

# A nested hybridizable discontinuous Galerkin method for computing second-harmonic generation in three-dimensional metallic nanostructures

F. Vidal-Codina<sup>a,\*</sup>, N.-C. Nguyen<sup>a</sup>, C. Ciraci<sup>b</sup>, S.-H. Oh<sup>c</sup>, J. Peraire<sup>a</sup>

<sup>a</sup>*Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

<sup>b</sup>*Center for Biomolecular Nanotechnologies, Istituto Italiano di Tecnologia, Via Barsanti 14, 73010 Arnesano (LE), Italy*

<sup>c</sup>*Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA*

---

## Abstract

We develop a nested hybridizable discontinuous Galerkin (HDG) method to numerically solve the Maxwell's equations coupled with a hydrodynamic model for the conduction-band electrons in metals. The HDG method leverages static condensation to eliminate the degrees of freedom of the approximate solution defined in the elements, yielding a linear system in terms of the degrees of freedom of the approximate trace defined on the element boundaries. This article presents a computational method that relies on a degree-of-freedom reordering such that the HDG linear system accommodates an additional static condensation step to eliminate a large portion of the degrees of freedom of the approximate trace, thereby yielding a much smaller linear system. For the particular metallic structures considered in this article, the resulting linear system obtained by means of nested static condensations is a block tridiagonal system, which can be solved efficiently. We apply the nested HDG method to compute second harmonic generation on a triangular coaxial periodic nanogap structure. This nonlinear optics phenomenon features rapid field variations and extreme boundary-layer structures that span a wide range of length scales. Numerical results show that the ability to identify structures which exhibit resonances at  $\omega$  and  $2\omega$  is essential to excite the second harmonic response.

*Keywords:* Hybridizable discontinuous Galerkin method, Maxwell's equations, hydrodynamic model for metals, nonlinear plasmonics, nonlocal electrodynamics, second-harmonic generation

---

\*Corresponding author

*Email addresses:* [fvidal@mit.edu](mailto:fvidal@mit.edu) (F. Vidal-Codina), [cuongng@mit.edu](mailto:cuongng@mit.edu) (N.-C. Nguyen), [ciraci@iit.edu](mailto:ciraci@iit.edu) (C. Ciraci), [sang@umn.edu](mailto:sang@umn.edu) (S.-H. Oh), [peraire@mit.edu](mailto:peraire@mit.edu) (J. Peraire)

## 1. Introduction

Nonlinear plasmonics [11, 13, 42, 66, 74] studies the effects that arise when nonlinear media in a plasmonic structure cause the polarization to depend nonlinearly on the electric field. These effects are inherently weak, but can be amplified by the collective excitation of conduction-band electrons that occur on metallic nanostructures, commonly known as plasmon resonances. These excitations produce strong near-field enhancements of the incident wave by confining light in small volumes whose critical length scales are several orders of magnitude smaller than the wavelength of light. The combination of nonlinear effects and localized plasmon resonances offers new opportunities for the generation and manipulation of light at the nanoscale. The most common nonlinear effect is second-harmonic generation (SHG), whereby two photons at the incident frequency interact to generate a single photon at twice the incident frequency. Within classical electromagnetics, SHG occurs when the polarization has a quadratic dependence on the electric field. In common nonlinear optical materials, this quadratic dependence is provided by the absence of an inversion symmetry in the crystalline lattice, which generates an asymmetric response with respect to the applied electric field orientation. In contrast, plasmonic metals (e.g. gold, silver, copper and aluminum) are centrosymmetric and do not possess an intrinsic second-order susceptibility. A correct description of metal nonlinearities has to account for the complex dynamics of free-electrons [71]. It is well-known in fact that free-electrons provide both bulk and surface mechanisms for SHG [38, 73, 85]. Surface contributions arise from locally broken centrosymmetry while bulk contributions are ascribed to convective and Lorentz-force interactions in the electron fluid [21, 71]. Contrarily to nonlinear crystals, the nonlinear response of metals becomes highly dependent on the device geometry [14, 44]. In fact, unless centrosymmetry is broken at the larger scale of the device structure, the second-harmonic fields stays highly localized and destructively interferes in the far-field, becoming too weak to be observed. Plenty of research has been devoted to designing and fabricating shapes that enhance SHG, see [13, 42, 46, 66] and the references therein.

The ability to accurately model and simulate nonlinear plasmonic phenomena requires computational capabilities that challenge traditional simulation techniques. From a modelling perspective, plasmonic phenomena can be described by Maxwell's equations coupled with a hydrodynamic model to account for the nonlocal effect of the conduction-band electrons in the metallic materials, which become relevant for sub-10 nanometer features. The problems of interest involve the interaction of long-wavelength electromagnetic waves ( $\mu\text{m}$  and  $\text{mm}$ ) with nanometer-wide features for potential applications in sensing and spectroscopy. Moreover, the electromagnetic fields are confined in deep-subwavelength cavities, and Ångstrom-thin accumulation charge layers develop at the metal-dielectric interface. As a consequence, the discretizations required to attain accurate simulations need to be highly adapted (to properly capture the extremely localized fields) and anisotropic (to be computationally tractable).

The finite-difference time-domain (FDTD) algorithm [48, 76] is a well known computational method for wave propagation. The most popular FDTD method utilizes Yee's scheme [86] to discretize space and time with staggered cartesian grids and second-order schemes. The main shortcoming of FDTD is modeling geometries with complex features, since the stair-casing at the interfaces not aligned with the Cartesian grid severely impacts the accuracy. In addition, the mismatch in length scales that is characteristic of nonlinear plasmonics is a severe hindrance, since resolving for the smallest phenomena while keeping a uniform discretization may require grids having prohibitively large number of degrees of freedom. The FDTD method has been used for nonlocal [31, 57] and second-harmonic generation [2, 16, 54, 90] simulations.

Finite element (FE) methods [40] have also been widely used in electromagnetics, due to their ability to handle heterogeneous media and intricate geometries with unstructured discretizations, as well as  $h/p$  adaptation for increased accuracy. The family of face/edge elements introduced by Nédélec [60] have been extensively used to simulate electromagnetic wave propagation, and have been shown to avoid the problem of spurious modes [10] by appropriately choosing the approximation spaces. A commonly used implementation of edge elements for Maxwell’s equations is the one provided by the RF Module of COMSOL Multiphysics [26]. Many groups have used this platform to implement their own version of the hydrodynamic model [19, 77, 78], SHG on nanoparticles and nanoantennas [4, 6, 15, 32, 91] as well as SHG on periodic arrays of nanostructures [17, 21, 45, 69].

Discontinuous Galerkin (DG) methods [25, 34] have also been widely used to simulate nonlocal effects in electromagnetics. In DG methods, the domain is discretized onto a collection of disjoint elements, and each field component is independently approximated within each element using standard finite element spaces. Solutions are therefore discontinuous across elements, and flux continuity is enforced at the interfaces. The DG method in time domain has been developed for many nanophotonics applications on metallic and lossless media [12, 39, 49, 55, 64, 72], as well as to simulate SHG on nanoparticles [35, 47, 58], resonators and antennas [1, 33, 52, 84]. The main caveat of DG methods for 3-D applications in the frequency domain or in the time domain with implicit time integration is the high computational burden, stemming from the duplication of degrees of freedom at the interfaces. This shortcoming is circumvented by the hybridizable discontinuous Galerkin (HDG) method, first developed in [22–24] and later extended to acoustics and elastodynamics [62, 70] as well as time-harmonic Maxwell’s equations [51, 63] and the hydrodynamic model for metals [50, 81]. In addition, unlike other DG methods HDG exhibits optimal convergence rates for both the solution and the flux. As a result, the solution may be locally post-processed to gain an additional order of convergence, a phenomenon known as superconvergence. In the recent years, the HDG method has been successfully applied for 2-D and 3-D metallic nanostructures to simulate plasmonic phenomena [50, 67, 80–82, 88, 89].

This article presents a nested hybridizable discontinuous Galerkin (nHDG) method for the Maxwell’s equations coupled with a hydrodynamic model for the conduction-band electrons in metals. By means of a static condensation to eliminate the degrees of freedom of the approximate solution defined within the elements, the HDG method yields a linear system in terms of the degrees of freedom of the approximate trace defined on the element boundaries. The nested HDG is a computational method that builds on top of classical HDG, whereby an additional static condensation is performed in order to eliminate a large portion of the degrees of freedom of the HDG linear system. Consequently, this nested strategy gives rise to a much smaller linear system, encompassing the degrees of freedom of a reduced number of element boundaries. Furthermore, if these element boundaries are judiciously selected, the nested HDG yields a linear system that is block-tridiagonal, and can thus be solved efficiently. [The article also presents the formulation and implementation of the HDG method to compute SHG under the assumption of non-depleted pump approximation, that is, the fundamental wave is not affected by the generated harmonic \[11\].](#) We apply the nested HDG method to compute the SHG on a triangular coaxial periodic nanogap structure, which is a computationally intensive task since nonlinear optics phenomena feature rapid field variations and extreme boundary-layer structures that span a wide range of length scales. In addition, we propose strategies to partition the mesh in order to achieve an efficient nested static condensation.

This article is organized as follows. In Section 2, we introduce the equations and notation used throughout the article. In Section 3, we review the formulation and implementation of the HDG

method to solve the hydrodynamic model for metals in frequency domain, describe the modifications needed to simulate SHG and present the nested hybridization strategy and algorithm. In Section 4, we present numerical results to assess the performance of the proposed method and present some concluding remarks in Section 5.

## 2. Second harmonic generation in metallic nanostructures

We now derive the equations that will be used throughout this article. The overline  $\bar{\phantom{x}}$  denotes dimensional variables and constants, whereas the quantities without overlines are their non-dimensional counterparts. The only dimensional quantities without overlines are the reference quantities that we use to non-dimensionalize the problem:  $L_c$  is a reference length scale in meters,  $\alpha$  is a reference magnetic field in Ampere/meter,  $\varepsilon_0$  is the free-space permittivity in Farad/meter,  $\mu_0$  is the free-space permeability in Henry/meter,  $c_0 = 1/\sqrt{\varepsilon_0\mu_0}$  is the free-space speed of light in meters/second and  $Z_0 = \sqrt{\mu_0/\varepsilon_0}$  is the free-space impedance in Volt/Ampere.

### 2.1. Maxwell's equations in time domain

The electric  $\bar{\mathcal{E}}(\bar{\mathbf{x}}, \bar{t})$  and magnetic  $\bar{\mathcal{H}}(\bar{\mathbf{x}}, \bar{t})$  fields, along with the electric displacement  $\bar{\mathcal{D}}$  and magnetic flux density  $\bar{\mathcal{B}}$ , satisfy Maxwell's equations in a metallic domain  $\Omega^m$

$$\begin{aligned}\bar{\nabla} \times \bar{\mathcal{E}} + \partial_{\bar{t}} \bar{\mathcal{B}} &= 0 \quad (\text{Ampère's law}), \\ \bar{\nabla} \times \bar{\mathcal{H}} - \partial_{\bar{t}} \bar{\mathcal{D}} &= \bar{\mathcal{J}}_{\text{ext}} \quad (\text{Faraday's law}), \\ \bar{\nabla} \cdot \bar{\mathcal{D}} &= \bar{\rho}_{\text{ext}}, \quad (\text{Gauss's law}), \\ \bar{\nabla} \cdot \bar{\mathcal{B}} &= 0, \quad (\text{magnetic Gauss's law}),\end{aligned}\tag{1}$$

where  $\bar{\mathcal{J}}_{\text{ext}}$  represents the external electric current and  $\bar{\rho}_{\text{ext}}$  the external volume charge density. For simplicity of exposition, we assume there are no external current  $\bar{\mathcal{J}}_{\text{ext}} = 0$  and external charge density  $\bar{\rho}_{\text{ext}} = 0$ . In addition, we have the following constitutive relations

$$\begin{aligned}\bar{\mathcal{B}} &= \bar{\mu} \bar{\mathcal{H}}, \\ \bar{\mathcal{D}} &= \varepsilon_0 \bar{\mathcal{E}} + \bar{\mathcal{P}} + \bar{\mathcal{P}}_{\infty} = \bar{\varepsilon}_{\infty} \bar{\mathcal{E}} + \bar{\mathcal{P}}, \\ \bar{\nabla} \cdot \bar{\mathcal{P}} &= -\bar{\rho} \\ \partial_{\bar{t}} \bar{\mathcal{P}} &= \bar{\mathcal{J}}.\end{aligned}\tag{2}$$

The polarization density  $\bar{\mathcal{P}}$  represents the density of permanent or induced electric dipole moments due to free electrons. Conversely, the background polarization  $\bar{\mathcal{P}}_{\infty} = (\bar{\varepsilon}_{\infty} - \varepsilon_0) \bar{\mathcal{E}}$  represents the polarization of the bound electrons in the valence band. The last two relations relate the polarization density  $\bar{\mathcal{P}}$  to the internal current  $\bar{\mathcal{J}}$  and internal charge density  $\bar{\rho}$ .

### 2.2. A nonlinear hydrodynamic model

The above set of equations is closed once we specify the polarization density of the material in response to applied electromagnetic fields. In the simplest case,  $\bar{\mathcal{P}}$  depends locally on the electric field through a linear relationship. To account for nonlocal effects which become important at nanoametric scales, a hydrodynamic model (HM) for the free electron gas was proposed in [30]. This model, despite neglecting quantum phenomena such as quantum tunneling and quantum

oscillations, introduces a hydrodynamic pressure term that accounts for the nonlocal coupling of the conduction-band electrons. Below we provide a brief description of the model and refer to [8, 20, 68] for additional details.

The electron density  $\bar{n}(\bar{\mathbf{x}}, \bar{t})$  and the hydrodynamic velocity  $\bar{\mathbf{v}}(\bar{\mathbf{x}}, \bar{t})$  are related by the continuity equation as  $\partial_{\bar{t}}\bar{n} = -\bar{\nabla} \cdot (\bar{n}\bar{\mathbf{v}})$ . In addition, the equation of motion for the electron fluid under a macroscopic electromagnetic field is described as

$$\bar{m}_e(\partial_{\bar{t}} + \bar{\mathbf{v}} \cdot \bar{\nabla} + \bar{\gamma})\bar{\mathbf{v}} = e(\bar{\mathcal{E}} + \bar{\mathbf{v}} \times \bar{\mathcal{B}}) - \frac{\bar{\nabla}\bar{p}}{\bar{n}}, \quad (3)$$

where  $m_e$  is the effective electron mass,  $e$  is the electron charge so that  $\bar{\mathcal{J}} = e\bar{n}\bar{\mathbf{v}}$  and  $\bar{\gamma}$  is a damping constant related to the collision rate of the electrons. The electron pressure  $\bar{p}(\bar{\mathbf{x}}, \bar{t})$  is given by [27]

$$p(\bar{\mathbf{x}}, \bar{t}) = \bar{p}_0 \left( \frac{\bar{n}(\bar{\mathbf{x}}, \bar{t})}{\bar{n}_0} \right)^{5/3} \quad (4)$$

where  $\bar{p}_0 = \frac{18}{25}\bar{n}_0\bar{E}_F$ ,  $\bar{E}_F$  is the Fermi energy and  $\bar{n}_0$  is the equilibrium charge density.

After combining equations (3), (4) and the continuity equation, we follow [21] and retaining only first- and second-order terms, we obtain a nonlinear nonlocal model for the polarization in response to an applied electromagnetic field

$$\begin{aligned} -\bar{\beta}^2\bar{\nabla}(\bar{\nabla} \cdot \bar{\mathcal{P}}) + \partial_{\bar{t}\bar{t}}\bar{\mathcal{P}} + \bar{\gamma}\partial_{\bar{t}}\bar{\mathcal{P}} - \bar{\omega}_p^2\varepsilon_0\bar{\mathcal{E}} = & -\frac{\bar{\omega}_p^2\varepsilon_0}{\bar{n}_0\bar{e}}\bar{\mathcal{E}}(\bar{\nabla} \cdot \bar{\mathcal{P}}) + \frac{\bar{\omega}_p^2}{c_0^2\bar{n}_0\bar{e}}\partial_{\bar{t}}\bar{\mathcal{P}} \times \bar{\mathcal{H}} \\ & - \frac{1}{\bar{n}_0\bar{e}}\bar{\nabla} \cdot (\partial_{\bar{t}}\bar{\mathcal{P}} \otimes \partial_{\bar{t}}\bar{\mathcal{P}}) - \frac{\bar{\beta}^2}{3\bar{n}_0\bar{e}}\bar{\nabla}(\bar{\nabla} \cdot \bar{\mathcal{P}})^2, \end{aligned} \quad (5)$$

where  $\bar{\omega}_p = \sqrt{\bar{n}_0\bar{e}^2/\bar{m}_e\varepsilon_0}$  is the metal's plasma frequency. For incoming fields above this frequency the metal behaves like a lossy dielectric since electron mobility is not sufficient to react and cancel the incoming wave. The nonlocal parameter  $\bar{\beta}$  is given by  $\bar{\beta} = \sqrt{6\bar{E}_F/5\bar{m}_e}$ . Given the definition of Fermi kinetic energy  $\bar{E}_F = \bar{m}_e\bar{v}_F^2/2$ , the nonlocal parameter reduces to  $\bar{\beta} = \sqrt{3/5}\bar{v}_F$  [53], where  $\bar{v}_F$  is the Fermi velocity. The nonlinear terms in the right hand side of equation (5) are known as Coulomb force, magnetic Lorentz force, nonlinear convective force and nonlinear pressure force. We note that the simplest form of the hydrodynamic model, based on the Thomas-Fermi approximation accounting only for the linearized kinetic energy, corresponds to neglecting the nonlinear terms in the right hand side of (5) to recover a linear expression. The implementation of this nonlocal linear model with a hybridizable discontinuous Galerkin method was presented in [81].

Equation (5) needs to be solved simultaneously with Maxwell's equations (1). Before proceeding, it is convenient to non-dimensionalize the problem variables using the following scalings

$$\begin{aligned} \bar{\mathbf{x}} = L_c\mathbf{x}, \quad \bar{t} = L_ct/c_0, \quad \bar{\mathcal{E}} = \alpha Z_0\mathcal{E}, \quad \bar{\mathcal{H}} = \alpha\mathcal{H}, \\ \bar{\mathcal{D}} = \varepsilon_0\alpha Z_0\mathcal{D}, \quad \bar{\mathcal{B}} = \mu_0\alpha\mathcal{B}, \quad \bar{\mathcal{J}} = \alpha\mathcal{J}/L_c, \quad \bar{\mathcal{P}} = \alpha\mathcal{P}/c_0. \end{aligned} \quad (6)$$

For a non-magnetic medium ( $\bar{\mu} = \mu_0$ ), applying the scalings above to Maxwell's equations (1), the

constitutive relations (2) and the hydrodynamic pressure equation (5), we obtain

$$\begin{aligned}
\nabla \times \mathcal{E} + \partial_t \mathcal{H} &= 0, \\
\nabla \times \mathcal{H} - \partial_t(\varepsilon_\infty \mathcal{E}) &= \mathcal{J}, \\
\beta^2 \nabla \rho + \partial_t \mathcal{J} + \gamma \mathcal{J} - \omega_p^2 \mathcal{E} &= \frac{\omega_p^2}{n_0 e} \mathcal{E} \rho + \frac{\omega_p^2}{n_0 e} \mathcal{J} \times \mathcal{H} - \frac{1}{n_0 e} \nabla \cdot (\mathcal{J} \otimes \mathcal{J}) - \frac{\beta^2}{3n_0 e} \nabla \rho^2,
\end{aligned} \tag{7}$$

with the non-dimensional optical constants  $\varepsilon_\infty = \bar{\varepsilon}_\infty / \varepsilon_0$ ,  $\omega_p = \bar{\omega}_p L_c / c_0$ ,  $\gamma = \bar{\gamma} L_c / c_0$ ,  $\beta = \bar{\beta} / c_0$  and electron constants  $e = \bar{e} c_0 / (\alpha L_c^2)$  and  $n_0 = \bar{n}_0 L_c^3$ .

We now assume a time-harmonic response of the non-dimensional electromagnetic fields, where  $*$  denotes the complex conjugate

$$\begin{aligned}
\mathcal{E}(\mathbf{x}, t) &= \sum_{n \in \mathbb{N}} \Re\{\mathbf{E}_n(\mathbf{x}) \exp(-in\omega t)\} = \frac{1}{2} \sum_{n \in \mathbb{N}} \mathbf{E}_n(\mathbf{x}) \exp(-in\omega t) + \mathbf{E}_n^*(\mathbf{x}) \exp(in\omega t), \\
\mathcal{H}(\mathbf{x}, t) &= \sum_{n \in \mathbb{N}} \Re\{\mathbf{H}_n(\mathbf{x}) \exp(-in\omega t)\} = \frac{1}{2} \sum_{n \in \mathbb{N}} \mathbf{H}_n(\mathbf{x}) \exp(-in\omega t) + \mathbf{H}_n^*(\mathbf{x}) \exp(in\omega t), \\
\mathcal{J}(\mathbf{x}, t) &= \sum_{n \in \mathbb{N}} \Re\{\mathbf{J}_n(\mathbf{x}) \exp(-in\omega t)\} = \frac{1}{2} \sum_{n \in \mathbb{N}} \mathbf{J}_n(\mathbf{x}) \exp(-in\omega t) + \mathbf{J}_n^*(\mathbf{x}) \exp(in\omega t), \\
\rho(\mathbf{x}, t) &= \sum_{n \in \mathbb{N}} \Re\{\rho_n(\mathbf{x}) \exp(-in\omega t)\} = \frac{1}{2} \sum_{n \in \mathbb{N}} \rho_n(\mathbf{x}) \exp(-in\omega t) + \rho_n^*(\mathbf{x}) \exp(in\omega t).
\end{aligned} \tag{8}$$

To derive a system for the amplitudes of the first harmonic, we introduce the expansions (8) into (7) considering only  $n \leq 2$ , multiply by  $\exp(-i\omega t)$  and invoke orthogonality of the Fourier modes to obtain following set of nonlinear equations

$$\begin{aligned}
\nabla \times \mathbf{E}_1 - i\omega \mathbf{H}_1 &= \mathbf{0}, \\
\nabla \times \mathbf{H}_1 + i\omega \varepsilon_\infty \mathbf{E}_1 - \mathbf{J}_1 &= \mathbf{0}, \\
\beta^2 \nabla \rho_1 + (\gamma - i\omega) \mathbf{J}_1 - \omega_p^2 \mathbf{E}_1 &= \mathbf{f}_1, \\
i\omega \rho_1 - \nabla \cdot \mathbf{J}_1 &= 0.
\end{aligned} \tag{9}$$

The nonlinear source term  $\mathbf{f}_1$  is given by

$$\begin{aligned}
2\mathbf{f}_1 &= \frac{\omega_p^2}{n_0 e} (\mathbf{E}_1^* \rho_2 + \mathbf{E}_2 \rho_1^*) + \frac{\omega_p^2}{n_0 e} (\mathbf{J}_1^* \times \mathbf{H}_2 + \mathbf{J}_2 \times \mathbf{H}_1^*) \\
&\quad - \frac{1}{n_0 e} \nabla \cdot (\mathbf{J}_1^* \otimes \mathbf{J}_2 + \mathbf{J}_2 \otimes \mathbf{J}_1^*) - \frac{\beta^2}{3n_0 e} \nabla (\rho_1^* \rho_2 + \rho_2 \rho_1^*).
\end{aligned} \tag{10}$$

Since we expect the power of the first harmonic to be several orders of magnitude stronger than that of the second harmonic ( $\|\mathbf{E}_2\| \ll \|\mathbf{E}_1\|$ ), we hereafter assume that  $\mathbf{f}_1 = \mathbf{0}$ . This approximation is known as the non-depleted pump approximation [11], which emphasizes that the second harmonic does not deplete the fundamental wave. This assumption simplifies the calculation of the second-harmonic and is justified in instances where the efficiency of the second harmonic generation is sufficiently weak. For later use, we refer to the system of equations (9) with the simplification  $\mathbf{f}_1 = \mathbf{0}$ , as  $\mathcal{L}^m(\mathbf{E}_1, \mathbf{H}_1, \mathbf{J}_1, \rho_1; \omega) = \mathbf{0}$ .

The Drude model [28] can be recovered from (9) by setting the nonlocal parameter  $\beta$  to zero, in which case Ohm's law is recovered  $\mathbf{J}_1 = i\omega_p^2 \mathbf{E}_1 / (\omega + i\gamma)$  and the complex Drude permittivity is expressed as  $\varepsilon_m(\omega) = \varepsilon_\infty - \omega_p^2 / (\omega(\omega + i\gamma))$ . The Drude model is more computationally efficient, since (9) simplifies to Maxwell's equations with a complex-valued permittivity, at the expense of neglecting the nonlocal electron interactions that become relevant for sub-10 nm features.

### 2.3. Metallic nanostructures

The above formulation is extended to consider the more general case of a metallic nanostructure, comprising both a metal  $\Omega^m$  described by the HM and a dielectric  $\Omega^d$  with permittivity  $\varepsilon_d$  described solely by Maxwell's equations. The solution within the metallic structure is governed by (9) and (12), whereas the response in the dielectric  $\Omega^d$  is given by regular time-harmonic Maxwell's equations, namely

$$\begin{aligned} \nabla \times \mathbf{E}_1 - i\omega \mathbf{H}_1 &= \mathbf{0}, \\ \nabla \times \mathbf{H}_1 + i\omega \varepsilon_d \mathbf{E}_1 &= \mathbf{0}. \end{aligned} \quad (11)$$

The boundary conditions for a metallic nanostructure can be expressed as

$$\begin{aligned} \mathbf{n} \times \mathbf{E}_1 \times \mathbf{n} &= \mathbf{0}, \quad \text{on } \partial\Omega_E, \\ \mathbf{n} \times \mathbf{H}_1 &= \mathbf{0}, \quad \text{on } \partial\Omega_H, \\ \mathbf{n} \cdot \mathbf{J}_1 &= 0, \quad \text{on } \partial\Omega_J^m \cup \Gamma_{\text{md}}, \\ \rho_1 &= 0, \quad \text{on } \partial\Omega_p^m, \\ \mathbf{H}_1 \times \mathbf{n} - \sqrt{\varepsilon_d} \mathbf{n} \times \mathbf{E}_1 \times \mathbf{n} &= \mathbf{H}_{\text{inc}} \times \mathbf{n} - \sqrt{\varepsilon_d} \mathbf{n} \times \mathbf{E}_{\text{inc}} \times \mathbf{n} := \mathbf{f}_{\text{inc}}, \quad \text{on } \partial\Omega_{\text{rad}}^d. \end{aligned} \quad (12)$$

The first and second boundary condition prescribe perfect electric conductor (PEC) and perfect magnetic conductor (PMC) behavior, which allows us to impose symmetries in periodic structures. The third and fourth conditions also prescribe symmetry conditions in periodic structures for the electric current and the electron charge, and are only applicable on the metallic subdomain. In addition, the third boundary condition is also applied at the metal-dielectric interface  $\Gamma_{\text{md}} = \Omega^m \cap \Omega^d$  to preclude the electrons from leaving the metal (since the normal component of the electric current vanishes), also known as no electron spill-out condition [7]. Quantum effects such as electron tunneling are therefore not modeled by the HM.

The last equation is the first-order Silver-Müller radiation condition [59, 75], preventing outgoing waves from reflecting at the computational boundary and coming back into the domain. A common alternative to the radiation condition are the perfectly matched layers (PMLs) [5, 41]. Even though PMLs can be more effective at absorbing waves, they are more computationally intensive and require parameter tuning. We have developed both alternatives and found no significant differences for the metallic nanostructures considered within the frequency regimes of interest, hence we resorted to Silver-Müller conditions. Illumination is prescribed as a  $\mathbf{p}$ -polarized plane wave propagating in the  $\mathbf{d}$ -direction, that is  $\mathbf{E}_{\text{inc}} = \mathbf{p} \exp(i\omega \sqrt{\varepsilon_d} \mathbf{d} \cdot \mathbf{x})$  and  $\mathbf{H}_{\text{inc}} = -i/\omega \nabla \times \mathbf{E}_{\text{inc}}$ . For an example of how the boundary conditions are assigned see Fig. 4 (c), where a periodic triangular coaxial nanostructure is shown.

We refer to the above system of equations (11) to be solved on the dielectric as  $\mathcal{L}^d(\mathbf{E}_1, \mathbf{H}_1; \omega) = \mathbf{0}$ , and the boundary equations (12) to be prescribed as  $b(\mathbf{E}_1, \mathbf{H}_1, \mathbf{J}_1, \rho_1; \omega) = \mathbf{f}_{\text{inc}}$ , respectively. In order to numerically solve the above systems with the HDG method we shall also impose continuity of the tangential component of the magnetic field along the entire domain and continuity of the

normal component of the electric current along the metal. These two additional conditions are explained and derived within the HDG discretization, see Section 3.

#### 2.4. Second-harmonic generation

Once the solution for the fundamental harmonic  $(\mathbf{E}_1, \mathbf{H}_1, \mathbf{J}_1, \rho_1)$  has been determined by simultaneously solving (9), (11) and (12), we turn our attention to the second harmonic. [Similarly as before](#), we introduce expansions (8) with  $n \leq 2$  into (7), multiply by  $\exp(-2i\omega t)$ , invoke orthogonality of the Fourier basis and obtain the following set of equations for the amplitudes of the second harmonic in the metallic domain. .

$$\begin{aligned} \nabla \times \mathbf{E}_2 - 2i\omega \mathbf{H}_2 &= \mathbf{0}, \\ \nabla \times \mathbf{H}_2 + 2i\omega \varepsilon_\infty \mathbf{E}_2 - \mathbf{J}_2 &= \mathbf{0}, \\ \beta^2 \nabla \rho_2 + (\gamma - 2i\omega) \mathbf{J}_2 - \omega_p^2 \mathbf{E}_2 &= \mathbf{f}_2 \\ 2i\omega \rho_2 - \nabla \cdot \mathbf{J}_2 &= 0, \end{aligned} \tag{13}$$

where

$$2\mathbf{f}_2 = \frac{\omega_p^2}{n_0 e} \mathbf{E}_1 \rho_1 + \frac{\omega_p^2}{n_0 e} \mathbf{J}_1 \times \mathbf{H}_1 - \frac{1}{n_0 e} \nabla \cdot (\mathbf{J}_1 \otimes \mathbf{J}_1) - \frac{\beta^2}{3n_0 e} \nabla \rho_1^2 \tag{14}$$

The nonlinear source term in (14) depends only on the fundamental fields and represents the motion of the electron fluid under an electromagnetic field.

[The assumption of non-depleted pump approximation simplifies second-harmonic calculations, since instead of solving a coupled system for the two harmonics, the SHG may be obtained by sequentially solving Maxwell's equations first for the fundamental wave and then for the second harmonic with a nonlinear source term involving only the fundamental fields \[21\].](#) Using the operator  $\mathcal{L}^m$  introduced earlier, we can express equation (13) as  $\mathcal{L}^m(\mathbf{E}_2, \mathbf{H}_2, \mathbf{J}_2, \rho_2; 2\omega) = \mathbf{f}_2$ . The boundary conditions for the SHG are then simply  $b(\mathbf{E}_2, \mathbf{H}_2, \mathbf{J}_2, \rho_2; 2\omega) = \mathbf{0}$ , since no incident light is shone and the only response is due to the nonlinear current.

Summarizing, the non-depleted SHG can be simulated as a two-step process, namely

1. Solve

$$\begin{aligned} \mathcal{L}^m(\mathbf{E}_1, \mathbf{H}_1, \mathbf{J}_1, \rho_1; \omega) &= \mathbf{0}, & \text{in } \Omega^m, \\ \mathcal{L}^d(\mathbf{E}_1, \mathbf{H}_1; \omega) &= \mathbf{0}, & \text{in } \Omega^d, \\ \text{with } b(\mathbf{E}_1, \mathbf{H}_1, \mathbf{J}_1, \rho_1; \omega) &= \mathbf{f}_{\text{inc}}, \end{aligned} \tag{15}$$

to compute  $(\mathbf{E}_1, \mathbf{H}_1, \mathbf{J}_1, \rho_1)$ .

2. Solve

$$\begin{aligned} \mathcal{L}^m(\mathbf{E}_2, \mathbf{H}_2, \mathbf{J}_2, \rho_2; 2\omega) &= \mathbf{f}_2(\mathbf{E}_1, \mathbf{H}_1, \mathbf{J}_1, \rho_1), & \text{in } \Omega^m, \\ \mathcal{L}^d(\mathbf{E}_2, \mathbf{H}_2; 2\omega) &= \mathbf{0}, & \text{in } \Omega^d, \\ \text{with } b(\mathbf{E}_2, \mathbf{H}_2, \mathbf{J}_2, \rho_2; 2\omega) &= \mathbf{0}, \end{aligned} \tag{16}$$

to compute  $(\mathbf{E}_2, \mathbf{H}_2, \mathbf{J}_2, \rho_2)$ .

Here, the plane-wave illumination is used only to compute the fundamental fields, ensuring the second harmonic fields are solely generated by the nonlinear source (14). [We point out that if the](#)



non-depleted approximation were not justified, the nonlinear source term (10) would need to be retained in (15) and in this case, a simple fixed-point iterative algorithm involving (15) and (16) could be devised.

### 3. Nested HDG method for the hydrodynamic model

#### 3.1. Approximation spaces

We first review the notation, operators and approximation spaces needed for the HDG method following [63]. We denote by  $\mathcal{T} = \mathcal{T}^m \cup \mathcal{T}^d$  a triangulation of disjoint regular elements  $K$  that partition a nanostructure consisting of a metallic and dielectric subdomains  $\mathcal{D} = \Omega^m \cup \Omega^d \in \mathbb{R}^3$ . The set of element boundaries is then defined as  $\partial\mathcal{T} := \{\partial K : K \in \mathcal{T}\}$ . For an arbitrary element  $K \in \mathcal{T}$ ,  $F = \partial K \cap \partial\mathcal{D}$  is a boundary face if it has a non-zero 2-D Lebesgue measure. Any pair of elements  $K^+$  and  $K^-$  share an interior face  $F = \partial K^+ \cap \partial K^-$  if its 2-D Lebesgue measure is non-zero. We finally denote by  $\mathcal{E}_h^o$  and  $\mathcal{E}_h^\partial$  the set of interior and boundary faces, respectively, and their union by  $\mathcal{E}_h = \mathcal{E}_h^o \cup \mathcal{E}_h^\partial$ .

Let  $\mathbf{n}^+$  and  $\mathbf{n}^-$  be the outward-pointing unit normal vectors on the neighboring elements  $K^+$ ,  $K^-$ , respectively. We further use  $\mathbf{u}^\pm$  to denote the trace of  $\mathbf{u}$  on  $F$  from the interior of  $K^\pm$ . The jump  $[[\cdot]]$  for an interior face  $F \in \mathcal{E}_h^o$  is defined as

$$[[\mathbf{u} \odot \mathbf{n}]] = \mathbf{u}^+ \odot \mathbf{n}^+ + \mathbf{u}^- \odot \mathbf{n}^-,$$

and for a boundary face  $F \in \mathcal{E}_h^\partial$  with outward normal  $\mathbf{n}$  as

$$[[\mathbf{u} \odot \mathbf{n}]] = \mathbf{u} \odot \mathbf{n}.$$

Here, the binary operation  $\odot$  represents either  $\cdot$  or  $\times$ . The tangential  $\mathbf{u}^t$  and normal  $\mathbf{u}^n$  components of  $\mathbf{u}$ , such that  $\mathbf{u} = \mathbf{u}^t + \mathbf{u}^n$ , are given by

$$\mathbf{u}^t = \mathbf{n} \times \mathbf{u} \times \mathbf{n}, \quad \mathbf{u}^n = \mathbf{n}(\mathbf{u} \cdot \mathbf{n}).$$

Let  $\mathbf{L}^2(\mathcal{D}) \equiv [L^2(\mathcal{D})]^3$  denote the Lebesgue space of three dimensional square integrable vector functions and  $H^1(\mathcal{D})$  the Hilbert space  $H^1(\mathcal{D}) = \{v \in L^2(\mathcal{D}) : \int_{\mathcal{D}} |\nabla v|^2 < \infty\}$ . We introduce the curl-conforming space

$$\mathbf{H}^{\text{curl}}(\mathcal{D}) = \{\mathbf{u} \in \mathbf{L}^2(\mathcal{D}) : \nabla \times \mathbf{u} \in \mathbf{L}^2(\mathcal{D})\}$$

with associated norm  $\|\mathbf{u}\|_{\mathbf{H}^{\text{curl}}(\mathcal{D})}^2 = \int_{\mathcal{D}} |\mathbf{u}|^2 + |\nabla \times \mathbf{u}|^2$ , as well as the div-conforming space

$$\mathbf{H}^{\text{div}}(\mathcal{D}) = \{\mathbf{u} \in \mathbf{L}^2(\mathcal{D}) : \nabla \cdot \mathbf{u} \in L^2(\mathcal{D})\}$$

with associated norm  $\|\mathbf{u}\|_{\mathbf{H}^{\text{div}}(\mathcal{D})}^2 = \int_{\mathcal{D}} |\mathbf{u}|^2 + |\nabla \cdot \mathbf{u}|^2$ .

Let  $\mathcal{P}^p(\mathcal{D})$  denote the space of complex-valued polynomials of degree at most  $p$  on  $\mathcal{D}$ . We introduce the following approximation spaces

$$\begin{aligned} W_h &= \{w \in L^2(\mathcal{D}) : w|_K \in \mathcal{P}^p(K), \forall K \in \mathcal{T}_h\}, \\ \mathbf{W}_h &= \{\mathbf{w} \in \mathbf{L}^2(\mathcal{D}) : \mathbf{w}|_K \in [\mathcal{P}^p(K)]^3, \forall K \in \mathcal{T}_h\}, \\ M_h &= \{\mu \in L^2(\mathcal{E}_h) : \mu|_F \in \mathcal{P}^p(F), \forall F \in \mathcal{E}_h\}, \\ \mathbf{M}_h &= \{\boldsymbol{\mu} \in \mathbf{L}^2(\mathcal{E}_h) : \boldsymbol{\mu}|_F \in \mathcal{P}^p(F)\mathbf{t}_1 \oplus \mathcal{P}^p(F)\mathbf{t}_2, \forall F \in \mathcal{E}_h\}, \end{aligned}$$

where  $\mathbf{t}_1, \mathbf{t}_2$  are linearly independent vectors tangent to the face. We note that by construction,  $\boldsymbol{\mu} \in \mathbf{M}_h$  satisfies  $\boldsymbol{\mu} = \mathbf{n} \times \boldsymbol{\mu} \times \mathbf{n} = \mu_1 \mathbf{t}_1 + \mu_2 \mathbf{t}_2$ . The tangent vectors on a face  $F$  can be defined in terms of its normal  $\mathbf{n} = (n_1, n_2, n_3)$  as  $\mathbf{t}_1 = (-n_2/n_1, 1, 0)$  and  $\mathbf{t}_2 = (-n_3/n_1, 0, 1)$ . This definition assumes that  $|n_1| \geq \max(|n_2|, |n_3|)$  but analogous expressions can be obtained when  $|n_2| \geq \max(|n_1|, |n_3|)$  or  $|n_3| \geq \max(|n_1|, |n_2|)$  to avoid singularities. Boundary conditions are included by setting  $\mathbf{M}_h(\mathbf{u}_\partial|_{\partial\mathcal{D}}) = \{\boldsymbol{\mu} \in \mathbf{M}_h : \mathbf{n} \times \boldsymbol{\mu} \times \mathbf{n} = \Pi \mathbf{u}_\partial \text{ on } \partial\mathcal{D}\}$  and  $M_h(u_\partial|_{\partial\mathcal{D}}) = \{\mu \in M_h : \mu = \Pi u_\partial \text{ on } \partial\mathcal{D}\}$ , where  $\Pi \mathbf{u}_\partial$  (respectively,  $\Pi u_\partial$ ) is the projection of the prescribed value of  $\mathbf{u}, u_\partial$ , onto  $\mathbf{M}_h$  (respectively,  $u_\partial$  onto  $M_h$ ).

Finally, we define the various Hermitian products for the above finite element spaces. The volume inner products are defined as

$$(\eta, \zeta)_\mathcal{T} := \sum_{K \in \mathcal{T}} (\eta, \zeta)_K, \quad (\boldsymbol{\eta}, \boldsymbol{\zeta})_\mathcal{T} := \sum_{i=1}^3 (\eta_i, \zeta_i)_\mathcal{T},$$

and the surface inner products by

$$\langle \eta, \zeta \rangle_{\partial\mathcal{T}} := \sum_{K \in \mathcal{T}} \langle \eta, \zeta \rangle_{\partial K}, \quad \langle \boldsymbol{\eta}, \boldsymbol{\zeta} \rangle_{\partial\mathcal{T}} := \sum_{i=1}^3 \langle \eta_i, \zeta_i \rangle_{\partial\mathcal{T}}.$$

For two arbitrary scalar functions  $\eta$  and  $\zeta$ , its scalar product  $(\eta, \zeta)_\mathcal{D}$  is the integral of  $\eta\zeta^*$  on  $\mathcal{D}$ .

### 3.2. First hybridization

In this section, we describe the HDG discretization for a metallic nanostructure, introduced in [81], which is a necessary first step to develop the nested HDG method. For completeness, we will consider both plane-wave illumination – active only when computing first harmonic – as well as the nonlinear source term – active only when computing second harmonic.

The HDG discretization of  $\mathcal{L}^m, \mathcal{L}^d$  and  $b$  needs to be completed with two additional continuity condition, that is enforcing zero jump in the tangential component of  $\mathbf{H}_h$  and in the normal component of  $\mathbf{J}_h$ . For all test functions  $(\boldsymbol{\kappa}, \boldsymbol{\eta}, \boldsymbol{\xi}, \zeta, \boldsymbol{\mu}, \theta) \in \mathbf{W}_h \times \mathbf{W}_h \times \mathbf{W}_h \times W_h \times \mathbf{M}_h \times M_h$ , we seek approximate fields  $(\mathbf{H}_h, \mathbf{E}_h, \mathbf{J}_h, \rho_h, \widehat{\mathbf{E}}_h, \widehat{\rho}_h) \in \mathbf{W}_h \times \mathbf{W}_h \times \mathbf{W}_h \times W_h \times \mathbf{M}_h(\mathbf{0}|_{\partial\Omega_E}) \times M_h(0|_{\partial\Omega_p^m})$  such that

$$\begin{aligned} -i\omega(\mathbf{H}_h, \boldsymbol{\kappa})_{\mathcal{T}^m} + (\mathbf{E}_h, \nabla \times \boldsymbol{\kappa})_{\mathcal{T}^m} + \langle \widehat{\mathbf{E}}_h, \boldsymbol{\kappa} \times \mathbf{n} \rangle_{\partial\mathcal{T}^m \setminus \partial\Omega_E^m} &= \mathbf{0}, \\ -\beta^2(\rho_h, \nabla \cdot \boldsymbol{\eta})_{\mathcal{T}^m} + \beta^2 \langle \widehat{\rho}_h, \boldsymbol{\eta} \cdot \mathbf{n} \rangle_{\partial\mathcal{T}^m \setminus \partial\Omega_p^m} + (\gamma - i\omega)(\mathbf{J}_h, \boldsymbol{\eta})_{\mathcal{T}^m} - \omega_p^2(\mathbf{E}_h, \boldsymbol{\eta})_{\mathcal{T}^m} &= (\mathbf{f}, \boldsymbol{\eta})_{\mathcal{T}^m}, \\ (\mathbf{H}_h, \nabla \times \boldsymbol{\xi})_{\mathcal{T}^m} + \langle \widehat{\mathbf{H}}_h, \boldsymbol{\xi} \times \mathbf{n} \rangle_{\partial\mathcal{T}^m} + i\omega(\varepsilon_\infty \mathbf{E}_h, \boldsymbol{\xi})_{\mathcal{T}^m} - (\mathbf{J}_h, \boldsymbol{\xi})_{\mathcal{T}^m} &= \mathbf{0}, \\ i\omega(\rho_h, \zeta)_{\mathcal{T}^m} - \langle \widehat{\mathbf{J}}_h \cdot \mathbf{n}, \zeta \rangle_{\partial\mathcal{T}^m} + (\mathbf{J}_h, \nabla \zeta)_{\mathcal{T}^m} &= 0, \\ -\langle \mathbf{n} \times \widehat{\mathbf{H}}_h, \boldsymbol{\mu} \rangle_{\partial\mathcal{T}^m \setminus \partial\Omega_E^m} &= \mathbf{0}, \\ \langle \widehat{\mathbf{J}}_h \cdot \mathbf{n}, \theta \rangle_{\partial\mathcal{T}^m \setminus \partial\Omega_p^m} &= 0, \end{aligned} \tag{17}$$

are satisfied in the metal. The fields  $\widehat{\mathbf{E}}_h, \widehat{\mathbf{H}}_h, \widehat{\mathbf{J}}_h, \widehat{\rho}_h$  are single valued on the faces and approximate the tangential component of  $\mathbf{E}, \mathbf{H}, \mathbf{J}$  and the trace of  $\rho$ , respectively. The first four equations are the weak formulation of  $\mathcal{L}^m(\mathbf{H}, \mathbf{E}, \mathbf{J}, \rho; \omega) = \mathbf{f}_{\text{inc}}$ . The boundary equations are strongly prescribed on the approximation spaces for the electric field and electron charge and weakly prescribed on the

last two equations for the magnetic field and the electric current. In addition, the fifth equation enforces zero jump in the tangential component of  $\mathbf{H}_h$ , that is  $[[\mathbf{n} \times \widehat{\mathbf{H}}_h]] = \mathbf{0}$  along all elemental interfaces  $\mathcal{E}_h$ , and lastly the sixth equation enforces zero jump on the normal component of  $\mathbf{J}_h$  along all metal-metal interfaces  $\mathcal{E}_h^m$ , that is  $[[\mathbf{n} \cdot \widehat{\mathbf{J}}_h]] = 0$ .

Similarly, for the dielectric domain  $\Omega^d$  the following weak formulation is satisfied

$$\begin{aligned} -i\omega(\mathbf{H}_h, \boldsymbol{\kappa})_{\mathcal{T}^d} + (\mathbf{E}_h, \nabla \times \boldsymbol{\kappa})_{\mathcal{T}_h} + \langle \widehat{\mathbf{E}}_h, \boldsymbol{\kappa} \times \mathbf{n} \rangle_{\partial\mathcal{T}^d \setminus \partial\Omega_{\mathbb{E}}^d} &= \mathbf{0}, \\ (\mathbf{H}_h, \nabla \times \boldsymbol{\xi})_{\mathcal{T}^d} + \langle \widehat{\mathbf{H}}_h, \boldsymbol{\xi} \times \mathbf{n} \rangle_{\partial\mathcal{T}^d} + i\omega(\varepsilon_d \mathbf{E}_h, \boldsymbol{\xi})_{\mathcal{T}^d} &= \mathbf{0}, \\ -\langle \mathbf{n} \times \widehat{\mathbf{H}}_h, \boldsymbol{\mu} \rangle_{\partial\mathcal{T}^d \setminus \partial\Omega_{\mathbb{E}}^d} - \sqrt{\varepsilon_d} \langle \widehat{\mathbf{E}}_h, \boldsymbol{\mu} \rangle_{\partial\Omega_{\text{rad}}^d} &= \langle \mathbf{f}_{\text{inc}}, \boldsymbol{\mu} \rangle_{\partial\Omega_{\text{rad}}^d} \end{aligned} \quad (18)$$

We close the system by introducing expressions for the hybrid fluxes of the magnetic field and electric current field as

$$\begin{aligned} \widehat{\mathbf{H}}_h &= \mathbf{H}_h + \tau_t(\mathbf{E}_h - \widehat{\mathbf{E}}_h) \times \mathbf{n}, \\ \widehat{\mathbf{J}}_h \cdot \mathbf{n} &= \mathbf{J}_h \cdot \mathbf{n} - \tau_n i\omega(\rho_h - \widehat{\rho}_h). \end{aligned} \quad (19)$$

The parameters  $\tau_t, \tau_n$  are the stabilization parameters, defined globally to ensure the accuracy and stability of the HDG discretization. We propose the choice  $\tau_t = \sqrt{\varepsilon}$  (for each material's permittivity) and  $\tau_n = \omega_p/\beta$  which leads to numerically stable solutions even in the presence of tightly localized fields in the metal-dielectric interface.

Substituting (19) in (17) and integrating by parts, we write the final HDG discretization for the metallic domain

$$\begin{aligned} -i\omega(\mathbf{H}_h, \boldsymbol{\kappa})_{\mathcal{T}^m} + (\mathbf{E}_h, \nabla \times \boldsymbol{\kappa})_{\mathcal{T}^m} + \langle \widehat{\mathbf{E}}_h, \boldsymbol{\kappa} \times \mathbf{n} \rangle_{\partial\mathcal{T}^m \setminus \partial\Omega_{\mathbb{E}}^m} &= \mathbf{0}, \\ -\beta^2(\rho_h, \nabla \cdot \boldsymbol{\eta})_{\mathcal{T}^m} + \beta^2 \langle \widehat{\rho}_h, \boldsymbol{\eta} \cdot \mathbf{n} \rangle_{\partial\mathcal{T}^m \setminus \partial\Omega_{\mathbb{P}}^m} + (\gamma - i\omega)(\mathbf{J}_h, \boldsymbol{\eta})_{\mathcal{T}^m} - \omega_p^2(\mathbf{E}_h, \boldsymbol{\eta})_{\mathcal{T}^m} &= (\mathbf{f}_2, \boldsymbol{\eta})_{\mathcal{T}^m}, \\ (\nabla \times \mathbf{H}_h, \boldsymbol{\xi})_{\mathcal{T}^m} + \langle \tau_t[\mathbf{E}_h - \widehat{\mathbf{E}}_h], \mathbf{n} \times \boldsymbol{\xi} \times \mathbf{n} \rangle_{\partial\mathcal{T}^m} + i\omega(\varepsilon_\infty \mathbf{E}_h, \boldsymbol{\xi})_{\mathcal{T}^m} - (\mathbf{J}_h, \boldsymbol{\xi})_{\mathcal{T}^m} &= \mathbf{0}, \\ -(\nabla \cdot \mathbf{J}_h, \zeta)_{\mathcal{T}^m} + i\omega(\rho_h, \zeta)_{\mathcal{T}^m} + i\omega\tau_n \langle \rho_h, \zeta \rangle_{\partial\mathcal{T}^m} - i\omega\tau_n \langle \widehat{\rho}_h, \zeta \rangle_{\partial\mathcal{T}^m} &= 0, \\ -\langle \mathbf{n} \times \mathbf{H}_h, \boldsymbol{\mu} \rangle_{\partial\mathcal{T}^m \setminus \partial\Omega_{\mathbb{E}}^m} - \langle \tau_t \mathbf{E}_h, \boldsymbol{\mu} \rangle_{\partial\mathcal{T}^m \setminus \partial\Omega_{\mathbb{E}}^m} + \langle \tau_t \widehat{\mathbf{E}}_h, \boldsymbol{\mu} \rangle_{\partial\mathcal{T}^m \setminus \partial\Omega_{\mathbb{E}}^m} &= \mathbf{0}, \\ \langle \mathbf{J}_h \cdot \mathbf{n}, \theta \rangle_{\partial\mathcal{T}^m \setminus \partial\Omega_{\mathbb{P}}^m} - i\omega\tau_n \langle \rho_h, \theta \rangle_{\partial\mathcal{T}^m \setminus \partial\Omega_{\mathbb{P}}^m} + i\omega\tau_n \langle \widehat{\rho}_h, \theta \rangle_{\partial\mathcal{T}^m \setminus \partial\Omega_{\mathbb{P}}^m} &= 0. \end{aligned} \quad (20)$$

The nonlinear source also requires integration by parts, hence

$$\begin{aligned} (\mathbf{f}_2, \boldsymbol{\eta})_{\mathcal{T}^m} &= \frac{\omega_p^2}{2n_0 e} (\mathbf{E}_1 \rho_1, \boldsymbol{\eta})_{\mathcal{T}^m} + \frac{\omega_p^2}{2n_0 e} (\mathbf{J}_1 \times \mathbf{H}_1)_{\mathcal{T}^m} \\ &\quad - \frac{1}{2n_0 e} \left[ \langle (\widehat{\mathbf{J}}_1 \otimes \widehat{\mathbf{J}}_1) \cdot \mathbf{n}, \boldsymbol{\eta} \rangle_{\partial\mathcal{T}^m} - (\mathbf{J}_1 \otimes \mathbf{J}_1, \nabla \boldsymbol{\eta})_{\mathcal{T}^m} \right] \\ &\quad - \frac{\beta^2}{6n_0 e} \left[ \langle \widehat{\rho}_1^2, \boldsymbol{\eta} \cdot \mathbf{n} \rangle_{\partial\mathcal{T}^m} - (\rho_1^2, \nabla \cdot \boldsymbol{\eta})_{\mathcal{T}^m} \right]. \end{aligned}$$

The final HDG discretization for the dielectric is obtained after substituting (19) in (18) and

integrating by parts

$$\begin{aligned}
& -i\omega(\mathbf{H}_h, \boldsymbol{\kappa})_{\mathcal{T}^d} + (\mathbf{E}_h, \nabla \times \boldsymbol{\kappa})_{\mathcal{T}^d} + \langle \widehat{\mathbf{E}}_h, \boldsymbol{\kappa} \times \mathbf{n} \rangle_{\partial\mathcal{T}^d \setminus \partial\Omega_{\mathbb{E}}^d} = \mathbf{0}, \\
& (\nabla \times \mathbf{H}_h, \boldsymbol{\xi})_{\mathcal{T}^d} + \langle \tau_t[\mathbf{E}_h - \widehat{\mathbf{E}}_h], \mathbf{n} \times \boldsymbol{\xi} \times \mathbf{n} \rangle_{\partial\mathcal{T}^d} + i\omega(\varepsilon_d \mathbf{E}_h, \boldsymbol{\xi})_{\mathcal{T}^d} = \mathbf{0}, \\
& -\langle \mathbf{n} \times \mathbf{H}_h, \boldsymbol{\mu} \rangle_{\partial\mathcal{T}^d \setminus \partial\Omega_{\mathbb{E}}^d} - \langle \tau_t \mathbf{E}_h, \boldsymbol{\mu} \rangle_{\partial\mathcal{T}^d \setminus \partial\Omega_{\mathbb{E}}^d} + \\
& + \langle \tau_t \widehat{\mathbf{E}}_h, \boldsymbol{\mu} \rangle_{\partial\mathcal{T}^d \setminus \partial\Omega_{\mathbb{E}}^d} - \sqrt{\varepsilon_d} \langle \widehat{\mathbf{E}}_h, \boldsymbol{\mu} \rangle_{\partial\Omega_{\text{rad}}^d} = \langle \mathbf{f}_{\text{inc}}, \boldsymbol{\mu} \rangle_{\partial\Omega_{\text{rad}}^d}.
\end{aligned} \tag{21}$$

The weak formulations (20)-(21) are then discretized using the corresponding basis functions on all the elements and faces of  $\mathcal{T}$ , thus giving rise to the linear system

$$\begin{bmatrix}
-i\omega\mathbb{A} & 0 & \mathbb{B} & 0 & \vdots & \mathbb{C} & 0 \\
0 & (\gamma - i\omega)\mathbb{A} & -\omega_p^2\mathbb{A} & -\beta^2\mathbb{P} & \vdots & 0 & \beta^2\mathbb{O} \\
\mathbb{B}^T & -\mathbb{A} & \mathbb{D} + i\omega\mathbb{A}_\varepsilon & 0 & \vdots & -\mathbb{E} & 0 \\
0 & -\mathbb{P}^T & 0 & i\omega\mathbb{H} & \vdots & 0 & -i\omega\mathbb{N} \\
\text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\
\mathbb{C}^T & 0 & -\mathbb{E}^T & 0 & \vdots & \mathbb{M} & 0 \\
0 & \mathbb{O}^T & 0 & -i\omega\mathbb{N}^T & \vdots & 0 & i\omega\mathbb{T}
\end{bmatrix}
\begin{bmatrix}
\mathbf{H}_m \\
\mathbf{J}_m \\
\mathbf{E}_m \\
\rho_m \\
\widehat{\mathbf{E}}_m \\
\widehat{\rho}_m
\end{bmatrix}
=
\begin{bmatrix}
0 \\
\mathbf{f}_2 \\
0 \\
0 \\
0 \\
0
\end{bmatrix}, \tag{22}$$

for the metal domain, and

$$\begin{bmatrix}
-i\omega\mathbb{A} & \mathbb{B} & \vdots & \mathbb{C} \\
\mathbb{B}^T & \mathbb{D} + i\omega\mathbb{A}_\varepsilon & \vdots & -\mathbb{E} \\
\text{---} & \text{---} & \text{---} & \text{---} \\
\mathbb{C}^T & -\mathbb{E}^T & \vdots & \mathbb{M}
\end{bmatrix}
\begin{bmatrix}
\mathbf{H}_d \\
\mathbf{E}_d \\
\widehat{\mathbf{E}}_d
\end{bmatrix}
=
\begin{bmatrix}
0 \\
0 \\
\mathbf{f}_{\text{inc}}
\end{bmatrix}, \tag{23}$$

for the dielectric domain. Here  $(\mathbf{H}_m, \mathbf{E}_m, \widehat{\mathbf{E}}_m)$  and  $(\mathbf{H}_d, \mathbf{E}_d, \widehat{\mathbf{E}}_d)$  represent the vectors of degrees of freedom of  $(\mathbf{H}_h, \mathbf{E}_h, \widehat{\mathbf{E}}_h)$  in metal and dielectric, respectively. Similarly,  $(\mathbf{J}_m, \rho_m, \widehat{\rho}_m)$  represent the vectors of degrees of freedom of  $(\mathbf{J}_h, \rho_h, \widehat{\rho}_h)$  in metal. It is important to point out that the two systems (22) and (23) have to be solved simultaneously because  $\widehat{\mathbf{E}}_m$  and  $\widehat{\mathbf{E}}_d$  share degrees of freedom for the faces located on the metal-dielectric interface.

We note that due to the discontinuous nature of the approximation spaces, the local variables  $\mathbf{E}_h, \mathbf{H}_h, \mathbf{J}_h, \rho_h$  (defined in the interior of each element) are only coupled globally through the global variables  $\widehat{\mathbf{E}}, \widehat{\rho}$  (defined on the element faces). This means that we can eliminate these local variables at the element level, which in matrix form corresponds to eliminating the upper-left submatrices, indicated with the dashed lines in (22) and (23), thus only a reduced matrix for the global variables needs to be assembled. This numerical strategy, also known as hybridization or static condensation, is essential to achieve an efficient implementation of the HDG method [22–24, 61–63]. Specifically, the hybridization procedure yields the following global linear system

$$\Upsilon \widehat{\mathbf{U}} = \mathbf{b}, \tag{24}$$

where the vector  $\widehat{\mathbf{U}}$  consists of the degrees of freedom of  $(\widehat{\mathbf{E}}_h, \widehat{\rho}_h)$ . In practice, both the matrix  $\Upsilon$  and the vector  $\mathbf{b}$  are formed by a standard finite element assembly procedure by computing

the elemental quantities and assembling them in an element-by-element fashion. The detailed implementation can be found in [79, 81].

The elimination of local degrees of freedom through hybridization renders a linear system where the global degrees of freedom are defined on the faces only, thus drastically reducing the size of the linear system that must be solved. After solving the linear system (24) for the global unknowns, the local unknowns can be efficiently recovered at the element level [79, 81], an operation that is trivially parallelizable.

### 3.3. Nested hybridization

Thus far, we have recreated the formulation and implementation of the HDG method for a metal-dielectric domain introduced in [81]. In this section, we describe a nested hybridization method to efficiently solve the global linear system that stems from the first hybridization (24) by exploiting the geometry of the problems of interest. Indeed, for large 3-D structures the direct solution of (24) may be challenging. Iterative methods, on the other hand, have found limited success for stiff indefinite problems of the type considered here.

The idea behind the nested hybridization is to partition the global degrees of freedom  $\widehat{\mathbf{U}} = \{\mathbf{W}, \mathbf{V}\}$ , such that the  $\mathbf{W}$  degrees of freedom can be statically condensed to yield the following linear system

$$\Psi \mathbf{V} = \mathbf{d} \quad (25)$$

for  $\mathbf{V}$ . Specifically, reordering the global system (24) using the  $\widehat{\mathbf{U}} = \{\mathbf{W}, \mathbf{V}\}$  allows us to write of  $\Psi$ ,  $\mathbf{d}$  as a function of  $\Upsilon$ ,  $\mathbf{b}$ , namely

$$\begin{bmatrix} \Upsilon_{\mathbf{W}\mathbf{W}} & \Upsilon_{\mathbf{W}\mathbf{V}} \\ \Upsilon_{\mathbf{V}\mathbf{W}} & \Upsilon_{\mathbf{V}\mathbf{V}} \end{bmatrix} \begin{bmatrix} \mathbf{W} \\ \mathbf{V} \end{bmatrix} = \begin{bmatrix} \mathbf{b}_{\mathbf{W}} \\ \mathbf{b}_{\mathbf{V}} \end{bmatrix}. \quad (26)$$

Assuming that  $\Upsilon_{\mathbf{W}\mathbf{W}}$  is invertible, the above system may be recast as

$$\mathbf{W} = \Upsilon_{\mathbf{W}\mathbf{W}}^{-1} [\mathbf{b}_{\mathbf{W}} - \Upsilon_{\mathbf{W}\mathbf{V}} \mathbf{V}] \quad (27)$$

$$[\Upsilon_{\mathbf{V}\mathbf{V}} - \Upsilon_{\mathbf{V}\mathbf{W}} \Upsilon_{\mathbf{W}\mathbf{W}}^{-1} \Upsilon_{\mathbf{W}\mathbf{V}}] \mathbf{V} = \mathbf{b}_{\mathbf{V}} - \Upsilon_{\mathbf{V}\mathbf{W}} \Upsilon_{\mathbf{W}\mathbf{W}}^{-1} \mathbf{b}_{\mathbf{W}}. \quad (28)$$

Thus we have  $\Psi = \Upsilon_{\mathbf{V}\mathbf{V}} - \Upsilon_{\mathbf{V}\mathbf{W}} \Upsilon_{\mathbf{W}\mathbf{W}}^{-1} \Upsilon_{\mathbf{W}\mathbf{V}}$  and  $\mathbf{d} = \mathbf{b}_{\mathbf{V}} - \Upsilon_{\mathbf{V}\mathbf{W}} \Upsilon_{\mathbf{W}\mathbf{W}}^{-1} \mathbf{b}_{\mathbf{W}}$ .

For this hybridization to be computationally efficient, we need to ensure the inverse of  $\Upsilon_{\mathbf{W}\mathbf{W}}$  may be efficiently evaluated by means of a judicious choice of the global degrees of freedom. To that end, we target specifically metallic nanostructures that can be discretized by extruding a 2-D discretization. There are many examples of such structures in the literature [18, 19, 21, 67, 81, 87, 88, 92]. Under this assumption, the degrees of freedom in the planes perpendicular to the extrusion direction are assigned to  $\mathbf{V}$ , and the remaining degrees of freedom are assigned to  $\mathbf{W}$ . This partitioning gives rise to  $N$  blocks  $\{\Omega_n\}_{n=1}^N$ , whose unknowns are  $\{\mathbf{W}_n\}_{n=1}^N$ ; and  $N + 1$  interfaces  $\{\Gamma_n\}_{n=0}^N$ , whose unknowns are  $\{\mathbf{V}_n\}_{n=0}^N$ , as illustrated in Fig. 1. Consequently, the matrix  $\Upsilon_{\mathbf{W}\mathbf{W}}$  can be inverted efficiently since it is a block-diagonal matrix, namely

$$\Upsilon_{\mathbf{W}\mathbf{W}} = \text{diag}(\Upsilon_{\mathbf{W}_1\mathbf{W}_1}, \Upsilon_{\mathbf{W}_2\mathbf{W}_2}, \dots, \Upsilon_{\mathbf{W}_N\mathbf{W}_N}).$$



The blocks may be computed explicitly following (27) and (28) to obtain

$$\begin{aligned}
\alpha_0 &= \Upsilon_{V_0 V_0} - \Upsilon_{V_0 W_1} \Upsilon_{W_1 W_1}^{-1} \Upsilon_{W_1 V_0}, \\
\beta_0 &= \Upsilon_{V_0 V_1} - \Upsilon_{V_0 W_1} \Upsilon_{W_1 W_1}^{-1} \Upsilon_{W_1 V_1}, \\
\gamma_0 &= \Upsilon_{V_1 V_0} - \Upsilon_{V_1 W_1} \Upsilon_{W_1 W_1}^{-1} \Upsilon_{W_1 V_0}, \\
\mathbf{d}_0 &= \mathbf{b}_{V_0} - \Upsilon_{V_0 W_1} \Upsilon_{W_1 W_1}^{-1} \mathbf{b}_{W_1}, \\
\alpha_N &= \Upsilon_{V_N V_N} - \Upsilon_{V_N W_N} \Upsilon_{W_N W_N}^{-1} \Upsilon_{W_N V_N}, \\
\mathbf{d}_N &= \mathbf{b}_{V_N} - \Upsilon_{V_N W_N} \Upsilon_{W_N W_N}^{-1} \mathbf{b}_{W_N},
\end{aligned} \tag{30}$$

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

$$\begin{aligned}
\alpha_n &= \Upsilon_{V_n V_n} - \Upsilon_{V_n W_n} \Upsilon_{W_n W_n}^{-1} \Upsilon_{W_n V_n} - \Upsilon_{V_n W_{n+1}} \Upsilon_{W_{n+1} W_{n+1}}^{-1} \Upsilon_{W_{n+1} V_n}, \\
\beta_n &= \Upsilon_{V_n V_{n+1}} - \Upsilon_{V_n W_{n+1}} \Upsilon_{W_{n+1} W_{n+1}}^{-1} \Upsilon_{W_{n+1} V_{n+1}}, \\
\gamma_n &= \Upsilon_{V_{n+1} V_n} - \Upsilon_{V_{n+1} W_{n+1}} \Upsilon_{W_{n+1} W_{n+1}}^{-1} \Upsilon_{W_{n+1} V_n}, \\
\mathbf{d}_n &= \mathbf{b}_{V_n} - \Upsilon_{V_n W_n} \Upsilon_{W_n W_n}^{-1} \mathbf{b}_{W_n} - \Upsilon_{V_n W_{n+1}} \Upsilon_{W_{n+1} W_{n+1}}^{-1} \mathbf{b}_{W_{n+1}}.
\end{aligned} \tag{31}$$

The block tridiagonal system is never formed in practice, but rather solved on-the-fly using Thomas method (forward Gaussian elimination for tridiagonal matrices). The first equation in (29) is recast as

$$\mathbf{V}_0 = \alpha_0^{-1} (\mathbf{d}_0 - \beta_0 \mathbf{V}_1) = \tilde{\mathbf{d}}_0 + \tilde{\alpha}_0 \mathbf{V}_1. \tag{32}$$

Using this relation, the second equation reads

$$\mathbf{V}_1 = (\alpha_1 + \gamma_0 \tilde{\alpha}_0)^{-1} (\mathbf{d}_1 - \gamma_0 \tilde{\mathbf{d}}_0 - \beta_1 \mathbf{V}_2) = \tilde{\mathbf{d}}_1 + \tilde{\alpha}_1 \mathbf{V}_2, \tag{33}$$

and we can thus establish an analogous expression for the subsequent equations  $n = 2, \dots, N-1$ , that is

$$\mathbf{V}_n = (\alpha_n + \gamma_{n-1} \tilde{\alpha}_{n-1})^{-1} (\mathbf{d}_n - \gamma_{n-1} \tilde{\mathbf{d}}_{n-1} - \beta_n \mathbf{V}_{n+1}) = \tilde{\mathbf{d}}_n + \tilde{\alpha}_n \mathbf{V}_{n+1}. \tag{34}$$

Hence, from the last equation we can retrieve the value for  $\mathbf{V}_N$  as

$$\mathbf{V}_N = (\alpha_N + \gamma_{N-1} \tilde{\alpha}_{N-1})^{-1} (\mathbf{d}_N - \gamma_{N-1} \tilde{\mathbf{d}}_{N-1}) = \tilde{\alpha}_N \tilde{\mathbf{d}}_N, \tag{35}$$

and we then march backwards to recover the remaining  $\{\mathbf{V}_n\}_{n=N-1}^0$  leveraging (34)–(32).

### 3.4. Implementation

In this section, we describe how to efficiently compute (30)–(31) and solve (29). The first step is to judiciously partition the computational domain in the extrusion direction to define the blocks and interfaces. For a given block  $\Omega_n$ , the values on  $\Upsilon_{W_n W_n}$  depend solely on the spatial dimensions of the block and the material properties ( $\varepsilon_d$  for dielectric and  $\varepsilon_\infty, \omega_p, \gamma, \beta$  for metal) and for a given interface  $\Gamma_n$  the values  $\Upsilon_{V_n V_n}$  depend solely on the spatial dimensions of the interface and the adjacent blocks' material properties. Hence, accounting for these features when partitioning the domain allows us to define unique block and interface types  $\{\bar{\Omega}_\ell\}_{\ell=1}^L, \{\bar{\Gamma}_m\}_{m=1}^M$  that may translate into important computational savings, similar to the strategy developed by Huynh *et al.*

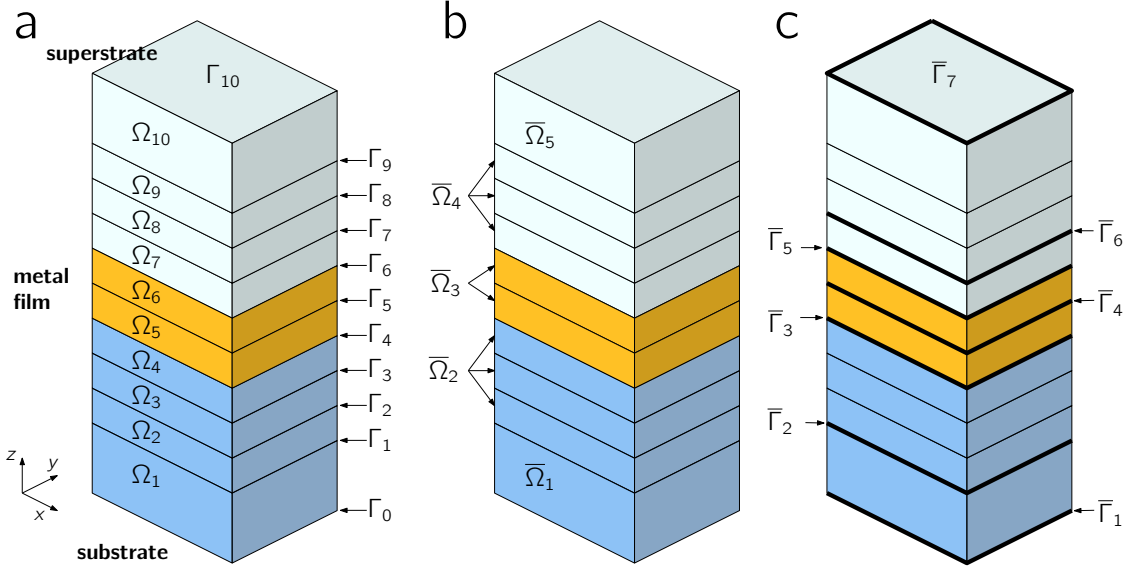


Figure 2: Substrate-metal-superstrate sample structure. (a) Block and interface definition. (b) Block types. (c) Interface types.

Blocks		Interfaces	
$\Omega_n$	$\xrightarrow{\sigma_\Omega} \bar{\Omega}_\ell$	$\Gamma_n$	$\xrightarrow{\sigma_\Gamma} \bar{\Gamma}_m$
1	1	0	1
2-4	2	1-3	2
5-6	3	4	3
7-9	4	5	4
10	5	6	5
		7-9	6
		10	7

Table 1: Block and interface type assignment for structure in Fig. 2.

in the context of efficient model order reduction for structured problems [29, 36, 37, 83]. The type assignments can be mathematically expressed with a tuple of maps  $(\sigma_\Omega, \sigma_\Gamma)$ , such that  $\sigma_\Omega(n) = \ell$  if block  $n$  belongs to type  $\ell$ , and analogously for the interfaces with  $\sigma_\Gamma$ . We have illustrated the type definition on a substrate-metal-superstrate structure shown in Fig. 2, where the block partitioning is done along the  $z$  axis. After discretization, we are left with 4, 2 and 4 substrate, metal and superstrate blocks respectively, along with 11 interfaces. Based on the dimensions and the material properties, the type assignment is summarized in Table 1.

Once the type maps have been established and the matrix and forcing term  $\mathbf{Y}$ ,  $\mathbf{b}$  have been computed, the degrees of freedom  $\hat{\mathbf{U}}$  following Algorithm 1, a procedure that we refer to as nested HDG with full assembly. Since an integral part of this nested hybridization method is the ability to reuse computations by virtue of the block and interface types, we shall define six lists  $\{v_i\}_{i=1}^6$  where all relevant computations are stored and can thus be accessed whenever required by Algo-



---

**Algorithm 1** Nested hybridization with full assembly to solve (24)

---

**Require:** HDG matrix and forcing term  $\Upsilon$ ,  $\mathbf{b}$  of entire domain  $\Omega$ ; type maps  $(\sigma_\Omega, \sigma_\Gamma)$  of blocks  $\{\Omega_n\}_{n=1}^N$  and interfaces  $\{\Gamma_n\}_{n=0}^N$ ; six empty lists  $v_1, \dots, v_6$

```

1: for  $n = 1 : N$  do
2:   Set  $\ell = \sigma_\Omega(n)$ ,  $m_0 = \sigma_\Gamma(n-1)$  and  $m_1 = \sigma_\Gamma(n)$ 
3:   if  $v_1(\ell) = \emptyset$  then
4:     Compute LU decomposition of  $\Upsilon_{W_\ell W_\ell} = LU$  and store  $LU \mapsto v_1(\ell)$ 
5:     Solve linear system for multiple right-hand-sides in parallel
           
$$v_1(\ell)[\mathbf{x}_0, \mathbf{x}_1, \mathbf{y}] = [\Upsilon_{W_\ell, V_{m_0}}, \Upsilon_{W_\ell, V_{m_1}}, \mathbf{b}_{W_n}],$$

           store  $\mathbf{x}_0 \mapsto v_2[\ell, m_0]$ ,  $\mathbf{x}_1 \mapsto v_3[\ell, m_1]$  and set  $\mathbf{x} = [\mathbf{x}_0, \mathbf{x}_1, \mathbf{y}]$ 
6:   else
7:     if  $v_2(\ell, m_0) = \emptyset$  and  $v_3(\ell, m_1) \neq \emptyset$  then
8:       Solve linear system for multiple right-hand-sides in parallel
           
$$v_1(\ell)[\mathbf{x}_0, \mathbf{y}] = [\Upsilon_{W_\ell, V_{m_0}}, \mathbf{b}_{W_n}],$$

           store  $\mathbf{x}_0 \mapsto v_2[\ell, m_0]$  and set  $\mathbf{x} = [\mathbf{x}_0, v_3[\ell, m_1], \mathbf{y}]$ 
9:     else if  $v_2(\ell, m_0) \neq \emptyset$  and  $v_3(\ell, m_1) = \emptyset$  then
10:      Solve linear system for multiple right-hand-sides in parallel
           
$$v_1(\ell)[\mathbf{x}_1, \mathbf{y}] = [\Upsilon_{W_\ell, V_{m_1}}, \mathbf{b}_{W_n}],$$

           store  $\mathbf{x}_1 \mapsto v_3[\ell, m_1]$  and set  $\mathbf{x} = [v_2[\ell, m_0], \mathbf{x}_1, \mathbf{y}]$ 
11:    else
12:      Solve linear system  $v_1(\ell)\mathbf{b} = \mathbf{b}_{W_n}$  and set  $\mathbf{x} = [v_2[\ell, m_0], v_3[\ell, m_1], \mathbf{y}]$ 
13:    end if
14:  end if
15:  Multiply and store  $\Upsilon_{V_{m_0}, W_\ell} \mathbf{x} \mapsto v_4[\ell, m_0]$ ,  $\Upsilon_{V_{m_1}, W_\ell} \mathbf{x} \mapsto v_5[\ell, m_1]$ 
16:  Compute  $\alpha_{n-1}$ ,  $\beta_{n-1}$ ,  $\gamma_{n-1}$  and  $\mathbf{d}_{n-1}$  using (30)-(31)
17:  Compute and store  $\tilde{\alpha}_{n-1} \mapsto v_6[n, 1]$  and  $\tilde{\mathbf{d}}_{n-1} \mapsto v_6[n, 2]$  using (32)-(34)
18:  if  $n = N$  then
19:    Compute  $\alpha_N$ ,  $\mathbf{d}_N$  using (30) and  $\tilde{\alpha}_N$ ,  $\tilde{\mathbf{d}}_N$  using (35)
20:    Evaluate  $\mathbf{V}_N = \tilde{\alpha}_N \tilde{\mathbf{d}}_N$ 
21:  end if
22: end for
23: for  $n = N : -1 : 1$  with  $\ell = \sigma_\Omega(n)$ ,  $m_0 = \sigma_\Gamma(n-1)$ ,  $m_1 = \sigma_\Gamma(n)$  do

```

$$\begin{aligned} \mathbf{V}_{n-1} &= v_6(n, 2) + v_6(n, 1)\mathbf{V}_n \\ v_1(\ell)\mathbf{W}_n &= \mathbf{b}_{W_n} - \Upsilon_{W_\ell V_{m_0}} \mathbf{V}_{n-1} - \Upsilon_{W_\ell V_{m_1}} \mathbf{V}_n \end{aligned}$$

```

24: end for

```

---

rithm 1. These lists will contain the following items:  $v_1$  stores the LU decomposition of  $\Upsilon_{WW}$  for each block type, and are indexed by block type;  $v_2, \dots, v_5$  contain the building blocks of the Schur

decomposition at the block-interface level, and are indexed by both block and interface type, see Algorithm 1 for the exact expressions; finally,  $v_6$  contains each intermediate factor in the solution of the tridiagonal system (29) using forward block Gaussian elimination, that is  $\{\tilde{\boldsymbol{\alpha}}_n, \tilde{\boldsymbol{d}}_n\}_{n=0}^N$ , which are required to compute the degrees of freedom for  $\mathbf{V}$  with equations (32)–(34).

We make the following remarks regarding Algorithm 1: (i) instead of precomputing the HDG matrix  $\boldsymbol{\Upsilon}$  and forcing  $\mathbf{b}$ , which may require significant storage, the matrix and forcing term can be partially assembled on-the-fly after operation 2, that is only for the degrees of freedom  $\{\mathbf{W}_\ell, \mathbf{V}_{m_0}, \mathbf{V}_{m_1}\}$  and purged after each iteration, a variation that is referred to hereafter as nested HDG with partial assembly; however, if this strategy is pursued both matrix and forcing term need to be partially re-assembled in operation 23 to recover the solution field, hence the saving in memory (the HDG matrix is never entirely assembled) comes at the expense of a higher computational runtime; (ii) since storing and reusing computations is an integral part of the algorithm, we may also eliminate items from  $v_2, \dots, v_5$  as soon as they are no longer needed; (iii) the solution of the linear systems for multiple right-hand-sides in steps 5, 8 and 10 of the algorithm is the most computationally intensive, although it can be trivially parallelized; (iv) the storage of  $v_6$  requires significant memory storage since all steps are needed to recover the interface degrees of freedom as in operation 23; and (v) this hybridization results in a strong compression of the original problem, since once  $\{\mathbf{V}_n\}_{n=0}^N$  have been recovered we may evaluate  $\{\mathbf{W}_n\}_{n=1}^N$  through operation 23 block-wise at minimal cost (the LU decompositions are already stored in  $v_1$ ) and then obtain the local variables  $\{\mathbf{E}_h, \mathbf{H}_h, \mathbf{J}_h, \rho_h\}$  with the classical HDG static condensation expressions at the element level, see [79, 81].

The computational strategy described above is a purely algebraic construction, hence it is not only applicable to the HDG discretization of Maxwell’s equations augmented with the hydrodynamic model, but to any linear system arising from an HDG discretization. In order to minimize the computational and memory requirements that stem from the nested HDG method described in Algorithm 1, a carefully designed mesh and judicious degree-of-freedom choice is critical.

## 4. Numerical results

The numerical results presented in this section have been simulated with the MATLAB implementation of the nested HDG method described above and the classical HDG for Maxwell’s equations introduced in [63, 81]. The computational times and memory footprint of the simulations correspond to a 512GB Linux 18.04 machine with 16 AMD Opteron(tm) Processors 6320x15 that has been used to perform the simulations.

### 4.1. Plane wave through layered media

In this section, we perform a numerical test to verify the implementation and accuracy of the nested HDG. To that end, we use a plane wave propagating through a sapphire-silica-air layered medium under normal incidence, for which the exact solution is known, and compare the errors of classical HDG, *i.e.* direct solution of (24), with nested hybridization described by Algorithm 1.

The refractive indices of the layers are 3.31 (sapphire), 1.98 (silica) and 1 (air), with respective thicknesses of 500, 250 and 250 nm. The plane wave is an  $x$ -polarized 1 micron wavelength plane wave propagating in the positive  $z$ -direction, impinging from the sapphire layer. The computational domain is a prism of  $250 \times 250 \times 1000$  nm discretized in  $6 \times 6 \times n$  isotropic cubes, and we prescribe  $\mathbf{E} \times \mathbf{n} = \mathbf{0}$  on the  $x$ -constant boundaries and  $\mathbf{H} \times \mathbf{n} = \mathbf{0}$  on the  $y$ -constant boundaries, as well as

$p$	$n$	HDG				nested HDG			
		$\ \mathbf{E}_0 - \mathbf{E}_h\ _{L^2}$		$\ \mathbf{E}_0 - \mathbf{E}_h\ _{\mathbf{H}^{\text{curl}}}$		$\ \mathbf{E}_0 - \mathbf{E}_h\ _{L^2}$		$\ \mathbf{E}_0 - \mathbf{E}_h\ _{\mathbf{H}^{\text{curl}}}$	
		Error	Order	Error	Order	Error	Order	Error	Order
2	8	5.0e-2	–	2.2e-1	–	5.0e-2	–	2.2e-1	–
	16	2.6e-3	4.27	4.7e-2	2.20	2.6e-3	4.27	4.7e-2	2.20
	32	1.2e-4	4.42	1.2e-2	2.01	1.2e-4	4.42	1.2e-2	2.01
	64	9.7e-6	3.63	2.9e-3	2.00	9.7e-6	3.63	2.9e-3	2.00
3	8	2.9e-3	–	4.5e-2	–	2.9e-3	–	4.5e-2	–
	16	6.6e-5	5.47	5.6e-3	3.01	6.6e-5	5.47	5.6e-3	3.01
	32	3.0e-6	4.46	7.2e-4	2.97	3.0e-6	4.46	7.2e-4	2.97
	64	1.7e-7	4.14	9.0e-5	2.99	1.7e-7	4.14	9.0e-5	2.99
4	8	1.6e-4	–	6.6e-3	–	1.6e-4	–	6.6e-3	–
	16	3.5e-6	5.51	4.8e-4	3.79	3.5e-6	5.51	4.8e-4	3.79
	32	9.3e-8	5.23	3.1e-5	3.96	9.3e-8	5.23	3.1e-5	3.96
	64	2.8e-9	5.08	1.9e-6	3.99	2.8e-9	5.08	1.9e-6	3.99

Table 2: History of convergence for the HDG and nested HDG solution.

first-order absorbing conditions on the top and bottom boundaries. We focus on  $p = 2, 3, 4$  and several  $n$  values, and compute the  $L^2(\mathcal{T})$  and  $\mathbf{H}^{\text{curl}}(\mathcal{T})$  errors of  $\mathbf{E}_h$ , collected in Table 2 along with a convergence analysis. Note that since the cubes in the  $z$ -direction are isotropic, there are only three different types of blocks (sapphire, silica, air) and seven interfaces (upper-lower boundaries, sapphire-sapphire, sapphire-silica, silica-silica, silica-air and air-air).

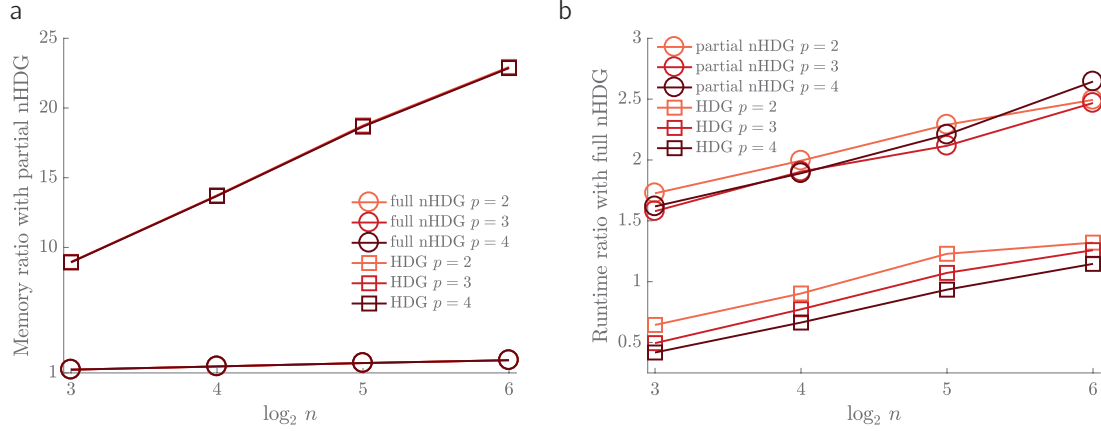


Figure 3: Comparison of computational costs between methods, for  $p = 2, 3, 4$  and  $n = 8, 16, 32, 64$ : (a) Memory requirements with respect to nested HDG with partial assembly. (b) Runtime with respect to nested HDG with full assembly.

As expected, the solutions computed by both methods have the exact same errors and orders of convergence, since the nested HDG is just an algebraic modification that enables a more efficient solution of the classical HDG linear system. We now compare the computational costs and memory requirements of HDG to the nested version, both with full and partial assembly. The computational

runtimes correspond to the wall time averages of 10 individual simulations for each method; the memory is measured in terms of RAM GB required to execute the algorithms. The most memory-efficient method is the nested HDG with partial assembly, and the differences become starker as the mesh is refined, requiring half the memory of that of nested HDG with full assembly and about 20 times less than HDG for the finest mesh, see Fig. 3(a). In terms of computational runtime, the nested HDG with full assembly, which we use to benchmark in Fig. 3(b), is obviously faster than nested HDG with partial assembly due to the cost of operation 23. Even though solutions computed with classical HDG are faster for coarser meshes, as the discretization is refined our implementation of nested HDG with full assembly becomes faster due to the reuse of computations at the block level. The main takeaway from this example is that the nested HDG produces the same solutions as classical HDG while exhibiting significantly lower memory requirements; the differences in computational runtime depend on a myriad of factors, namely the resolution of the 2-D mesh (the interfaces), the amount of unique blocks, the number of processors available and the efficiency of the implementation, to name a few. However, based on the results in this article, we can conclude that the nested HDG has the potential of resolving the linear HDG system faster than via direct solution.

#### 4.2. Triangular nanocoaxial aperture

We now consider a metallic nanostructure that produces extraordinary optical transmission and can excite second harmonic fields. This structure consists of periodic arrays of subwavelength triangular apertures of a dielectric material patterned in a metallic film, and unlike arrays of annular nanogap structures that have been simulated with HDG in previous works [67, 81, 88], triangular apertures are not centrosymmetric, a requirement to excite second-order effects.

##### 4.2.1. Structure definition

The structure that will be analyzed is a gold thin-film with triangular coaxial nanogaps arranged according to the symmetries of the square, see Fig. 4(a), for wavelengths ranging from visible to low infra-red. The metal film is deposited over a sapphire substrate, a transparent material in these frequency regimes, and the nanogap is filled with alumina. The structure is illuminated from below with an  $x$ -polarized plane wave, and we can exploit the symmetries of the structure and solve only for the domain shown in Fig. 4 containing half of the triangular nanogap. Under these symmetry conditions, we prescribe  $\mathbf{E} \times \mathbf{n} = \mathbf{0}$ ,  $\rho = 0$  on the  $x$ -constant boundaries and  $\mathbf{H} \times \mathbf{n} = \mathbf{0}$ ,  $\mathbf{J} \cdot \mathbf{n} = 0$  on the  $y$ -constant boundaries. First-order radiation conditions are imposed on the  $z$ -constant boundaries. To assess the efficacy of SHG, we monitor the transmittance  $\tau$  through the structure and the second harmonic transmittance, computed as

$$\tau_1 = 100 \cdot \frac{\left| \int_{A_o} \Re [\mathbf{E}_1 \times \mathbf{H}_1^*] \cdot \mathbf{n} \, dA \right|}{\left| \int_{A_i} \Re [\mathbf{E}_{\text{inc}} \times \mathbf{H}_{\text{inc}}^*] \cdot \mathbf{n} \, dA \right|}, \quad \tau_2 = 100 \cdot \frac{\left| \int_{A_o} \Re [\mathbf{E}_2 \times \mathbf{H}_2^*] \cdot \mathbf{n} \, dA \right|}{\left| \int_{A_i} \Re [\mathbf{E}_{\text{inc}} \times \mathbf{H}_{\text{inc}}^*] \cdot \mathbf{n} \, dA \right|}, \quad (36)$$

where  $A_i$  is an arbitrary  $xy$  plane below the gold film and  $A_o$  an arbitrary  $xy$  plane above the gold film.

The discretization consists of 18K hexahedral cubic elements, and is constructed by extruding in the  $z$ -direction the 2-D curved mesh in Fig. 4(b). The 2-D curved mesh, with 330 elements, is devised such that the rounded corners are properly represented, and a boundary-layer type discretization is used for the region surrounding the gap. To that end, we place 2-D coaxial layers at distances 0.5,

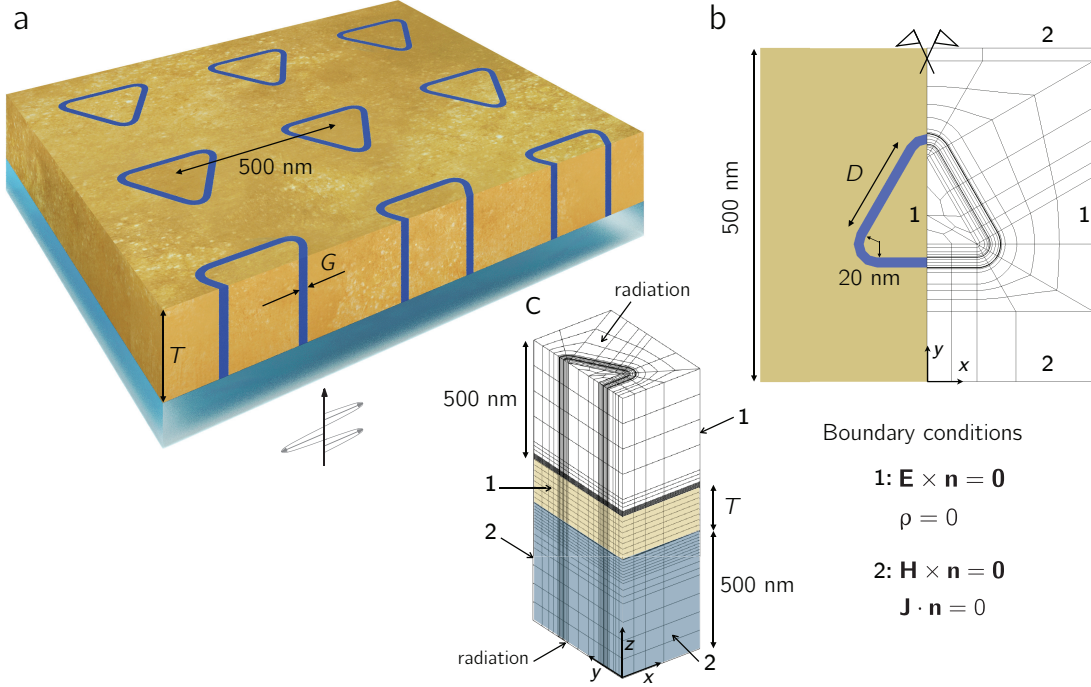


Figure 4: Triangular coaxial nanostructure: (a) 3-D model of periodic array with relevant measures. (b) Cross-sectional view and 2-D curved mesh of unit cell, with symmetry axis and boundary condition specification. (c) 3-D computational domain of unit cell.

1, 2, 3 and 5 nm on both sides of the gap-metal interfaces, ensuring enough resolution for both first and second order phenomena. In the vertical direction, we set both the substrate and superstrate thickness to 500 nm which is sufficient to properly represent illumination conditions and domain unboundedness. For each stratum, we use 19 blocks divided among 4 types, increasing the thickness of each block type as we move further away from the gold film. This computational strategy allows us to capture the rapidly-varying near-field effects in the vicinity of the metal surface ( $< 5$  nm) as well as to smoothly transition towards the far-field values of transmittance. For the gold stratum, we use one thin block type for the regions near the metal-substrate and metal-superstrate interfaces (4 blocks of 0.25 nm thickness for both upper and lower areas) to capture the boundary-layer features that develop and one thick block type for the rest (8 blocks). Hence, the total number of blocks is 54, split into 10 type blocks and 7 type interfaces, see Fig. 4(c). The mesh topology we have described is the same for the different geometric parameters discussed below, and since the hydrodynamic density profile is independent of the gap size we use the same values of 2-D coaxial layers and 3-D boundary-layer detail at the upper and lower metal surfaces for all gaps. Numerical accuracy is verified by carrying out grid convergence studies on consecutively refined meshes, until the relative error for the SHG transmittance is below 1%. This highly anisotropic mesh, along with the nested HDG method, allows us to efficiently solve for the full 3-D EM wave field.

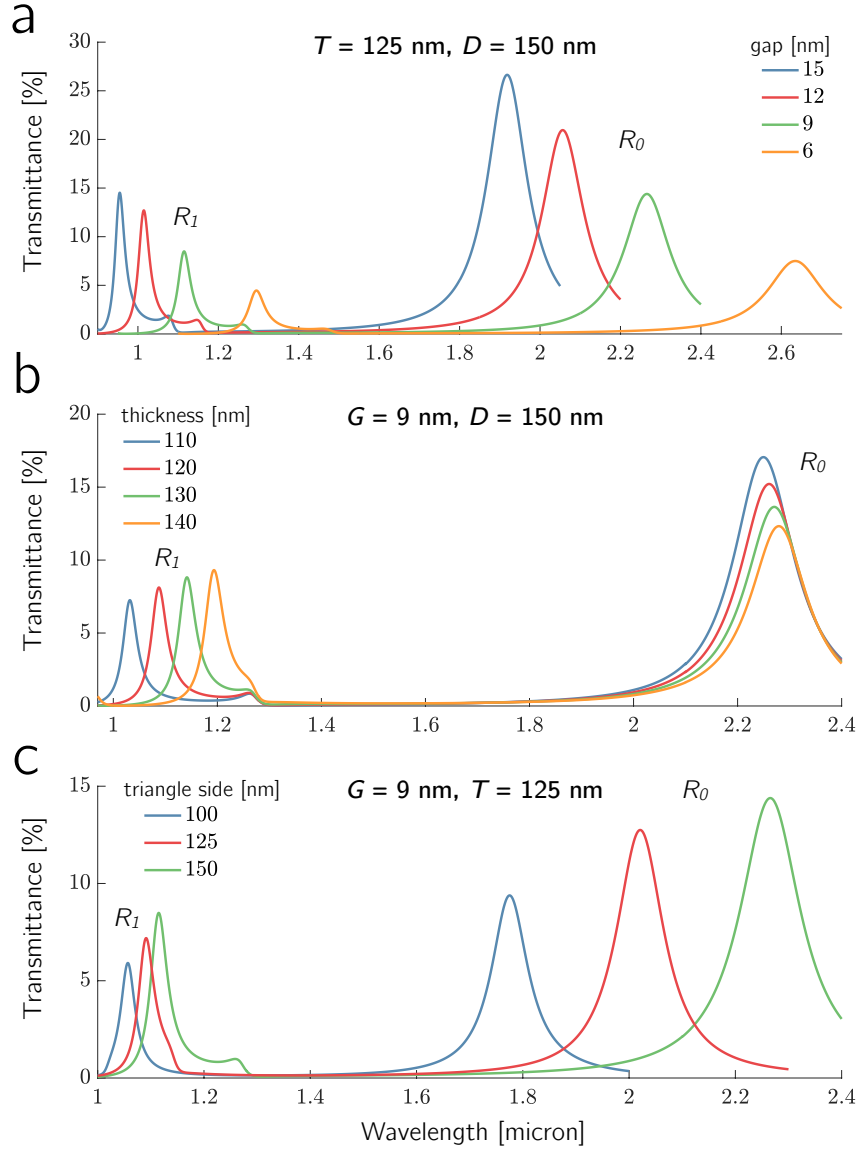


Figure 5: Wavelength [ $\mu\text{m}$ ] - transmittance [%] curves (x-axes shared between subfigures): (a) Effect of decreasing gap. (b) Effect of increasing thickness. (c) Effect of increasing triangle side length.

#### 4.2.2. Optimal geometry

For the structure under consideration, we set the array periodicity to 500 nm and the radius of curvature at the triangle corners to 20 nm for fabrication purposes. The remaining geometric features, namely the gap size  $G$ , the triangle side  $D$  and the film thickness  $T$  need to be numerically determined so as to achieve a doubly-resonant structure. The idea behind such a structure is that resonances are excited at both  $\omega$  and  $2\omega$ , thus amplifying the second-order effects that are generated

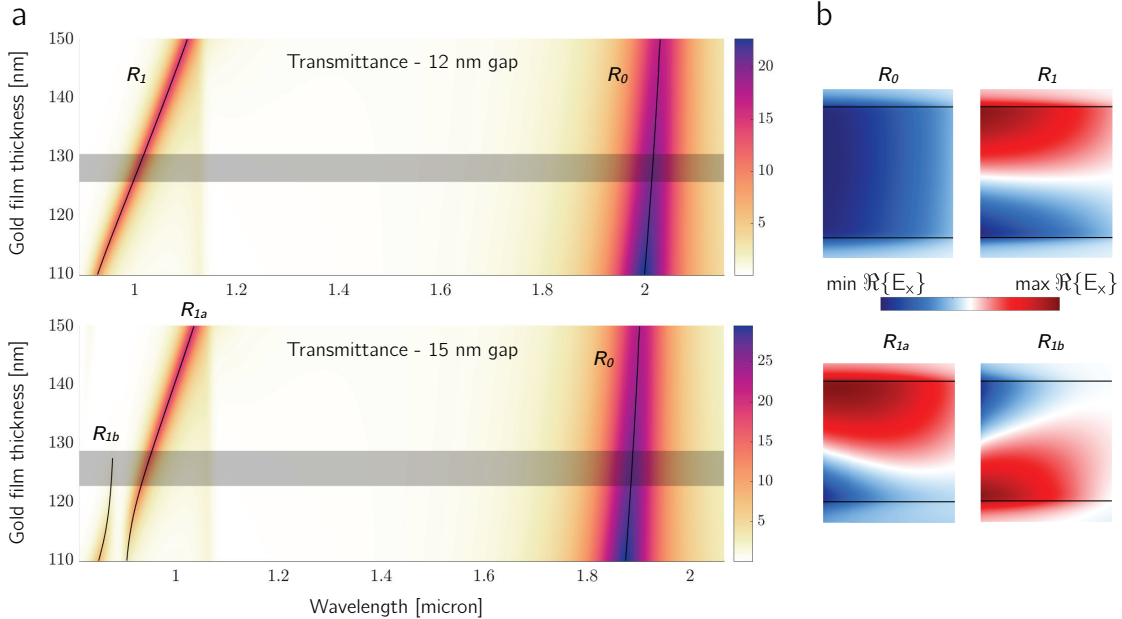


Figure 6: (a) Transmittance [%]-wavelength [ $\mu\text{m}$ ]-thickness [nm] heatmaps for 12 nm (top) and 15 nm gap (bottom), with lines to visually aid resonance tracking and gray shaded area to identify thicknesses that lead to doubly-resonant structures. (b) Real part of  $E_x$  for the different resonances along the mid-gap diagonally-vertical plane. Black lines correspond to upper and lower film surfaces.

as a consequence of its non-centrosymmetric nature.

Firstly, we need to understand the effect of  $G$ ,  $D$  and  $T$  on the resonances. In order to alleviate the computational burden of this parametric study, we perform the simulations using the Drude model (by setting  $\beta = 0$  in (15)) instead of the hydrodynamic model; in the mid-infrared the Drude model predicts a transmittance spectrum that is qualitatively identical to that of the hydrodynamic (only red-shifted), hence it suffices to understand the impact of the geometry parameters. Furthermore, the high-resolution mesh defined above is specifically tailored to capture SHG, hence we may use a coarser mesh just for these geometry simulations since the Drude model does not solve for the Ångstrom-thin accumulation charge layers at the metal-dielectric interfaces. To that end, we use a mesh similar to that of Fig. 4(b,c), but with only 112 2-D elements for a total of 2K hexahedral cubic elements.

For a single wavelength, simulating the full electromagnetic response using the Drude model on the coarse mesh requires solving a sparse linear system of size 190K and 62M non-zeros (1.5GB of RAM), which takes 9 minutes to solve using MATLAB’s backslash operation. We simulate the full spectra of transmittance (0.9-2.8 micron) for several gaps, triangle side lengths and film thicknesses. The impact of gap is shown in Fig. 5(a) for  $T = 125$  nm and  $D = 150$  nm, the impact of thickness is shown in Fig. 5(b) for  $G = 9$  nm and  $D = 150$  nm and the impact of triangle side length is shown in Fig. 5(c) for  $T = 125$  nm and  $G = 9$  nm. The wavelength  $\lambda_0^*$  of resonance  $R_0$  is blue-shifted as  $D$  decreases, whereas it is almost insensitive to the thickness. Conversely, the wavelength  $\lambda_1^*$  of resonance  $R_1$  is much more sensitive to changes in  $T$  (blue-shift for decreasing thickness)

than to changes in  $D$ . Finally, both resonances are similarly affected by gap size modifications. Consequently, for a given gap size one should fix either  $D$  (resp.  $T$ ) and vary  $T$  (resp.  $D$ ) to attain a geometric configuration that is doubly-resonant at frequencies  $\omega$  and  $2\omega$ .

For this structure, we choose to fix  $D$  and optimize the double resonance as  $T$  varies, and leverage our previous work to compute parametrized solutions of plasmonic structures [82], whereby a small number of high-fidelity simulations can be used to construct an accurate reduced order model (ROM) that enables the inexpensive computation of approximate solutions. The reduced order model is constructed upon two parameters: the incident wavelength, to obtain spectrum profiles; and the film thickness, to evaluate how variations in the film thickness impact the resonances. To achieve a parametric representation of the thickness, we build a mapping using  $\mathcal{C}^2$  splines that prescribes deformations in the z-direction, thus ensuring that thickness variations starting from a reference thickness value may be accommodated and are continuous and differentiable. Further details on how to parametrize geometry in plasmonics using deformation mappings may be found in [82].

For each gap size of interest, we set the triangle side to  $D = 150$  nm and build a ROM by first computing 200 high-fidelity simulations (Drude model on the coarse mesh) and then combining these solutions, or snapshots, to form a low-dimensional approximation space, see [82]. This is commonly known as offline stage, which is computationally intensive (each of the 200 solutions takes 9 minutes for a total of 30 hours) but done only once. After completing this stage, the main advantage of ROMs is that they can be queried for any value of  $\lambda \in [0.8, 22]$  micron and  $T \in [110, 150]$  nm –these are the prescribed intervals of interest for the triangular coax– and produce an approximate full-wave 3-D solution of (15) in less than 0.1 seconds. This multi-query process, known as the online stage, will obviously exhibit lower accuracy since instead of the high-fidelity solver we employ a surrogate model. However, for the reduced order models under consideration we report relative errors in transmittance of less than 5% when comparing the ROM solution to the high-fidelity HDG solution, hence the ROM are a suitable computational tool to study the impact of thickness in the resonances of this triangular coaxial structure.

In this case, the burden of the offline stage is greatly compensated by the efficiency of the online stage, since transmittance-wavelength-thickness heatmaps can be obtained by inexpensively querying the ROM for multiple  $(\lambda, T)$  combinations, which would otherwise require a full HDG 3-D simulation for each  $(\lambda, T)$ . These heatmaps are paramount to track the resonances as a function of the thickness and to identify, for each gap, the metal film thickness that gives rise to a doubly-resonant structure. We show the transmittance heatmaps for 12 and 15 nm nanogaps with  $D = 150$  nm in Fig 6, where we notice that the  $R_1$  resonance splits between two for 15 nm gaps and above, whereas it remains a single resonance for gaps below 15 nm. A field plot of the real part of the  $x$ -electric amplitude is provided in Fig. 6, where the field is shown along the mid-gap diagonal plane –that is, the vertical plane that runs along the middle of the gap and is parallel to the longer side of the triangle as shown in Fig. 5(b)– where it can be observed that the  $R_1$  splitting gives rise to modes  $R_{1a}$ ,  $R_{1b}$  that are not constant along the diagonal direction, as opposed the  $R_1$  mode. The shaded gray area corresponds to thickness values that lead to double resonances.

#### 4.2.3. Second-harmonic simulations

Once we have identified, for each gap, the values of  $D$  and  $T$  that excite modes at  $\omega$  and  $2\omega$ , we can apply the computational strategy summarized in (15)-(16) on the fine mesh and with the hydrodynamic model ( $\beta > 0$ ) to compute SHG, where we solve the global HDG linear systems by



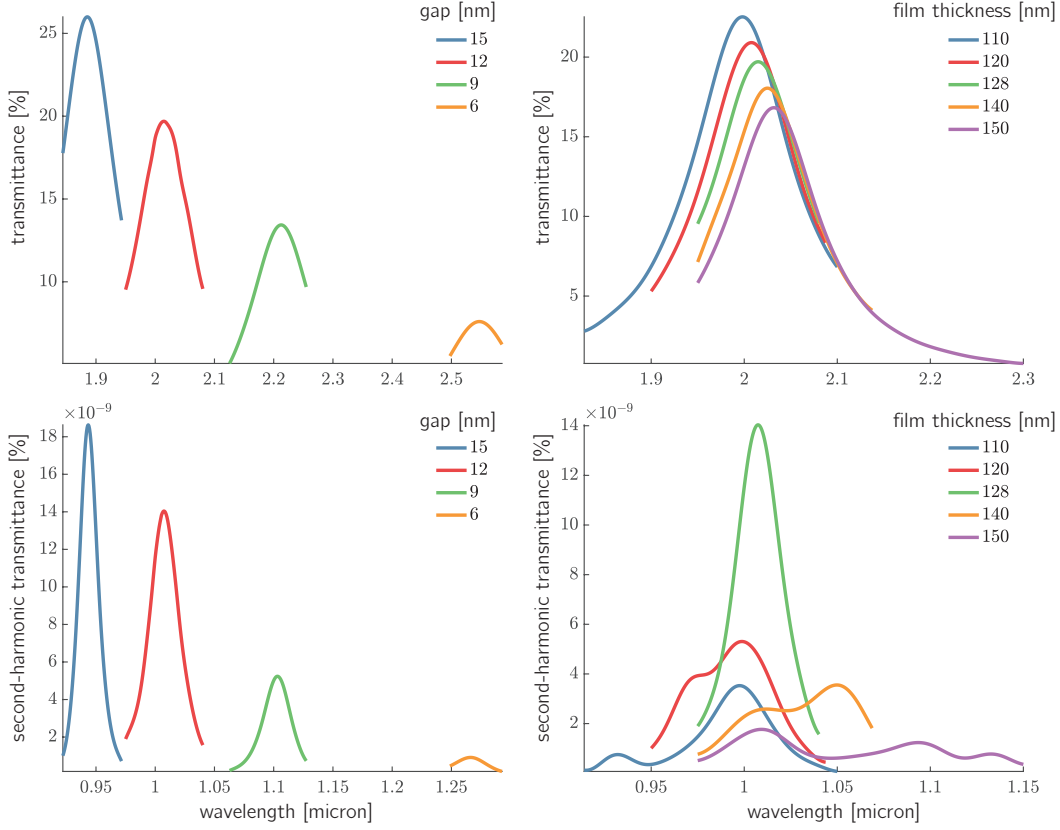


Figure 7: First and second harmonic transmittance for  $D = 150$  nm. (a,b) Transmittance profile  $\tau_1$  (a) and  $\tau_2$  (b)-wavelength [ $\mu\text{m}$ ] for 6, 9, 12 and 15 nm gap and optimal thickness to realize double resonances. (c,d) Transmittance profile  $\tau_1$  (c) and  $\tau_2$  (d)-wavelength [ $\mu\text{m}$ ] for the 12 nm gap and several film thicknesses. Optimal thickness leading to double resonance is 128 nm.

means of Algorithm 1. In order to simulate the spectra shown in Fig. 7, for each gap size and film thickness we solve (15)-(16) for 24 different wavelength values. These simulations are expensive, taking around 19 hours –9.5 hours for each (15) and (16)– per wavelength, for a grand total of 19 days of nonstop computation to recover the first and second harmonic spectrum for a given gap and thickness. We now discuss the breakdown of simulation costs of either (15) or (16) for one wavelength into operations as per Algorithm 1, using the nested HDG with partial assembly and purging the lists  $v_2, v_5$  of unnecessary computations after each iteration. The LU decomposition (operation 4) takes 25 min, solving the linear system (operations 5, 8, 10, 12) takes 250 min, forming and operating the full matrices that result from the second hybridization (operations 16, 17, 20) take 282 min and the  $\{\mathbf{W}, \mathbf{V}\}$  recovery (operation 23) takes 3 min. In terms of storage, the partial assembly of  $\mathcal{Y}$  requires 0.5 GB, the LU decompositions in  $v_1$  requires 13 GB for all types,  $v_2$  and  $v_3$  combined require a maximum of 29 GB,  $v_4$  and  $v_5$  combined require a maximum of 28 GB and finally  $v_6$  requires 128 GB. The storage requirement at any given algorithm step does not exceed 150 GB thanks to the type blocks and interfaces definition and since the information on  $v_2, \dots, v_5$

can be eliminated as the algorithm progresses. However, the storage for  $v_1, v_6$  keeps increasing throughout the main loop, and it can only be deleted after it has been used to recover  $\{\mathbf{W}, \mathbf{V}\}$  in operation 23. Unfortunately, no specific cost comparison can be drawn with classical HDG because the direct solution of the fully assembled sparse matrix exceeds the RAM capacity of our machine (512 GB). If fully assembled, the HDG linear system  $\Upsilon$  is of dimension  $2M$ , with 750M non-zeros for a total of 18 GB in storage.

The bottleneck for the nested HDG method is the discretization of the 2-D mesh rather than the amount of blocks in the third dimension. The resolution of the 2-D mesh will dictate the amount of degrees of freedom at the interfaces, which is directly related to the amount of forcing terms when solving the linear systems (operations 5, 8, 10), as well as to the dimension of the matrices  $\alpha_n$  that need to be inverted to solve the tridiagonal system, which are no longer sparse. A possible strategy to mitigate the computational costs is to increase the number of processors, since operations 5, 8, 10 are embarrassingly parallel with respect to the number of forcing terms. However, further research is needed to reduce the computational burden of the forward Gaussian elimination of the block tridiagonal matrix (Thomas algorithm), summarized in operations 16, 17, 20.

For all simulations, we choose an illumination intensity of  $100 \text{ MW/cm}^2$ , which corresponds to a reference magnetic field  $\alpha = 7.29 \cdot 10^4 \text{ A/m}$ , and a reference lengthscale  $L_c = 10^{-9} \text{ m}$ . The values for gold optical constants are  $\epsilon_\infty = 1$ ,  $\hbar\bar{\omega}_p = 8.45 \text{ eV}$ ,  $\hbar\bar{\gamma} = 0.047 \text{ eV}$  [65], the Fermi velocity  $\bar{v}_F = 1.39 \cdot 10^6 \text{ m/s}$ , the equilibrium charge density  $\bar{n}_0 = 5.91 \cdot 10^{28} \text{ m}^{-3}$  [3] and finally the electron charge  $\bar{e} = 1.062 \cdot 10^{-19} \text{ C}$ . The dielectric constant values for alumina as a function of the incident wavelength are given by [9] for  $\lambda < 1.5 \text{ micron}$  and [43] for  $\lambda > 1.5 \text{ micron}$ , whereas the permittivity of sapphire is taken from [56].

The first and second order transmittances  $\tau_1, \tau_2$  are shown in Figs. 7(a) and (b) for 6, 9, 12 and 15 nm gap. Setting  $D = 150 \text{ nm}$ , we capitalize on the ROMs constructed in Section 4.2.2 and find that thickness values of 128.5, 129, 128 and 126 nm lead to double resonances for these gaps, respectively. Even though the ROM is constructed for the Drude model only, the doubly-resonant thickness is still valid for the nonlocal calculations, since the blue-shift introduced by the hydrodynamic model depends only on the gap width and triangle side length. Stronger second-harmonic transmittance correlates with stronger first harmonic transmittance, suggesting that ultranarrow gaps (below 10 nm) may not be the best candidate structures to observe SHG, despite exhibiting larger field enhancements due to the increased confinement.

In order to highlight the importance of having a doubly resonant structure for enhanced SHG, we compute for  $D = 150 \text{ nm}$  and  $G = 12 \text{ nm}$  the transmittance profile for several film thicknesses, namely 110, 120, 128, 140 and 150 nm, shown in Figs. 7(c) and (d). For this gap and triangle side, thicknesses of approximately 128 nm are the ones that guarantee double resonances, as shown in Fig. 6(a) (top), thus we can expect the second harmonic transmittance  $\tau_2$  to peak for this geometric configuration. Indeed, when comparing the peak transmittance of 120 and 128 nm thicknesses, we see that despite exhibiting lower first harmonic transmittance  $\tau_1^{128}/\tau_1^{120} = 0.95$ , the 128 nm thickness attains a second harmonic transmittance nearly four times larger than that of 120 nm  $\tau_2^{128}/\tau_2^{120} = 3.97$ . This boosting in second harmonic transmittance is a consequence of an optimal geometric configuration that excites resonances at precisely  $\omega$  and  $2\omega$ . The field enhancement for this optimal geometry at the resonant wavelength is shown in Fig. 8, where the extreme confinement and boundary-layer structure of both the first and second harmonic at the curved metal-alumina interface can be appreciated.

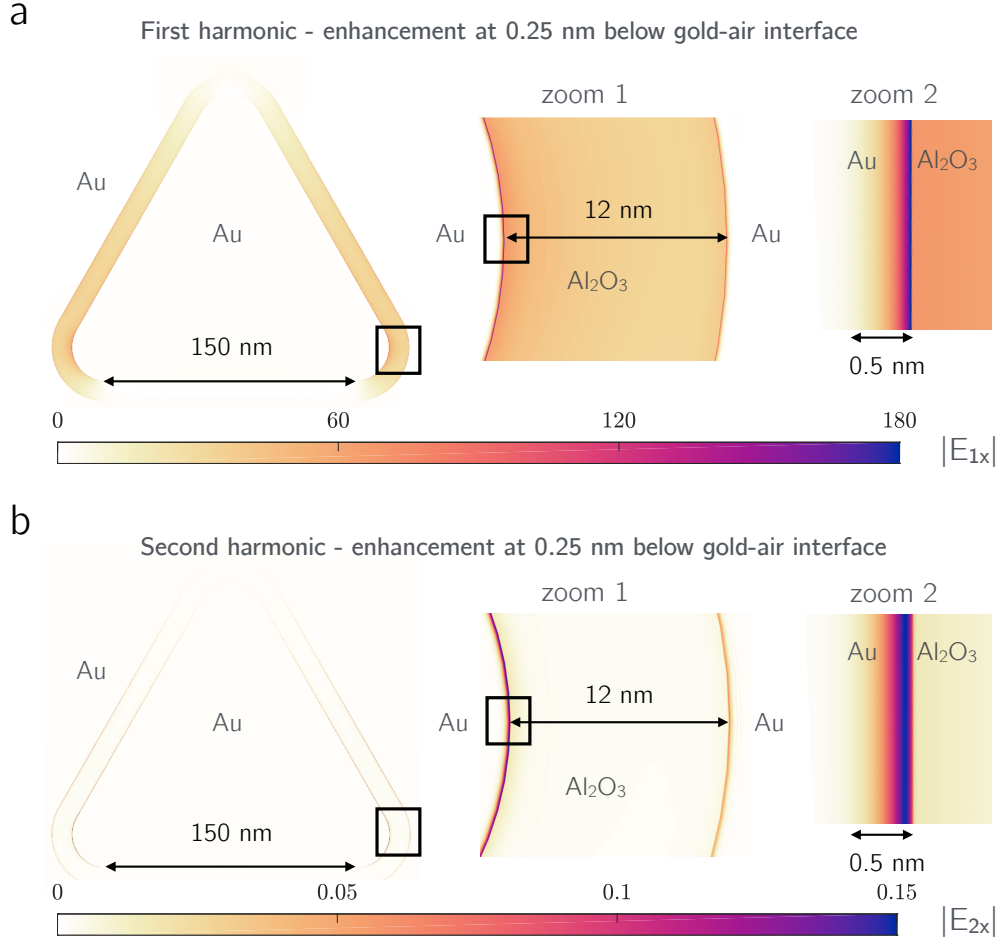


Figure 8: Cross-section image, with two zooms to show detail at the rounded vertex, of  $x$ -component of electric field for  $D = 150$  nm,  $G = 12$  nm and  $T = 128$  nm (double resonant structure) at resonant wavelength  $\lambda = 2.009$  micron, computed 0.25 nm below the gold-air interface. Black squares indicate area of zoom. (a) First harmonic  $|E_{1x}|$  (b) Second harmonic  $|E_{2x}|$ .

## 5. Conclusions

The hybridizable discontinuous Galerkin method for Maxwell's equations augmented with the hydrodynamic model for metals is specially suited to simulate nonlinear plasmonics phenomena, owing to its high-order accuracy and its ability to handle the very large disparity in length scales and the extreme localization of electromagnetic fields. For complex structures the required spatial discretization gives rise to a system that cannot be directly solved due to storage limitations. On the other hand we have not found a robust and effective iterative algorithm to solve the large indefinite HDG linear systems. In this article, we have presented a computational strategy to efficiently solve linear systems of equations that arise from the HDG method by performing a nested hybridization. In computational terms, we consider discretizations that results form the extrusion of 2-D meshes

and substitute one single large linear system solve for multiple smaller linear systems that stem from partitioning the original mesh into non-overlapping blocks of mesh elements following the extrusion direction. Furthermore, a judicious block partition enables us to reuse computations, thus making the nested HDG more efficient than classical HDG both from the storage and the computational perspective.

### Acknowledgements

F. V.-C., N.- C. N and J. P. acknowledge support from the AFOSR Grant No. FA9550-19-1-0240. S.-H.O. acknowledge support from the NSF Grant No. ECCS 1809240 and ECCS 1809723. F. V.-C. acknowledges Vimworks for the design of Fig. 4 (a).

### Bibliography

- [1] Alberti, J., Linnenbank, H., Linden, S., Grynko, Y., Förstner, J., 2016. The role of electromagnetic interactions in second harmonic generation from plasmonic metamaterials. *Applied Physics B* 122 (2), 45.
- [2] Aouani, H., Navarro-Cia, M., Rahmani, M., Sidiropoulos, T. P., Hong, M., Oulton, R. F., Maier, S. A., 2012. Multiresonant broadband optical antennas as efficient tunable nanosources of second harmonic light. *Nano Letters* 12 (9), 4997–5002.
- [3] Ashcroft, N. W., Mermin, N. D., 1976. *Solid State Physics*. Holt, Rinehart and Winston.
- [4] Bachelier, G., Russier-Antoine, I., Benichou, E., Jonin, C., Brevet, P.-F., 2008. Multipolar second-harmonic generation in noble metal nanoparticles. *JOSA B* 25 (6), 955–960.
- [5] Berenger, J.-P., 1994. A perfectly matched layer for the absorption of electromagnetic waves. *Journal of Computational Physics* 114 (2), 185–200.
- [6] Berthelot, J., Bachelier, G., Song, M., Rai, P., Des Francs, G. C., Dereux, A., Bouhelier, A., 2012. Silencing and enhancement of second-harmonic generation in optical gap antennas. *Optics Express* 20 (10), 10498–10508.
- [7] Boardman, A., Paranjape, B., Nakamura, Y. O., 1976. Surface plasmon-polaritons in a spatially dispersive inhomogeneous medium. *Physica Status Solidi (b)* 75 (1), 347–359.
- [8] Boardman, A. D., 1982. *Electromagnetic surface modes*. John Wiley & Sons.
- [9] Boidin, R., Halenkovič, T., Nazabal, V., Beneš, L., Němec, P., 2016. Pulsed laser deposited alumina thin films. *Ceramics International* 42 (1), 1177–1182.
- [10] Bossavit, A., 1990. Solving Maxwell equations in a closed cavity, and the question of 'spurious modes'. *IEEE Transactions on Magnetics* 26 (2), 702–705.
- [11] Boyd, R. W., 2003. *Nonlinear Optics*. Elsevier.
- [12] Busch, K., König, M., Niegemann, J., 2011. Discontinuous Galerkin methods in nanophotonics. *Laser & Photonics Reviews* 5 (6), 773–809.

- [13] Butet, J., Brevet, P.-F., Martin, O. J., 2015. Optical second harmonic generation in plasmonic nanostructures: from fundamental principles to advanced applications. *ACS Nano* 9 (11), 10545–10562.
- [14] Canfield, B. K., Husu, H., Laukkanen, J., Bai, B., Kuittinen, M., Turunen, J., Kauranen, M., May 2007. Local field asymmetry drives second-harmonic generation in noncentrosymmetric nanodimers. *Nano Letters* 7 (5), 1251–1255.
- [15] Carletti, L., Locatelli, A., Stepanenko, O., Leo, G., De Angelis, C., 2015. Enhanced second-harmonic generation from magnetic resonance in AlGaAs nanoantennas. *Optics Express* 23 (20), 26544–26550.
- [16] Celebrano, M., Wu, X., Baselli, M., Großmann, S., Biagioni, P., Locatelli, A., De Angelis, C., Cerullo, G., Osellame, R., Hecht, B., et al., 2015. Mode matching in multiresonant plasmonic nanoantennas for enhanced second harmonic generation. *Nature Nanotechnology* 10 (5), 412.
- [17] Chandrasekar, R., Emani, N. K., Lagutchev, A., Shalaev, V. M., Ciraci, C., Smith, D. R., Kildishev, A. V., 2015. Second harmonic generation with plasmonic metasurfaces: direct comparison of electric and magnetic resonances. *Optical Materials Express* 5 (11), 2682–2691.
- [18] Chen, X., Park, H.-R., Pelton, M., Piao, X., Lindquist, N. C., Im, H., Kim, Y. J., Ahn, J. S., Ahn, K. J., Park, N., Kim, D.-S., Oh, S.-H., 2013. Atomic layer lithography of wafer-scale nanogap arrays for extreme confinement of electromagnetic waves. *Nature Communications* 4, 2361.
- [19] Ciraci, C., Hill, R., Mock, J., Urzhumov, Y., Fernández-Domínguez, A., Maier, S., Pendry, J., Chilkoti, A., Smith, D., 2012. Probing the ultimate limits of plasmonic enhancement. *Science* 337 (6098), 1072–1074.
- [20] Ciraci, C., Pendry, J. B., Smith, D. R., 2013. Hydrodynamic model for plasmonics: a macroscopic approach to a microscopic problem. *ChemPhysChem* 14 (6), 1109–1116.
- [21] Ciraci, C., Poutrina, E., Scalora, M., Smith, D. R., 2012. Second-harmonic generation in metallic nanoparticles: Clarification of the role of the surface. *Physical Review B* 86 (11), 115451.
- [22] Cockburn, B., Dong, B., Guzmán, J., 2008. A superconvergent LDG-hybridizable Galerkin method for second-order elliptic problems. *Mathematics of Computation* 77 (264), 1887–1916.
- [23] Cockburn, B., Gopalakrishnan, J., Lazarov, R., 2009. Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems. *SIAM Journal on Numerical Analysis* 47 (2), 1319–1365.
- [24] Cockburn, B., Guzmán, J., Wang, H., 2009. Superconvergent discontinuous Galerkin methods for second-order elliptic problems. *Mathematics of Computation* 78 (265), 1–24.
- [25] Cockburn, B., Shu, C.-W., 1998. The local discontinuous Galerkin method for time-dependent convection-diffusion systems. *SIAM Journal on Numerical Analysis* 35 (6), 2440–2463.
- [26] COMSOL Inc, 1986. COMSOL Multiphysics Modeling Software. <http://www.comsol.com/>.

- [27] Crouseilles, N., Hervieux, P.-A., Manfredi, G., 2008. Quantum hydrodynamic model for the nonlinear electron dynamics in thin metal films. *Physical Review B* 78 (15), 155412.
- [28] Drude, P., 1900. Zur elektronentheorie der metalle. *Annalen der Physik* 306 (3), 566–613.
- [29] Eftang, J. L., Patera, A. T., 2013. Port reduction in parametrized component static condensation: approximation and a posteriori error estimation. *International Journal for Numerical Methods in Engineering* 96 (5), 269–302.
- [30] Eguiluz, A., Quinn, J. J., Aug 1976. Hydrodynamic model for surface plasmons in metals and degenerate semiconductors. *Physical Review B* 14, 1347–1361.
- [31] Fang, M., Huang, Z., Sha, W. E., Xiong, X. Y., Wu, X., 2016. Full hydrodynamic model of nonlinear electromagnetic response in metallic metamaterials. *Progress In Electromagnetics Research* 157, 63–78.
- [32] Ginzburg, P., Krasavin, A. V., Wurtz, G. A., Zayats, A. V., 2015. Nonperturbative hydrodynamic model for multiple harmonics generation in metallic nanostructures. *ACS Photonics* 2 (1), 8–13.
- [33] Grynko, Y., Förstner, J., 2017. Simulation of second harmonic generation from photonic nanostructures using the discontinuous Galerkin time domain method. In: *Recent Trends in Computational Photonics*. Springer, pp. 261–284.
- [34] Hesthaven, J. S., Warburton, T., 2002. Nodal high-order methods on unstructured grids: I. Time-domain solution of Maxwell’s equations. *Journal of Computational Physics* 181 (1), 186–221.
- [35] Hille, A., Moferdt, M., Wolff, C., Matyssek, C., Rodríguez-Oliveros, R., Prohm, C., Niegemann, J., Grafström, S., Eng, L. M., Busch, K., 2016. Second harmonic generation from metal nano-particle resonators: Numerical analysis on the basis of the hydrodynamic drude model. *The Journal of Physical Chemistry C* 120 (2), 1163–1169.
- [36] Huynh, D. B. P., Knezevic, D. J., Patera, A. T., 2013. A static condensation reduced basis element method: approximation and a posteriori error estimation. *ESAIM: Mathematical Modelling and Numerical Analysis* 47 (01), 213–251.
- [37] Huynh, D. B. P., Knezevic, D. J., Patera, A. T., 2013. A static condensation reduced basis element method: complex problems. *Computer Methods in Applied Mechanics and Engineering* 259, 197–216.
- [38] Jha, S., 1965. Theory of optical harmonic generation at a metal surface. *Physical Review* 140 (6A), A2020.
- [39] Ji, X., Cai, W., Zhang, P., 2007. High-order DGTD methods for dispersive Maxwell’s equations and modelling of silver nanowire coupling. *International Journal for Numerical Methods in Engineering* 69 (2), 308–325.
- [40] Jin, J.-M., 2015. *The finite element method in electromagnetics*. John Wiley & Sons.
- [41] Johnson, S. G., 2008. Notes on perfectly matched layers (PMLs). Lecture notes, Massachusetts Institute of Technology, Massachusetts 5 (5.3), 2.

- [42] Kauranen, M., Zayats, A. V., 2012. Nonlinear plasmonics. *Nature Photonics* 6 (11), 737.
- [43] Kischkat, J., Peters, S., Gruska, B., Semtsiv, M., Chashnikova, M., Klinkmüller, M., Fedosenko, O., Machulik, S., Aleksandrova, A., Monastyrskiy, G., et al., 2012. Mid-infrared optical properties of thin films of aluminum oxide, titanium dioxide, silicon dioxide, aluminum nitride, and silicon nitride. *Applied Optics* 51 (28), 6789–6798.
- [44] Klein, M. W., Enkrich, C., Wegener, M., Linden, S., 2006. Second-harmonic generation from magnetic metamaterials. *Science* 313 (5786), 502–504.
- [45] Kolkowski, R., Szeszko, J., Dwir, B., Kapon, E., Zyss, J., 2016. Non-centrosymmetric plasmonic crystals for second-harmonic generation with controlled anisotropy and enhancement. *Laser & Photonics Reviews* 10 (2), 287–298.
- [46] Krasavin, A. V., Ginzburg, P., Zayats, A. V., 2018. Free-electron optical nonlinearities in plasmonic nanostructures: A review of the hydrodynamic description. *Laser & Photonics Reviews* 12 (1), 1700082.
- [47] Kullock, R., Hille, A., Haußmann, A., Grafström, S., Eng, L. M., 2011. SHG simulations of plasmonic nanoparticles using curved elements. *Optics Express* 19 (15), 14426–14436.
- [48] Kunz, K. S., Luebbers, R. J., 1993. *The finite difference time domain method for electromagnetics*. CRC press.
- [49] Lanteri, S., Scheid, C., 2013. Convergence of a discontinuous Galerkin scheme for the mixed time-domain Maxwell’s equations in dispersive media. *IMA Journal of Numerical Analysis* 33 (2), 432–459.
- [50] Li, L., Lanteri, S., Mortensen, N. A., Wubs, M., 2017. A hybridizable discontinuous Galerkin method for solving nonlocal optical response models. *Computer Physics Communications* 219, 99–107.
- [51] Li, L., Lanteri, S., Perrussel, R., 2013. A hybridizable discontinuous galerkin method for solving 3d time-harmonic maxwell’s equations. In: *Numerical Mathematics and Advanced Applications 2011*. Springer, pp. 119–128.
- [52] Linden, S., Niesler, F., Förstner, J., Grynko, Y., Meier, T., Wegener, M., 2012. Collective effects in second-harmonic generation from split-ring-resonator arrays. *Physical Review Letters* 109 (1), 015502.
- [53] Lindhard, J., 1954. On the properties of a gas of charged particles. *Dan. Vid. Selsk Mat.-Fys. Medd.* 28, 8.
- [54] Liu, J., Brio, M., Zeng, Y., Zakharian, A. R., Hoyer, W., Koch, S. W., Moloney, J. V., 2010. Generalization of the FDTD algorithm for simulations of hydrodynamic nonlinear Drude model. *Journal of Computational Physics* 229 (17), 5921–5932.
- [55] Lu, T., Zhang, P., Cai, W., 2004. Discontinuous Galerkin methods for dispersive and lossy Maxwell’s equations and PML boundary conditions. *Journal of Computational Physics* 200 (2), 549–580.

- [56] Malitson, I., Dodge, M., 1972. Refractive-index and birefringence of synthetic sapphire. In: *Journal Of The Optical Society Of America*. Vol. 62. pp. 1405–1405.
- [57] McMahon, J. M., Gray, S. K., Schatz, G. C., 2010. Calculating nonlocal optical properties of structures with arbitrary shape. *Physical Review B* 82 (3), 035423.
- [58] Moeferd, M., Kiel, T., Sproll, T., Intravaia, F., Busch, K., 2018. Plasmonic modes in nanowire dimers: A study based on the hydrodynamic Drude model including nonlocal and nonlinear effects. *Physical Review B* 97 (7), 075431.
- [59] Mur, G., 1981. Absorbing boundary conditions for the finite-difference approximation of the time-domain electromagnetic-field equations. *Electromagnetic Compatibility, IEEE Transactions on* (4), 377–382.
- [60] Nédélec, J.-C., 1980. Mixed finite elements in  $\mathbb{R}^3$ . *Numerische Mathematik* 35 (3), 315–341.
- [61] Nguyen, N.-C., Peraire, J., Cockburn, B., 2009. An implicit high-order hybridizable discontinuous Galerkin method for linear convection-diffusion equations. *Journal of Computational Physics* 228 (9), 3232–3254.
- [62] Nguyen, N.-C., Peraire, J., Cockburn, B., 2011. High-order implicit hybridizable discontinuous Galerkin methods for acoustics and elastodynamics. *Journal of Computational Physics* 230 (10), 3695–3718.
- [63] Nguyen, N.-C., Peraire, J., Cockburn, B., 2011. Hybridizable discontinuous Galerkin methods for the time-harmonic Maxwell’s equations. *Journal of Computational Physics* 230 (19), 7151–7175.
- [64] Niegemann, J., Pernice, W., Busch, K., 2009. Simulation of optical resonators using DGTD and FDTD. *Journal of Optics A: Pure and Applied Optics* 11 (11), 114015.
- [65] Olmon, R. L., Slovick, B., Johnson, T. W., Shelton, D., Oh, S.-H., Boreman, G. D., Raschke, M. B., 2012. Optical dielectric function of gold. *Physical Review B* 86 (23), 235147.
- [66] Panoiu, N., Sha, W., Lei, D., Li, G., 2018. Nonlinear optics in plasmonic nanostructures. *Journal of Optics* 20 (8), 083001.
- [67] Park, H.-R., Chen, X., Nguyen, N.-C., Peraire, J., Oh, S.-H., 2015. Nanogap-enhanced terahertz sensing of 1 nm thick ( $\lambda/10^6$ ) dielectric films. *ACS Photonics* 2 (3), 417–424.
- [68] Pitarke, J., Silkin, V., Chulkov, E., Echenique, P., 2006. Theory of surface plasmons and surface-plasmon polaritons. *Reports on Progress in Physics* 70 (1), 1–87.
- [69] Razdolski, I., Chen, Y., Giles, A. J., Gewinner, S., Schöllkopf, W., Hong, M., Wolf, M., Gianini, V., Caldwell, J. D., Maier, S. A., et al., 2016. Resonant enhancement of second-harmonic generation in the mid-infrared using localized surface phonon polaritons in subdiffractive nanostructures. *Nano Letters* 16 (11), 6954–6959.
- [70] Saà-Seoane, J., Nguyen, N.-C., Men, H., Freund, R., Peraire, J., 2012. Binary optimization techniques for linear PDE-governed material design. *Applied Physics A* 109 (4), 1023–1030.



- [71] Scalora, M., Vincenti, M. A., de Ceglia, D., Roppo, V., Centini, M., Akozbek, N., Bloemer, M. J., Oct. 2010. Second- and third-harmonic generation in metal-based structures. *Physical Review A* 82 (4), 043828.
- [72] Schmitt, N., Scheid, C., Lanteri, S., Moreau, A., Viquerat, J., 2016. A DGTD method for the numerical modeling of the interaction of light with nanometer scale metallic structures taking into account non-local dispersion effects. *Journal of Computational Physics* 316, 396–415.
- [73] Sipe, J. E., So, V., Fukui, M., Stegeman, G., 1980. Analysis of second-harmonic generation at metal surfaces. *Physical Review B* 21 (10), 4389–4402.
- [74] Smirnova, D., Kivshar, Y. S., 2016. Multipolar nonlinear nanophotonics. *Optica* 3 (11), 1241–1255.
- [75] Sommerfeld, A., 1949. *Partial differential equations in physics*. Vol. 1. Academic Press.
- [76] Taflove, A., Hagness, S. C., 2005. *Computational electrodynamics: the finite-difference time-domain method*. Artech house.
- [77] Toscano, G., Raza, S., Jauho, A.-P., Mortensen, N. A., Wubs, M., 2012. Modified field enhancement and extinction by plasmonic nanowire dimers due to nonlocal response. *Optics Express* 20 (4), 4176–4188.
- [78] Toscano, G., Straubel, J., Kwiatkowski, A., Rockstuhl, C., Evers, F., Xu, H., Mortensen, N. A., Wubs, M., 2015. Resonance shifts and spill-out effects in self-consistent hydrodynamic nanoplasmonics. *Nature Communications* 6 (1), 1–11.
- [79] Vidal-Codina, F., 2017. *Simulation methods for plasmonic structures*. Ph.D. thesis, Massachusetts Institute of Technology.
- [80] Vidal-Codina, F., Martín-Moreno, L., Ciraci, C., Yoo, D., Nguyen, N.-C., Oh, S.-H., Paire, J., 2020. Terahertz and infrared nonlocality and field saturation in extreme-scale nanoslits. *Opt. Express* 28 (6), 8701–8715.
- [81] Vidal-Codina, F., Nguyen, N.-C., Oh, S.-H., Paire, J., 2018. A hybridizable discontinuous Galerkin method for computing nonlocal electromagnetic effects in three-dimensional metallic nanostructures. *Journal of Computational Physics* 355, 548–565.
- [82] Vidal-Codina, F., Nguyen, N.-C., Paire, J., 2018. Computing parametrized solutions for plasmonic nanogap structures. *Journal of Computational Physics* 366, 89–106.
- [83] Vidal-Codina, F., Saà-Seoane, J., Nguyen, N.-C., Paire, J., 2019. A multiscale continuous galerkin method for stochastic simulation and robust design of photonic crystals. *Journal of Computational Physics: X* 2, 100016.
- [84] von Cube, F., Irsen, S., Diehl, R., Niegemann, J., Busch, K., Linden, S., 2013. From isolated metaatoms to photonic metamaterials: evolution of the plasmonic near-field. *Nano Letters* 13 (2), 703–708.
- [85] Wang, F. X., Rodríguez, F. J., Albers, W. M., Ahorinta, R., Sipe, J. E., Kauranen, M., Dec. 2009. Surface and bulk contributions to the second-order nonlinear optical response of a gold film. *Physical Review B* 80 (23), 233402.

- [86] Yee, K. S., 1966. Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. *IEEE Transactions on Antennas and Propagation* 14 (3), 302–307.
- [87] Yoo, D., Mohr, D. A., Vidal-Codina, F., John-Herpin, A., Jo, M., Kim, S., Matson, J., Caldwell, J. D., Jeon, H., Nguyen, N.-C., et al., 2018. High-contrast infrared absorption spectroscopy via mass-produced coaxial zero-mode resonators with sub-10 nm gaps. *Nano Letters* 18 (3), 1930–1936.
- [88] Yoo, D., Nguyen, N.-C., Martin-Moreno, L., Mohr, D. A., Carretero-Palacios, S., Shaver, J., Peraire, J., Ebbesen, T. W., Oh, S.-H., 2016. High-throughput fabrication of resonant metamaterials with ultrasmall coaxial apertures via atomic layer lithography. *Nano Letters* 16 (3), 2040–2046.
- [89] Yoo, D., Vidal-Codina, F., Ciraci, C., Nguyen, N.-C., Smith, D. R., Peraire, J., Oh, S.-H., 2019. Modeling and observation of mid-infrared nonlocality in effective epsilon-near-zero ultranarrow coaxial apertures. *Nature Communications* 10, 4476.
- [90] Zeng, Y., Hoyer, W., Liu, J., Koch, S. W., Moloney, J. V., 2009. Classical theory for second-harmonic generation from metallic nanoparticles. *Physical Review B* 79 (23), 235109.
- [91] Zhang, Y., Grady, N. K., Ayala-Orozco, C., Halas, N. J., 2011. Three-dimensional nanostructures as highly efficient generators of second harmonic light. *Nano Letters* 11 (12), 5519–5523.
- [92] Zhu, W., Esteban, R., Borisov, A. G., Baumberg, J. J., Nordlander, P., Lezec, H. J., Aizpurua, J., Crozier, K. B., 2016. Quantum mechanical effects in plasmonic structures with subnanometre gaps. *Nature Communications* 7, 11495.