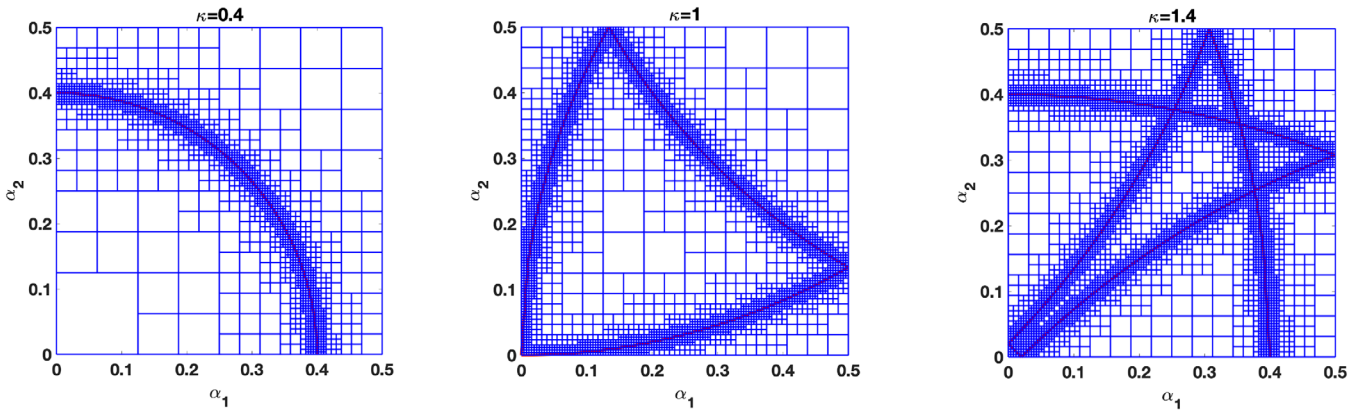


FIGURE 5 The generated adapted mesh $\mathcal{G}_6$ for different κ by Algorithm 1.

Proposition 8. *Let the square $K \in \mathcal{M}_n$ (for $n = 1, \dots, N$) with center $\boldsymbol{\xi}_K$, then*

$$\text{dist}(\boldsymbol{\xi}_K, \mathcal{S}) > \frac{1}{2^{n+1}}, \qquad n = 0, \dots, N-1,$$

$$\text{dist}(\boldsymbol{\xi}_K, \mathcal{S}) \le \frac{1}{2^n}\left(1 + \frac{\sqrt{2}}{2}h_0\right), \qquad n = 1, \dots, N.$$

Furthermore,

$$\text{dist}(\mathcal{R}_n, \mathcal{S}) \ge \frac{1}{2^{n+1}}\left(1 - \sqrt{2}h_0\right) =: d_{\min,n}, \qquad n = 0, \dots, N-1, \tag{23}$$

$$\sup_{x\in\mathcal{R}_n} \text{dist}(x, \mathcal{S}) \le \frac{1}{2^n}\left(1 + \sqrt{2}h_0\right) =: d_{\max,n}, \qquad n = 1, \dots, N. \tag{24}$$

Proof. Consider the square $K \in \mathcal{M}_n$, $n = 1, \dots, N$, with center $\boldsymbol{\xi}_K$. According to Algorithm 1, K is generated by refining a larger square $\widetilde{K} \in \mathcal{M}_{n-1}$. The center $\boldsymbol{\xi}_{\widetilde{K}}$ of $\widetilde{K}$ satisfies the condition

$$\text{dist}(\boldsymbol{\xi}_{\widetilde{K}}, \mathcal{S}) = |\kappa - |\boldsymbol{\xi}_{\widetilde{K}} - \boldsymbol{j}|| \le \frac{1}{2^n} \quad \text{for at least for one } \boldsymbol{j} \in \tilde{\boldsymbol{J}}. \tag{25}$$

Based on the refinement, we first conclude that $|\boldsymbol{\xi}_K - \boldsymbol{\xi}_{\widetilde{K}}| = (\sqrt{2}/4)\, h_{n-1}$, and hence from (25) that

$$\operatorname{dist}(\boldsymbol{\xi}_K, \mathcal{S}) \le \operatorname{dist}(\boldsymbol{\xi}_K, \boldsymbol{\xi}_{\widetilde{K}}) + \operatorname{dist}(\boldsymbol{\xi}_{\widetilde{K}}, \mathcal{S}) \le \frac{1}{2^n}\left(1 + \frac{\sqrt{2}}{2}\, h_0\right), \qquad n = 1, \dots, N\,. \tag{26}$$

A bound for $x \in K$ is obtained by adding half of the diameter of K,

$$\operatorname{dist}(x, \mathcal{S}) \le \frac{\sqrt{2}}{2}\, h_n + \frac{1}{2^n}\left(1 + \frac{\sqrt{2}}{2}\, h_0\right) = \frac{1}{2^n}\left(1 + \sqrt{2}\, h_0\right).$$

As the right-hand side is independent of K, it actually holds for all $x \in \mathcal{R}_n$.

On the other hand, any $K \in \mathcal{M}_n$, $n = 0, \dots, N-1$, that was not subject to the refinement in the $(n+1)$-th refinement step, it implies

$$\operatorname{dist}(\boldsymbol{\xi}_K, \mathcal{S}) > \frac{1}{2^{n+1}}, \qquad n = 0, \dots, N-1. \tag{27}$$

Hence, for any $x \in K$, we have

$$\operatorname{dist}(x, \mathcal{S}) \ge \operatorname{dist}(\boldsymbol{\xi}_K, \mathcal{S}) - \operatorname{diam}(K)/2 > \frac{1}{2^{n+1}} - \frac{\sqrt{2}}{2}\, h_n = \frac{1}{2^{n+1}}\left(1 - \sqrt{2}\, h_0\right).$$

As the right-hand side is independent of K, the estimate holds for any $x \in \mathcal{R}_n$. □

Remark 9. Proposition 8 shows that every set $\mathcal{R}_n$ is covered by annuli for which we have explicit bounds for inner and outer radius. As each $\mathcal{R}_n$ is the union of the equally sized squares in $\mathcal{M}_n$, we may estimate the number of squares in $\mathcal{M}_n$. For $n = N$, we have

$$|\mathcal{R}_N| \le \pi\left(\kappa + d_{\max,N}\right)^2 - \pi\left(\kappa - d_{\max,N}\right)^2 = 4\pi\kappa\, d_{\max,N} = \frac{4\pi\kappa}{2^N}\left(1 + \sqrt{2}\, h_0\right),$$

and hence

$$\#\mathcal{M}_N = \frac{|\mathcal{R}_N|}{h_N^2} \le \frac{4\pi\kappa}{h_0}\left(\sqrt{2} + \frac{1}{h_0}\right)2^N\,.$$

Similarly, for $n = 1, \dots, N-1$,

$$|\mathcal{R}_n| \le 4\pi\kappa\,(d_{\max,n} - d_{\min,n}) = \frac{2\pi\kappa}{2^n}\left(1 + 3\sqrt{2}\, h_0\right),$$

and

$$\#\mathcal{M}_n = \frac{|\mathcal{R}_n|}{h_n^2} \le \frac{2\pi\kappa}{h_0}\left(3\sqrt{2} + \frac{1}{h_0}\right)2^n\,.$$

We will now proceed with defining appropriate quadrature rules on each square in $\mathcal{G}_N$ and then analyze the corresponding error in computing the integral. We will strongly rely on the correspondence of the squares in the mesh to representations of the integrand w. In accordance with Theorem 6, we may use (15) for w on the smallest squares if $\mathcal{R}_N \subseteq \widetilde{\mathcal{S}}$ and if $h_N < \sqrt{2}\,\delta$. In the first step, we will use this observation to estimate the error of applying a composite trapezoidal rule on $\mathcal{R}_N$. Afterwards, we investigate the error of a P-point Gaussian quadrature rule applied on all other squares, making use of the representation as derived in Theorem 7. Finally, it is proved that combining both rules for approximating the inverse Floquet–Bloch transform is super-algebraically convergent.

Recall that it suffices to consider the case $\boldsymbol{j} = 0$ when approximating (20). Led by the properties of the transformed total field established in Section 3, let us first sum up all required assumptions for the integrand. Also recall the definition of the space $\mathcal{X}$ in (14).

Assumption 10. We assume that $w \in \mathcal{X}(H^1_{\mathrm{per}}(\Omega^{2\pi}_{\mathrm{H}}))$ and that r, δ denote the corresponding numbers from Theorem 6. Note that w then will also admit the representation (17).

4.2 | The trapezoidal rule on the smallest squares

We first consider a square $K \in \mathcal{M}_N$ with center $\boldsymbol{\xi}_K = (\xi_{K,1}, \xi_{K,2})$. The vertices of K are given by $\boldsymbol{\alpha}_{p,q} = \boldsymbol{\xi}_K + (p - \frac{1}{2})\, h_N\, \boldsymbol{e}^{(1)} + (q - \frac{1}{2})\, h_N\, \boldsymbol{e}^{(2)}$, $p,\ q = 0, 1$, where $\boldsymbol{e}^{(j)}$ denotes the j-th coordinate vector. The integral w over K is approximated by the trapezoidal rule

$$\int_K w(\boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{\alpha} = \frac{h_N^2}{4} \sum_{p,q=0}^{1} w(\boldsymbol{\alpha}_{p,q}) + \mathrm{E}_K^t w\,,$$

where $E_K^t w$ denotes the error. To estimate $E_K^t w$, we require the bilinear interpolation operator of the transformed field in the points $\boldsymbol{\alpha}_{p,q}$, which we shall denote by P_K. Well-known estimates for interpolation give

$$\max_{\boldsymbol{\alpha} \in K} |f(\boldsymbol{\alpha}) - P_K f(\boldsymbol{\alpha})| \le C \max_{\nu=1,2} \left\| \frac{\partial^2 f}{\partial \alpha_\nu^2} \right\|_\infty h_N^2, \tag{28}$$

for any $f \in C^2(\overline{K})$. This estimate of course generalizes to C^2-smooth functions on K with values in a Sobolev space.

Theorem 11. *Let w satisfy Assumption 10 and let h_0, N be chosen such that $d_{\max,N} < r$, $h_N \le \sqrt{2}\delta$. Then*

$$\max_{\boldsymbol{\alpha} \in K} \|w(\boldsymbol{\alpha}) - P_K w(\boldsymbol{\alpha})\|_{\widetilde{H}^1_{\mathrm{per}}(\Omega_{\mathrm{H}}^{2\pi})} \le \mathrm{C}\, 2^{-N/2}\,,$$

where the constant C depends on κ and the functions w_I appearing in (15) for all the centers of expansion from Remark 6.

Proof. According to Theorem 6, there exists $\boldsymbol{\alpha}_0 \in S$ such that the representation

$$w(\boldsymbol{\alpha}) = \sum_{I \subseteq \mathbf{J}(\boldsymbol{\alpha}_0)} \prod_{\boldsymbol{j} \in I} \sqrt{\kappa^2 - |\boldsymbol{\alpha} - \boldsymbol{j}|^2}\; v_I(\boldsymbol{\alpha})\,,$$

with analytic functions w_I, holds for all $\boldsymbol{\alpha} \in K$. To establish the assertion, it is necessary to distinguish between curves of singular points close to K and those at a larger distance. Hence, define

$$J_1 = \left\{ \boldsymbol{j} \in \mathbf{J}(\boldsymbol{\alpha}_0) \,:\, |\kappa - |\boldsymbol{\alpha} - \boldsymbol{j}|| \le d_{\max,N} \text{ for some } \boldsymbol{\alpha} \in K \right\},$$

and $J_2 = \mathbf{J}(\boldsymbol{\alpha}_0) \setminus J_1$. To abbreviate notation, we set $\gamma_{\boldsymbol{j}}(\boldsymbol{\alpha}) = \sqrt{\kappa^2 - |\boldsymbol{\alpha} - \boldsymbol{j}|^2}$ and introduce

$$v_{I_1}(\boldsymbol{\alpha}) = \begin{cases} w_\emptyset(\boldsymbol{\alpha}) + \displaystyle\sum_{\emptyset \ne I_2 \subseteq J_2} w_{I_2}(\boldsymbol{\alpha}) \prod_{\boldsymbol{j} \in I_2} \gamma_{\boldsymbol{j}}(\boldsymbol{\alpha})\,, & I_1 = \emptyset\,, \\ \displaystyle\sum_{I_2 \subseteq J_2} w_{I_1 \cup I_2}(\boldsymbol{\alpha}) \prod_{\boldsymbol{j} \in I_2} \gamma_{\boldsymbol{j}}(\boldsymbol{\alpha})\,, & I_1 \subseteq J_1\,,\ I_1 \ne \emptyset\,. \end{cases}$$

With this notation, the representation of w becomes

$$w(\boldsymbol{\alpha}) = \sum_{I_1 \subseteq J_1} v_{I_1}(\boldsymbol{\alpha}) \prod_{\boldsymbol{j} \in I_1} \gamma_{\boldsymbol{j}}(\boldsymbol{\alpha})\,. \tag{29}$$

The goal is thus to establish the asserted estimate for each term in (29). Throughout the arguments we shall make use of a generic C denoting constants that depend on κ, the maximum norms of derivatives of all w_I up to second order and on maximum norms of all v_I (but not their derivatives).

We start with terms for $I_1 = \emptyset$. For $w_\emptyset$, the estimate follows directly from (28). This is, in fact, also the initial step in an induction over the number of square root factors in a summand in the definition of v_{I_1}. For the induction step, assume that the estimate has been proven for some bounded continuous function z. Let $\boldsymbol{j} \in I_2$. From Lemma A2 and the definition

of J_2, we obtain

$$\left|\frac{\partial^2 \gamma_j(\boldsymbol{\alpha})}{\partial \alpha_\nu^2}\right| \le C\,\frac{(\kappa + |\boldsymbol{\alpha} - \boldsymbol{j}|)^{1/2}}{|\kappa - |\boldsymbol{\alpha} - \boldsymbol{j}||^{3/2}} \le \frac{C}{d_{\max,N}^{3/2}} \le C\,2^{3N/2}\,, \qquad \boldsymbol{\alpha} \in K\,, \quad \nu = 1,2\,. \tag{30}$$

By the induction and properties of P_K,

$$\begin{aligned}
&|P_K(\gamma_j z)(\boldsymbol{\alpha}) - \gamma_j(\boldsymbol{\alpha})\, z(\boldsymbol{\alpha})| \\
&\qquad \le |P_K(\gamma_j z)(\boldsymbol{\alpha}) - \gamma_j(\boldsymbol{\alpha})\, P_K z(\boldsymbol{\alpha})| + |\gamma_j(\boldsymbol{\alpha})\, P_K z(\boldsymbol{\alpha}) - \gamma_j(\boldsymbol{\alpha})\, z(\boldsymbol{\alpha})| \\
&\qquad \le |P_K(\gamma_j P_K z)(\boldsymbol{\alpha}) - \gamma_j(\boldsymbol{\alpha})\, P_K z(\boldsymbol{\alpha})| + C\left\|\gamma_j\right\|_{\infty;\mathrm{V}^*} 2^{-N/2}.
\end{aligned}$$

Now using Equation (28), the bilinearity of $P_K z$ and finally (30), the first term can be estimated by

$$\left|P_K(\gamma_j P_K z)(\boldsymbol{\alpha}) - \gamma_j(\boldsymbol{\alpha})\, P_K z(\boldsymbol{\alpha})\right| \le C\,\|z\|_\infty \max_{\nu=1,2}\left\|\frac{\partial^2 \gamma_j}{\partial \alpha_\nu^2}\right\|_{\infty;K} h_N^2 \le C\,2^{-N/2}\,.$$

Next, we establish the estimate for terms with $\mathcal{I}_1 \neq \emptyset$. Consider again a bounded continuous function z for which the asserted estimate is valid and let now $\boldsymbol{j} \in \mathcal{I}_1$. Similarly as before, we estimate

$$\begin{aligned}
&\left|P_K(\gamma_j z)(\boldsymbol{\alpha}) - \gamma_j(\boldsymbol{\alpha})\, z(\boldsymbol{\alpha})\right| \\
&\qquad \le |P_K(\gamma_j z)(\boldsymbol{\alpha}) - \gamma_j(\boldsymbol{\alpha})\, P_K z(\boldsymbol{\alpha})| + C\,\|\gamma_j\|_{\infty;K}\, 2^{-N/2} \le C\left(1 + 2^{-N/2}\right)\left\|\gamma_j\right\|_{\infty;\mathcal{R}_N}.
\end{aligned}$$

By the definition of J_1, it follows that

$$\left\|\gamma_j\right\|_{\infty;\mathcal{R}_N} \le C\,|\kappa - |\boldsymbol{\alpha} - \boldsymbol{j}|| \le C\left(d_{\max,N} + \operatorname{diam}(K)\right) \le C\left(2^{-N} + \sqrt{2}\,h_N\right) \le C\,2^{-N}\,.$$

By induction, the asserted estimate now follows for all terms in (29). □

It is now straightforward to obtain a bound for approximating the integral on the union of all $K \in \mathcal{M}_N$. The corresponding quadrature operator will be denoted by

$$I_N^T w = \sum_{K \in \mathcal{M}_N} \int_K P_K w(\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\alpha}\,.$$

Theorem 12. *Let w satisfy Assumption 10 and let N be chosen such that $d_{\max,N} < r$, $h_N \le \sqrt{2}\delta$. Then, the error of the trapezoidal rule over $\mathcal{R}_N$ is bounded by*

$$\left\|\int_{\mathcal{R}_N} w(\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\alpha} - I_N^T w\right\|_{\widetilde{H}^1_{\mathrm{per}}(\Omega_{\mathrm{H}}^{2\pi})} \le C\,2^{-3N/2}.$$

Proof. By using the triangle inequality and Theorem 11, we have

$$\begin{aligned}
\left\|\int_{\mathcal{R}_N} w(\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\alpha} - I_N^T w\right\|_{\widetilde{H}^1_{\mathrm{per}}(\Omega_{\mathrm{H}}^{2\pi})} &\le \sum_{K \in \mathcal{M}_N} \int_K \|(w - P_K w)(\boldsymbol{\alpha})\|_{\widetilde{H}^1_{\mathrm{per}}(\Omega_{\mathrm{H}}^{2\pi})}\,\mathrm{d}\boldsymbol{\alpha} \\
&\le \mathrm{C}(\kappa)\,(\#\mathcal{M}_N)\,h_N^2\,2^{-N/2}\,.
\end{aligned}$$

By using Remark 9, which establishes $\#\mathcal{M}_N \sim 2^N$, and the construction $h_N \sim 2^{-N}$, the assertion follows. □

4.3 | The Gauss–Legendre quadrature rule on all larger squares

On all squares $K \in \mathcal{M}_n$ for $n = 1, \dots, N-1$, we will use a P-point Gauss–Legendre quadrature rule in each coordinate direction to approximate the inverse Floquet–Bloch transform. We denote this rule applied to a function f by $I^G_{P,K} f$ and set $I^G_{P,\mathcal{R}_n} f = \sum_{K \in \mathcal{R}_n} I^G_{P,K} f$. In the next theorem, we present the well known general error estimate for applying such a rule.

Theorem 13. *Let $f \in C^{2P}(\overline{\mathcal{R}_n}; \widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_{\mathrm{H}}))$. Then, there is a constant C such that*

$$\left\| \int_{\mathcal{R}_n} f(\boldsymbol{\alpha}) \, \mathrm{d}\boldsymbol{\alpha} - I^G_{P,\mathcal{R}_n} f \right\|_{\widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_{\mathrm{H}})} \leq C \left(\frac{h_0}{2} \right)^{2P} \frac{2^{-(2P+1)n}}{(2P+1)!} \max_{\boldsymbol{\alpha} \in \mathcal{R}_n} \left(\sum_{\nu=1}^{2} \left\| \frac{\partial^{2P} f(\boldsymbol{\alpha})}{\partial \alpha_\nu^{2P}} \right\|_{\widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_{\mathrm{H}})} \right).$$

Proof. Extending standard estimate for the P-point Gauss–Legendre quadrature rule (see e.g., [28, Theorem 9.20]) to the two-dimensional case and our setting of functions mapping to a Sobolev space, gives

$$\left\| \int_{K} f(\boldsymbol{\alpha}) \, \mathrm{d}\boldsymbol{\alpha} - I^G_{P,K} f \right\|_{\widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_{\mathrm{H}})} \leq \frac{4}{(2P+1)!} \left(\frac{h_n}{2} \right)^{2P+2} \max_{\boldsymbol{\alpha} \in K} \left(\sum_{\nu=1}^{2} \left\| \frac{\partial^{2P} f(\boldsymbol{\alpha})}{\partial \alpha_\nu^{2P}} \right\|_{\widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_{\mathrm{H}})} \right).$$

Using the estimates from Remark 9, we obtain the asserted error bound. □

Based on Theorem 13, the error of the Gauss–Legendre rule for computing the integral of w over $\mathcal{R}_n$ depends on the $2P$-th partial derivatives of w with respect to either α_1 or α_2. Recalling the representation (17), it suffices to estimate the $2P$-th partial derivatives of $\prod_{j \in \mathcal{I}} \sqrt{\kappa^2 - |\boldsymbol{\alpha} - \mathbf{j}|^2}\, v_{\mathcal{I}}(\boldsymbol{\alpha})$ with respect to only one coordinate. We do so in the next lemma using some standard estimates for square root functions and their derivatives presented in the appendix.

Lemma 14. *For any fixed $\ell \in \mathbb{N}$, there is a constant C such that*

$$\max_{\boldsymbol{\alpha} \in \mathcal{R}_n} \left| \frac{\partial^\ell \sqrt{\kappa^2 - |\boldsymbol{\alpha} - \mathbf{j}|^2}}{\partial \alpha_\nu^\ell} \right| \leq \frac{C\, \ell! \left(d_{\max,n} \right)^{1/2}}{\left(d_{\min,n} \right)^{\ell}}, \qquad \text{for } n = 1, \dots, N-1, \quad \nu = 1, 2, \tag{31}$$

where $d_{\min,n}$ and $d_{\max,n}$ are defined by (23) and (24), respectively.

Proof. According to Lemma A2 in Appendix, for all $\boldsymbol{\alpha} \in \mathcal{R}_n$, $n = 1, \dots, N-1$, there is a constant C_1 such that

$$\left| \frac{\partial^\ell \sqrt{\kappa^2 - |\boldsymbol{\alpha} - \mathbf{j}|^2}}{\partial \alpha_\nu^\ell} \right| \leq \frac{C_1\, \ell! \left| \kappa + |\boldsymbol{\alpha} - \mathbf{j}| \right|^{1/2}}{\left| \kappa - |\boldsymbol{\alpha} - \mathbf{j}| \right|^{\ell - 1/2}}.$$

Hence, using Equations (23), (24), that is, $d_{\min,n} \leq |\kappa - |\boldsymbol{\alpha} - \mathbf{j}|| \leq d_{\max,n}$, leads to

$$\max_{\boldsymbol{\alpha} \in \mathcal{R}_n} \left| \frac{\partial^\ell \sqrt{\kappa^2 - |\boldsymbol{\alpha} - \mathbf{j}|^2}}{\partial \alpha_\nu^\ell} \right| \leq \frac{C\, \ell! \left(d_{\max,n} \right)^{1/2}}{\left(d_{\min,n} \right)^{\ell}}.$$

□

Theorem 15. *Let $\mathcal{I} \subseteq \mathbf{J}$ and denote by $\Lambda_{\mathcal{I}}(\boldsymbol{\alpha}) := \prod_{j \in \mathcal{I}} \sqrt{\kappa^2 - |\boldsymbol{\alpha} - \mathbf{j}|^2}\, v_{\mathcal{I}}(\boldsymbol{\alpha})$ one of the terms in (17). Let $m = \#\mathcal{I}$. Then, for every $\ell \in \mathbb{N}_0$ there exists $C_\ell > 0$ such that*

$$\max_{\boldsymbol{\alpha} \in \mathcal{R}_n} \left\| \frac{\partial^\ell \Lambda_{\mathcal{I}}}{\partial \alpha_\nu^\ell} \right\|_{\widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_{\mathrm{H}})} \leq \frac{C_\ell \left(d_{\max,n} \right)^{m/2}}{\left(d_{\min,n} \right)^{\ell}}, \qquad \nu = 1, 2. \tag{32}$$

Proof. From the generalized Leibniz formula, we obtain

$$\frac{\partial^\ell \Lambda_I(\boldsymbol{\alpha})}{\partial \alpha_\nu^\ell} = \sum_{\mathrm{K}_0+\cdots+\mathrm{K}_m=\ell} \frac{\ell!}{\mathrm{K}_0!\cdots\mathrm{K}_m!} \frac{\partial^{\mathrm{K}_0} v_I(\boldsymbol{\alpha})}{\partial \alpha_\nu^{\mathrm{K}_0}} \prod_{\mu=1}^{m} \frac{\partial^{\mathrm{K}_\mu}}{\partial \alpha_\nu^{\mathrm{K}_\mu}} \sqrt{\kappa^2 - |\boldsymbol{\alpha}-\boldsymbol{j}_\mu|^2}\,.$$

Using (18) and Lemma 14 yields for $\boldsymbol{\alpha} \in \mathcal{R}_n$

$$\left\| \frac{\partial^\ell \Lambda_I(\boldsymbol{\alpha})}{\partial \alpha_\nu^\ell} \right\|_{\widetilde{H}^1_{\mathrm{per}}(\Omega_{\mathrm{H}}^{2\pi})} \le C \sum_{\mathrm{K}_0+\cdots+\mathrm{K}_m=\ell} \frac{\ell!}{\mathrm{K}_0!\cdots\mathrm{K}_m!} \frac{C_{\mathrm{K}_0}}{\left(d_{\min,n}\right)^{\mathrm{K}_0}} \prod_{\mu=1}^{m} \frac{\mathrm{K}_\mu! \left(d_{\max,n}\right)^{1/2}}{\left(d_{\min,n}\right)^{\mathrm{K}_\mu}}.$$

Combining all constants gives the assertion. □

Theorem 16. *Let w satisfy Assumption 10. Then, for every $P \in \mathbb{N}$, there exists a constant C_P such that*

$$\left\| \sum_{n=1}^{N-1} \int_{\mathcal{R}_n} w(\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\alpha} - \sum_{n=1}^{N-1} I^G_{P,\mathcal{R}_n} w \right\|_{\widetilde{H}^1_{\mathrm{per}}(\Omega_{\mathrm{H}}^{2\pi})} \le C_P\, h_0^{2P}\,.$$

Proof. Combining Theorems 13 and 15, we obtain the estimate

$$\left\| \int_{\mathcal{R}_n} w(\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\alpha} - I^G_{P,\mathcal{R}_n} w \right\|_{\widetilde{H}^1_{\mathrm{per}}(\Omega_{\mathrm{H}}^{2\pi})} \le C_P \left(\frac{h_0}{2}\right)^{2P} \frac{2^{-(2P+1)n}}{(d_{\min,n})^{2P}},$$

with some constant C_P independent of h_0 and n. From (23), we have $d_{\min,n} \ge C\,2^{-n}$. Hence, we conclude

$$\left\| \int_{\mathcal{R}_n} w(\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\alpha} - I^G_{P,\mathcal{R}_n} w \right\|_{\widetilde{H}^1_{\mathrm{per}}(\Omega_{\mathrm{H}}^{2\pi})} \le C_P\, \frac{h_0^{2P}}{2^n}\,.$$

Summing over $n = 1, \ldots, N-1$ completes the proof. □

Now, we are going to provide the analysis of the total error in numerical solution of the main non-periodic scattering problem (1–4).

4.4 | The combined quadrature rule

It is now straightforward to combine the quadrature rules of both the previous two subsections to up to a super-algebraically convergent approximation to the Floquet–Bloch transform of the total field.

Corollary 17. *Let w satisfy Assumption 10 and fix $P \in \mathbb{N}$. Then there is $C_P > 0$ such that for every h_0 and N with $d_{\max,N} < r$, $h_N \le \sqrt{2}\delta$, there holds*

$$\left\| \int_{\mathrm{V}^*} w(\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\alpha} - I_N^T w - \sum_{n=1}^{N-1} I^G_{P,\mathcal{R}_n} w \right\|_{\widetilde{H}^1_{\mathrm{per}}(\Omega_{\mathrm{H}}^{2\pi})} \le C_P\left(2^{-3N/2} + h_0^{2P}\right).$$

Example 18. As examples for the performance achievable with our quadrature rule, we consider functions w that are simply products of the square root functions occurring in the representation (15). In this special case, all w_I are either constant 0 or 1 and thus analytic on V^*. From (19) and the estimates in the proof of Theorem 15 we expect the constant C_P to be independent of P in this case.

We apply the quadrature rule to the approximation of two integrals,

$$\mathbf{I}_1 = \int_{\mathrm{V}^*} \sqrt{\kappa^2 - |\boldsymbol{\alpha}-\boldsymbol{j}|^2}\,\mathrm{d}\boldsymbol{\alpha}\,, \qquad \kappa = 0.4\,,\ \boldsymbol{j} = (0,0)\,,$$

$$\mathbf{I}_2 = \int_{\mathrm{V}^*} \sqrt{\kappa^2 - |\boldsymbol{\alpha}-\boldsymbol{j}|^2}\,\sqrt{\kappa^2 - |\boldsymbol{\alpha}-\boldsymbol{l}|^2}\,\mathrm{d}\boldsymbol{\alpha}\,, \qquad \kappa = 1.4\,,\ \boldsymbol{j} = (-1,0)\,,\ \boldsymbol{l} = (-1,1)\,.$$

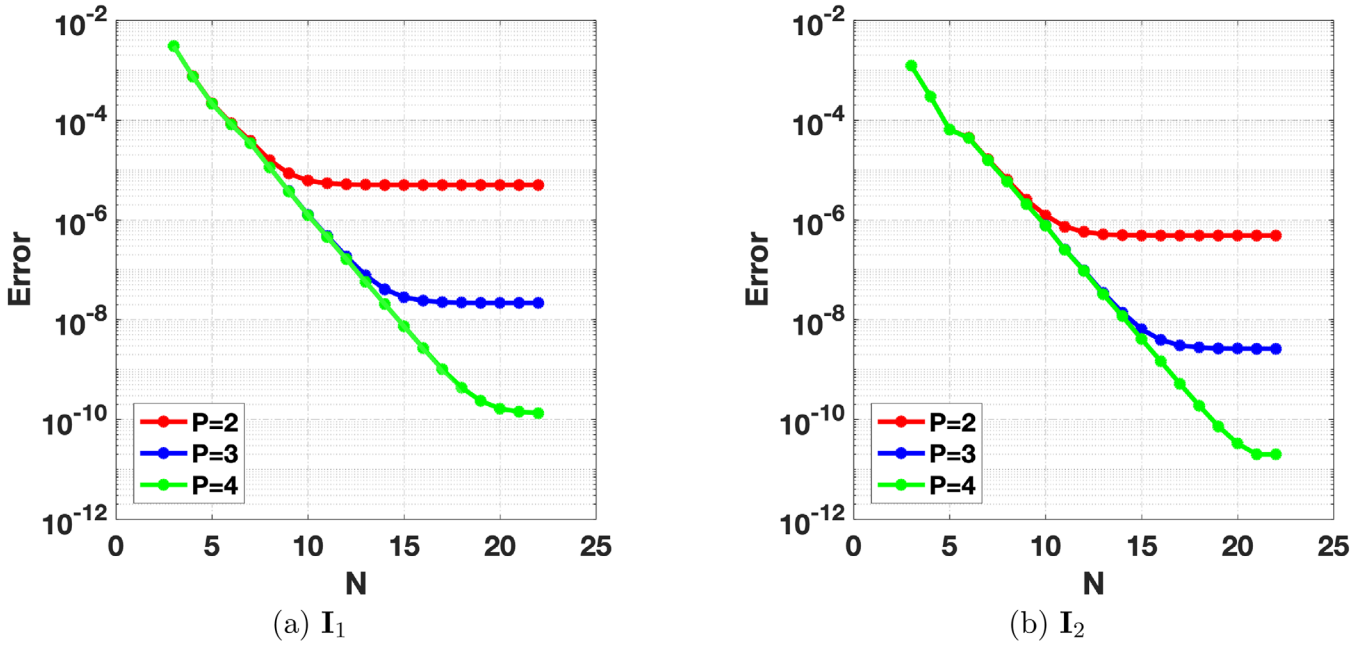


FIGURE 6 Difference between the value computed for $\mathbf{I}_j$, $j = 1, 2$ by using the quadrature rule for various N and P and the exact value ($j = 1$) or reference value ($j = 2$), respectively.

For the first integral, the set S is a single circle entirely contained in the set V^*. Hence, the exact value of the integral $\mathbf{I}_1$ can be obtained analytically. We have used Maple 2022 to carry out this task and then computed approximations using our quadrature rule for various values of N and P.

In the second integral, the integrand is singular along two circular arcs contained in the set V^*. The exact value of this integral is not available. Instead, we have computed a reference value for $N = 23$ and $P = 5$ and compare our results against this.

The results are presented in Figure 6. The theoretically predicted convergence rate from Corollary 17 is very well reflected, with exponential convergence with respect to N dominating the result for small N, until the error of the Gauss quadrature rule becomes dominant. The results also nicely illustrate our expectation that C_P is independent of P for these examples.

To conclude our analysis, we combine the result of Corollary 17 with error bounds for the Galerkin approximation of the solution of the variational Equation (12).

Theorem 19. *Let $\mathcal{F}(\boldsymbol{\alpha}) \in H^{-1/2}_{\mathrm{per}}(\Gamma^{2\pi}_{\mathrm{H}})$ and $w(\boldsymbol{\alpha})$ denote the exact solution of the variational formulation of (12) and $w_\tau(\boldsymbol{\alpha})$ its numerical approximation by the finite element method with mesh size τ. For sufficiently small τ,*

$$\|w(\boldsymbol{\alpha}) - w_\tau(\boldsymbol{\alpha})\|_{H^s(\Omega^{2\pi}_{\mathrm{H}})} \leq C\,\tau^{2-s}\,\|\mathcal{F}(\boldsymbol{\alpha})\|_{H^{-1/2}_{\mathrm{per}}(\Gamma^{2\pi}_{\mathrm{H}})}, \quad \text{for} \quad s = 0, 1\,,$$

where C is independent of $\boldsymbol{\alpha}$.

Proof. The proof is completely analogous to that of Theorem 16 presented in ref. [10]. □

Combing both error bounds yields the complete estimate for the proposed numerical method. To concisely formulate this result, we introduce operators

$$\mathrm{Y}_j\psi(\boldsymbol{\alpha}, \boldsymbol{x}) = \psi(\boldsymbol{\alpha}, \boldsymbol{x})\,\mathrm{e}^{\mathrm{i}\boldsymbol{\alpha}\cdot(\widetilde{\boldsymbol{x}}+2\pi\boldsymbol{j})} \quad \text{and} \quad \mathcal{J}^{-1}_{P,N,h_0}\psi(\widetilde{\boldsymbol{x}} + 2\pi\boldsymbol{j}, x_3) = \left(I^T_N + \sum_{n=1}^{N-1} I^G_{P,\mathcal{R}_n}\right)\mathrm{Y}_j\psi(\boldsymbol{x})\,.$$

Theorem 20. *Let $u^i \in H^1_r(\Omega_H)$ for some $|r| < 1$ and additionally $\mathcal{F} \in \mathcal{X}(H^{-1/2}_{\mathrm{per}}(\Gamma^{2\pi}_H))$. Let u denote the total field, that is, the solution to (5), and for any $\boldsymbol{\alpha} \in \mathrm{V}^*$ by $w_\tau(\boldsymbol{\alpha})$ the finite element approximation to the solution of (12) for sufficiently small mesh size τ. Let h_0 and N satisfy $d_{\max,N} < r$, $h_N \le \sqrt{2}\delta$ and fix $P \in \mathbb{N}$. Then there holds the error estimate*

$$\|u - \mathcal{J}^{-1}_{P,N,h_0} w_\tau\|_{H^s(\Omega^{2\pi}_H)} \le C\Big(\tau^{2-s} + 2^{-3N/2} + {h_0}^{2P}\Big), \qquad s = 0, 1\,,$$

where C depends on P and u^i.

Proof. For any $\boldsymbol{\alpha} \in \mathrm{V}^*$, denote by $w(\boldsymbol{\alpha})$ the exact solution to (12). By using the inverse Floquet–Bloch transform and then the triangle inequality, we have

$$\begin{aligned}\left\|u - \mathcal{J}^{-1}_{P,N,h_0} w_\tau\right\|_{H^s(\Omega^{2\pi}_H)} &= \left\|\mathcal{J}^{-1} w - \mathcal{J}^{-1}_{P,N,h_0} w_\tau\right\|_{H^s(\Omega^{2\pi}_H)} \\ &\le \left\|\left(\mathcal{J}^{-1} - \mathcal{J}^{-1}_{P,N,h_0}\right) w\right\|_{H^s(\Omega^{2\pi}_H)} + \left\|\mathcal{J}^{-1}_{P,N,h_0}(w - w_\tau)\right\|_{H^s(\Omega^{2\pi}_H)}. \end{aligned} \tag{33}$$

Note that application of Y_j is just a multiplication with an analytic function, hence $\mathrm{Y}_j w$ satisfies Assumption 10. For the first term of (33), Corollary 17 gives

$$\left\|\left(\mathcal{J}^{-1} - \mathcal{J}^{-1}_{P,N,h_0}\right) w\right\|_{H^s(\Omega^{2\pi}_H)} \le C_P\big(2^{-3N/2} + h_0^{2P}\big).$$

Denote by $\boldsymbol{\alpha}_\ell$, ϱ_ℓ, for $\ell = 1, \dots, Q$, all the quadrature points and corresponding weights appearing in the rules I^T_N and $I^G_{P,\mathcal{R}_n}$, respectively. It should be noted that all the weights are positive. Accordingly, we may write using Theorem 19,

$$\left\|\mathcal{J}^{-1}_{P,N,h_0}(w - w_\tau)\right\|_{H^s(\Omega^{2\pi}_H)} \le \sum_{\ell=1}^{Q} \varrho_\ell \|w(\boldsymbol{\alpha}_\ell) - w_\tau(\boldsymbol{\alpha}_\ell)\|_{H^s(\Omega^{2\pi}_H)} \le C\tau^{2-s} \sum_{\ell=1}^{Q} \varrho_\ell \, \|\mathcal{F}(\boldsymbol{\alpha}_\ell)\|_{H^{-1/2}_{\mathrm{per}}(\Gamma^{2\pi}_H)}.$$

As $\mathcal{F} \in \mathcal{X}(H^{-1/2}_{\mathrm{per}}(\Gamma^{2\pi}_H))$, we may use the same approach as in the proof of Theorem 7 to derive an expression analogous to (17) for $\mathcal{F}$ and conclude that $\sup_{\boldsymbol{\alpha}\in\mathrm{V}^*} \|\mathcal{F}(\boldsymbol{\alpha})\|_{H^{-1/2}_{\mathrm{per}}(\Omega^{2\pi}_H)} < \infty$. Then, using the fact that $\sum_{\ell=1}^{Q} \varrho_\ell = |\mathrm{V}^*| = 1$, the proof is completed. □

5 | NUMERICAL RESULTS

In this section, we present numerical examples to illustrate the performance of the proposed method for solving the three-dimensional scattering problems. To have access to an exact solution, we consider the case of a radiation problem: We assume that $\Gamma \subseteq \mathbb{R}^3_+$, where $\mathbb{R}^3_+ := \{\boldsymbol{x} \in \mathbb{R}^3_+ : x_3 > 0\}$ is the upper half-space and that u^i is the Dirichlet Green's function for this upper half-space for some source point $\boldsymbol{y}$ located between Γ and $x_3 = 0$,

$$u^i(\boldsymbol{x}) = \mathrm{G}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{4\pi}\left[\frac{\exp(\mathrm{i}\kappa|\boldsymbol{x} - \boldsymbol{y}|)}{|\boldsymbol{x} - \boldsymbol{y}|} - \frac{\exp(\mathrm{i}\kappa|\boldsymbol{x} - \widehat{\boldsymbol{y}}|)}{|\boldsymbol{x} - \widehat{\boldsymbol{y}}|}\right], \quad \boldsymbol{x} \in \mathbb{R}^3_+, \quad \boldsymbol{x} \ne \boldsymbol{y}.$$

As indicated we assume that $\boldsymbol{y} = (y_1, y_2, y_3)^\top$ satisfies $0 < y_3 < \xi(y_1, y_2)$, and $\widehat{\boldsymbol{y}} = (y_1, y_2, -y_3)^\top$ denotes the reflected point source. The reason for using this Green's function instead of the standard fundamental solution is its faster decay rate in vertically bounded strips. It is known that $u^i \in H^1_r(\Omega_H)$ for $r < 1$ [16]. As we are considering a radiation problem, the "scattered field" u^s satisfies $u^s = -u^i$ in Ω. Hence, we are able to compute explicitly the numerical approximation error in the scattered field u^s_τ obtained by Equation (11) for the vanishing total field in the bounded cell $\Omega^{2\pi}_H$.

We fix $H = 2$, and assume that Γ is given by the bi-periodic function

$$\xi(\tilde{\boldsymbol{x}}) = 0.6 + 0.3\sin(x_1)\cos(2x_2) + 0.2\sin(2x_1)\sin(3x_2), \ \tilde{\boldsymbol{x}} = (x_1, x_2) \in \mathbb{R}^2.$$

Moreover, we consider the point source $\boldsymbol{y} = (0, 0, 0.1)^\top$.

TABLE 1 Relative error and computational order with respect to τ by $N = 3$, $P = 2$.

	$\kappa = 0.4$		$\kappa = 1.4$		$\kappa = 3$	
τ	**Error**	$\mathrm{C_{order}}$	**Error**	$\mathrm{C_{order}}$	**Error**	$\mathrm{C_{order}}$
0.78	3.3438×10^{-2}	—	3.7156×10^{-2}	—	1.9390×10^{-1}	—
0.41	1.0870×10^{-2}	1.75	1.0788×10^{-2}	1.92	5.9628×10^{-2}	1.83
0.21	3.0854×10^{-3}	1.88	2.8671×10^{-3}	1.98	1.5824×10^{-2}	1.98
0.10	8.1722×10^{-4}	1.79	7.3826×10^{-4}	1.83	4.0295×10^{-3}	1.84

TABLE 2 Relative error with respect to P and N for wave number $\kappa = 0.4$.

	$\tau = 0.78$		$\tau = 0.21$	
P	$N = 2$	$N = 3$	$N = 2$	$N = 3$
2	3.3658×10^{-2}	3.3438×10^{-2}	3.4548×10^{-3}	3.0854×10^{-3}
3	3.3658×10^{-2}	3.3438×10^{-2}	3.4548×10^{-3}	3.0854×10^{-3}
4	3.3658×10^{-2}	3.3438×10^{-2}	3.4548×10^{-3}	3.0854×10^{-3}

To solve Equation (11) in $\mathrm{V}^* \times \Omega_2^{2\pi}$, we first generate an adapted square mesh in V^* by using Algorithm 1 and tetrahedral meshes in $\Omega_2^{2\pi}$ with $(\mathrm{M}+1)^2 \times (\mathrm{M}/2+1)$ nodes for $\mathrm{M} \in \{16, 32, 64, 128\}$ so that the maximum diameter τ for these four generated meshes is 0.78,0.41, 0.21 and 0.1, respectively. Note that these values for τ are smaller than the essential limit of one-tenth of the wavelength for each value of κ considered below. For each $\boldsymbol{\alpha} \in \mathrm{V}^*$, we approximate the solution $w(\boldsymbol{\alpha}, .)$ of (12) by $\mathtt{P1}-$conforming piecewise linear finite elements. The Floquet–Bloch transform of the incident field for each $\boldsymbol{\alpha} \in \mathrm{V}^*$ is computed by [16]

$$\mathcal{J}u^i(\boldsymbol{\alpha}, \boldsymbol{x}) = \sum_{\boldsymbol{j} \in \mathbb{Z}^2} e^{\mathrm{i}(\boldsymbol{\alpha}-\boldsymbol{j})\cdot(\tilde{\boldsymbol{x}}-\tilde{\boldsymbol{y}})} \begin{cases} e^{\mathrm{i}\sqrt{\kappa^2-|\boldsymbol{\alpha}-\boldsymbol{j}|^2}x_3}\, \mathrm{sinc}(\sqrt{\kappa^2-|\boldsymbol{\alpha}-\boldsymbol{j}|^2}y_3)\, y_3, & y_3 < x_3, \\ e^{\mathrm{i}\sqrt{\kappa^2-|\boldsymbol{\alpha}-\boldsymbol{j}|^2}y_3}\, \mathrm{sinc}(\sqrt{\kappa^2-|\boldsymbol{\alpha}-\boldsymbol{j}|^2}x_3)\, x_3, & \text{otherwise.} \end{cases}$$

The right-hand side of (11) is obtained the normal derivative and the DtN map of $\mathcal{J}u^i(\boldsymbol{\alpha}, \boldsymbol{x})$. Thus, the formula for $\mathcal{J}u^i$ above in particular shows that the assumptions of Theorem 20 are satisfied. The right hand side can be evaluated by truncating the infinite series if $|j_1|$ and $|j_2| > 40$. Eventually, we solve a sparse linear system for each $\boldsymbol{\alpha}$ by the GMRES iterative method with tolerance 1×10^{-6}.

Below, we will demonstrate the dependence of the numerical errors on the discretization parameters τ, N and P. In Table 1, the relative errors and the computational orders, which are computed the following formula

$$\mathrm{Error} = \frac{\|u^s - u^s_\tau\|_{L^2(\Omega_{\mathrm{H}}^{2\pi})}}{\|u^s\|_{L^2(\Omega_{\mathrm{H}}^{2\pi})}}, \qquad \mathrm{C_{order}} = \frac{\log(\mathrm{E}_1/\mathrm{E}_2)}{\log(\tau_1/\tau_2)},$$

are listed for different values of the finite element discretization parameter τ and wave number κ. This table indicates that the numerical results are consistent with the analytic results of Theorem 20 for each κ since the errors converge as τ decreases even with a low number of N and P. Note that for large values of the wave number κ, the structure of the singular curves becomes more complicated. For example for $\kappa = 3$ there are 20 curves of singular points in the domain V^*. Despite the complicated structure of the singular curves in α-space, the accurate results can still be obtained by using small values of N and P, only refining the spatial mesh τ, as reported in Table 1.

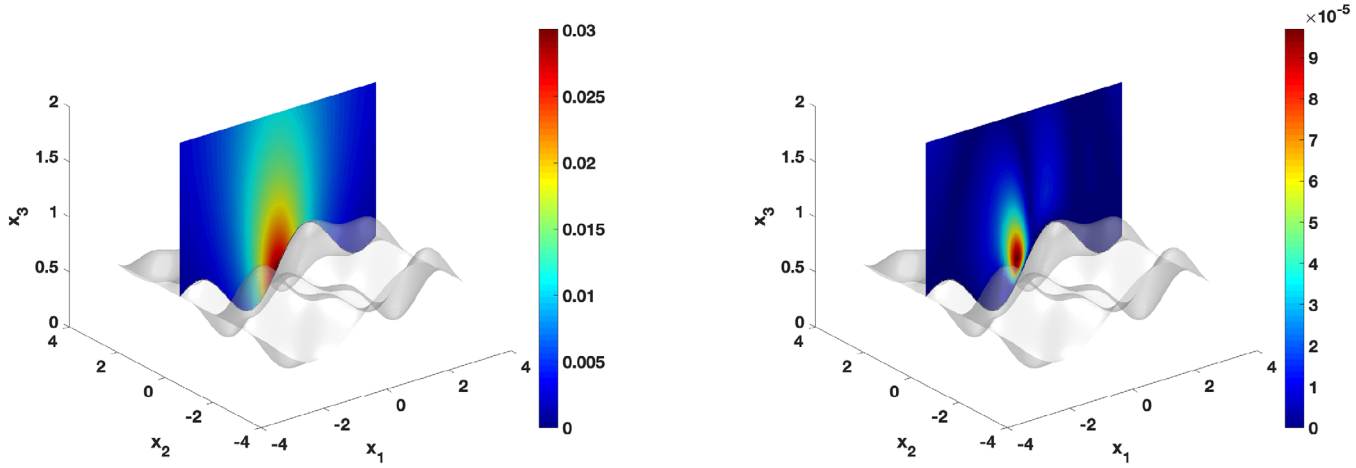
In Tables 2 and 3, we report the relative errors with respect to N and P for different values of τ. Since the error of the finite element method is dominated in the computational order, we can not see the exponential convergence of the proposed numerical integration method with respect to N and P.

In Figure 7, we show the numerical scattered field and its numerical error in L^2-norm for $\kappa = 1$ with the parameter $\tau = 0.21$, $N = 3$ and $P = 2$.

In conclusion, our method provides a way to very accurately approximate the inverse Floquet–Bloch transform for solutions to a non-periodic scattering problem. Even for very small values of P, the error from this approximation is already dominated by the error from the finite element method. Nevertheless, for larger wave numbers, the structure of the sin-

TABLE 3 Relative error with respect to N and τ for $\kappa = 1$, $P = 2$.

N	$\tau = 0.78$	$\tau = 0.41$	$\tau = 0.21$
2	3.4106×10^{-2}	1.1145×10^{-2}	3.5580×10^{-3}
3	3.4054×10^{-2}	1.1137×10^{-2}	3.2018×10^{-3}
4	3.3979×10^{-2}	1.1078×10^{-2}	3.1413×10^{-3}
5	3.3976×10^{-2}	1.1078×10^{-2}	3.1428×10^{-3}



(a) $|u_\tau^s|$ (b) $|u^s - u_\tau^s|$

FIGURE 7 Graphs of the numerical scattered field and its absolute error for $\kappa = 1$ in $\Omega_H^{2\pi}$.

gular curves quickly becomes quite complicated, making it necessary to use a large number of quadrature points. Thus, the accurate solution of non-periodic scattering problems in periodic domains remains a computational challenge.

ACKNOWLEDGMENTS

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 433126998 – SFB 1173. The work of N. Shafieeabyaneh and R. Zhang was also supported by DFG Grant Project-ID 258734477.

Open access funding enabled and organized by Projekt DEAL.

ORCID

Nasim Shafieeabyaneh https://orcid.org/0000-0002-2332-8733
Ruming Zhang https://orcid.org/0000-0003-2336-1020

REFERENCES

[1] Arens, T.: Scattering by biperiodic layered media: The integral equation approach. Habilitation (2010). https://doi.org/10.5445/IR/1000016241
[2] Johnson, S.G., Joannopoulos, J.D.: Photonic Crystals: The Road from Theory to Practice. Kluwer Academic, Boston (2002)
[3] Kirsch, A.: Diffraction by periodic structures. Inverse Problems in Mathematical Physics, pp. 87–102. Springer, Heidelberg (1993)
[4] Schmidt, G.: On the diffraction by biperiodic anisotropic structures. Appl. Anal. 82, 75–92 (2003). https://doi.org/10.1080/0003681031000068275
[5] Bao, G.: Finite element approximation of time harmonic waves in periodic structures. SIAM J. Numer. Anal. 32, 1155–1169 (1995). https://doi.org/10.1137/0732053
[6] Hsiao, G.C., Nigam, N., Pasciak, J.E., Liwei, X.: Error analysis of the DTN-FEM for the scattering problem in acoustics via fourier analysis. J. Comput. Appl. Math. 235, 4949–4965 (2011). https://doi.org/10.1016/j.cam.2011.04.020
[7] Nedelec, J.C., Starling, F.: Integral equation methods in a quasi-periodic diffraction problem for the time-harmonic maxwell's equations. SIAM J. Numer. Anal. 22, 1679–1701 (1991). https://doi.org/10.1137/0522104
[8] Coatléven, J.: Helmholtz equation in periodic media with a line defect. J. Comput. Phys. 1675–1704 (2012). https://doi.org/10.1016/j.jcp.2011.10.022

[9] Haddar, H., Nguyen, T.: A volume integral method for solving scattering problems from locally perturbed infinite periodic layers. Appl. Anal. 96, 130–158 (2017). https://doi.org/10.1080/00036811.2016.1221942
[10] Lechleiter, A., Zhang, R.: A convergent numerical scheme for scattering of aperiodic waves from periodic surfaces based on the Floquet–Bloch transform. SIAM J. Numer. Anal. 55, 713–736 (2017). https://doi.org/10.1137/16M1067524
[11] Lechleiter, A., Zhang, R.: A Floquet–Bloch transform based numerical method for scattering from locally perturbed periodic surfaces. SIAM J. Sci. Comput. 39, B819–B839 (2017). https://doi.org/10.1137/16M1104111
[12] Lechleiter, A.: The Floquet–Bloch transform and scattering from locally perturbed periodic surfaces. J. Math. Anal. Appl. 446, 605–627 (2017). https://doi.org/10.1016/j.jmaa.2016.08.055
[13] Lechleiter, A., Nguyen, D.L.: Scattering of Herglotz waves from periodic structures and mapping properties of the bloch transform. Proc. R. Soc. Edinb. A: Math. 145, 1283–1311 (2015). https://doi.org/10.1017/S0308210515000335
[14] Konschin, A.: Direkte und inverse elektromagnetische Streuprobleme fur lokal gestörte periodische Medien. Theses. Universität Bremen (2019). urn:nbn:de:gbv:46-00107835-13
[15] Konschin, A.: Electromagnetic wave scattering from locally perturbed periodic inhomogeneous layers. Math. Methods Appl. Sci. 44(18), 14126–14147 (2021). https://doi.org/10.1002/mma.7680
[16] Lechleiter, A., Zhang, R.: Non-periodic acoustic and electromagnetic, scattering from periodic structures in 3D. Comput. Math. Appl. 74, 2723–2738 (2017). https://doi.org/10.1016/j.camwa.2017.08.042
[17] Ehrhardt, M., Han, H., Zheng, C.: Numerical simulation of waves in periodic structures. Commun. Comput. Phys. 5, 849–870 (2009). http://global-sci.org/intro/article_detail/cicp/7767.html
[18] Ehrhardt, M., San, J., Zheng, C.: Evaluation of scattering operators for semi-infinite periodic arrays. Commun. Comput. Phys. 7, 347–364 (2009)
[19] Fliss, S.: Analyse mathématique et numérique de problèmes de propagation des ondes dans des milieux périodiques infinis localement perturbés. Ph.D. thesis. Ecole Polytechnique (2009). https://pastel.hal.science/pastel-00005464
[20] Fliss, S., Joly, P.: Wave propagation in locally perturbed periodic media (case with absorption): Numerical aspects. J. Comput. Phys. 231, 1244–1271 (2012). https://doi.org/10.1016/j.jcp.2011.10.007
[21] Joly, P., Li, J.R., Fliss, S.: Exact boundary conditions for periodic waveguides containing a local perturbation. Commun. Comput. Phys. 1, 945–973 (2006). http://global-sci.org/intro/article_detail/cicp/7989.html
[22] Zhang, R.: A high order numerical method for scattering from locally perturbed periodic surfaces. SIAM J. Sci. Comput. 40, A2286–A2314 (2018). https://doi.org/10.1137/17M1144945
[23] Arens, T., Zhang, R.: A nonuniform mesh method for wave scattered by periodic surfaces. In: Preprinted (2022). https://doi.org/10.48550/arXiv.2203.05792
[24] Chandler-Wilde, S.N., Elschner, J.: Variational approach in weighted Sobolev spaces to scattering by unbounded rough surfaces. SIAM J. Math. Anal. 42, 2554–2580 (2010). https://doi.org/10.1137/090776111
[25] Chandler-Wilde, S.N., Monk, P.: Existence, uniqueness, and variational methods for scattering by unbounded rough surfaces. SIAM J. Math. Anal. 37, 598–618 (2005). https://doi.org/10.1137/040615523
[26] Elschner, J., Schmidt, G.: Diffraction in periodic structures and optimal design of binary gratings. part I: direct problems and gradient formulas. Math. Methods Appl. Sci. 21, 1297–1342 (1998). https://doi.org/10.1002/(SICI)1099-1476(19980925)21:14⟨1297::AID-MMA997⟩3.0.CO;2-C
[27] Hörmander, L.: An Introduction to Complex Analysis in Several Variables. Amsterdam: North-Holland Pub. Co; New York; American Elsevier Pub. Co (1979)
[28] Kress, R.: Numerical Analysis. Springer, Berlin (1998)

How to cite this article: Arens, T., Shafieeabyaneh, N., Zhang, R.: A high-order numerical method for solving non-periodic scattering problems in three-dimensional bi-periodic structures. Z Angew Math Mech. e202300650 (2024). https://doi.org/10.1002/zamm.202300650

APPENDIX A: ESTIMATES FOR DERIVATIVES OF SQUARE ROOT FUNCTIONS

Lemma A1. *Let $s \in \mathbb{C}$, $\alpha \in \mathbb{R}$ such that $\alpha \neq s$. Then, for any $\ell \in \mathbb{N}$,*

$$\left| \frac{\mathrm{d}^{\ell}}{\mathrm{d}\alpha^{\ell}} \sqrt{s \pm \alpha} \right| \leq \ell! \left| s \pm \alpha \right|^{1/2-\ell}.$$

Proof. For any $\ell \geq 0$, a direct calculation yields

$$\left| \frac{\mathrm{d}^{\ell}}{\mathrm{d}\alpha^{\ell}} \sqrt{s \pm \alpha} \right| = \frac{|(2\ell-3)!!|}{2^{\ell}} \left| s \pm \alpha \right|^{1/2-\ell} \leq \frac{(2\ell)!!}{2^{\ell}} \left| s \pm \alpha \right|^{1/2-\ell} = \ell! \left| s \pm \alpha \right|^{1/2-\ell}.$$

□

Lemma A2. *Let $\nu \in \{1,2\}$. For any fixed $\ell \in \mathbb{N}$, there is a constant C such that*

$$\left|\frac{\partial^\ell \sqrt{\kappa^2 - |\boldsymbol{\alpha}|^2}}{\partial \alpha_\nu^\ell}\right| \leq \frac{C\,\ell!\,|\kappa + |\boldsymbol{\alpha}||^{1/2}}{|\kappa - |\boldsymbol{\alpha}||^{\ell-1/2}},$$

for all $\boldsymbol{\alpha} \in \mathbb{R}^2$ such that $|\boldsymbol{\alpha}| \neq \kappa$.

Proof. Without loss of generality, we treat the case $\nu = 1$. Consider $\sqrt{\kappa^2 - |\boldsymbol{\alpha}|^2} = \sqrt{s^2 - \alpha_1^2}$ where $s = \sqrt{\kappa^2 - \alpha_2^2}$. Using the Leibniz formula and Lemma A1 leads to

$$\begin{aligned}\left|\frac{\partial^\ell \sqrt{s^2 - \alpha_1^2}}{\partial \alpha_1^\ell}\right| &\leq \sum_{n=0}^{\ell} \binom{\ell}{n} \left|\frac{\partial^n \sqrt{s + \alpha_1}}{\partial \alpha_1^n}\right| \left|\frac{\partial^{\ell-n} \sqrt{s - \alpha_1}}{\partial \alpha_1^{\ell-n}}\right| \\ &\leq \sum_{n=0}^{\ell} \binom{\ell}{n} n!\,(\ell - n)!\,|s + \alpha_1|^{1/2-n} |s - \alpha_1|^{1/2-\ell+n} \\ &\leq \frac{C\,\ell! \sqrt{|s^2 - \alpha_1^2|}}{\left(\min\left\{|s + \alpha_1|, |s - \alpha_1|\right\}\right)^\ell}.\end{aligned}$$

Now, it remains to estimate $\min\left\{|s + \alpha_1|, |s - \alpha_1|\right\}$, and we can distinguish two cases as follows:

1. If $|\alpha_2| \geq \kappa$, then $s = \mathrm{i}\sqrt{\alpha_2^2 - \kappa^2}$. Hence,

$$|s + \alpha_1| = |s - \alpha_1| = \sqrt{\alpha_2^2 - \kappa^2 + \alpha_1^2} = \sqrt{|\kappa^2 - |\boldsymbol{\alpha}|^2|} \geq |\kappa - |\boldsymbol{\alpha}||.$$

2. If $|\alpha_2| < \kappa$, then $s = \sqrt{\kappa^2 - \alpha_2^2} > 0$. In this case, we write

$$\min\{|s + \alpha_1|, |s - \alpha_1|\} = |s - |\alpha_1|| = \frac{|\kappa^2 - |\boldsymbol{\alpha}|^2|}{\sqrt{\kappa^2 - \alpha_2^2} + |\alpha_1|}.$$

We conclude that

$$\min\{|s + \alpha_1|, |s - \alpha_1|\} \geq \frac{|\kappa^2 - |\boldsymbol{\alpha}|^2|}{\kappa + |\boldsymbol{\alpha}|} = |\kappa - |\boldsymbol{\alpha}||.$$

In both cases, we find by substituting $s^2 = \kappa^2 - \alpha_2^2$ in the estimate found above that

$$\left|\frac{\partial^\ell \sqrt{\kappa^2 - |\boldsymbol{\alpha}|^2}}{\partial \alpha_1^\ell}\right| \leq \frac{C\,\ell!\,|\kappa + |\boldsymbol{\alpha}||^{1/2}}{|\kappa - |\boldsymbol{\alpha}||^{\ell-1/2}}.$$

□