UNIVERSITY OF CALIFORNIA, SAN DIEGO

Plasmonic and Near-Field Phenomena in Low-Dimensional Nanostructures

A Dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Physics

by

Bor-Yuan Jiang

Committee in charge:

Professor Michael M. Fogler, Chair
Professor Dimitri N. Basov
Professor Michael J. Holst
Professor Zhaowei Liu
Professor Congjun Wu

2018

The Dissertation of Bor-Yuan Jiang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2018

EPIGRAPH

"The most important step a man can take

is the next one."

—Brandon Sanderson

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my advisor Prof. Michael Fogler, who has been unfailingly patient in guiding me toward the completion of my degree. Your brilliance and sense of humor will always be an inspiration to me. I would also like to thank Prof. Dimitri Basov, it was a privilege to work in close collaboration with your group and have access to data on the cutting-edge of near-field optics. I would like to thank my parents for their unconditional support during the past six years. Without you I would not have had the peace of mind to carry out my studies. Thanks to all my coauthors, especially Dr. Guang-Xin Ni, Prof. Eugene Mele, and Zach Addison, your input was instrumental in the successful publication of our work. Without your contributions I would not have made it to the high-profile journals as we did. Thanks also to Dr. Jhih-Shing Wu and Zhiyuan Sun for the many useful discussions we shared. Finally, I would like to thank all my friends in the U.S., you made my working hours less lonely and my leisure time enjoyable. Special thanks to Evan Chang for always making time for me when I'm back at Taiwan.

Chapter 2, in full, is a reprint of the material as it appears in Journal of Applied Physics: Jiang, B.-Y.; Zhang, L. M.; Castro Neto, A. H.; Basov, D. N.; Fogler, M. M. "Generalized spectral method for near-field optical microscopy", J. Appl. Phys. 119, 054305 (2016). The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the materials as they appear in Physical Review Letters: Jiang, B.-Y.; Ni, G.; Pan, C.; Fei, Z.; Cheng, B.; Lau, C.; Bockrath, M.; Basov, D.; Fogler, M. "Tunable Plasmonic Reflection by Bound 1D Electron States in a 2D Dirac Metal", Phys. Rev. Lett. 117, 086801 (2016), and in Physical Review B: Jiang, B.-Y.; Fogler, M. M. "Electronic response of graphene to linelike charge perturbations". Phys. Rev. B 91, 235422 (2015). The dissertation author was the primary investigator and author of these papers.

Chapter 4, in full, is a reprint of the material as it appears in Nano Letters: Jiang, B.-Y.; Ni, G.; Addison, Z.; Shi, J. K.; Liu, X; Zhao, S. Y. F.; Kim, P.; Mele, E. J.; Basov, D. N.; Fogler, M. M. "Plasmon reflections by topological electronic boundaries in bilayer graphene", Nano Lett.

17, 7080 (2017). The dissertation author was the primary investigator and author of this paper.

VITA

2008	B. S. in Physics, National Taiwan University
2018	Ph. D. in Physics, University of California, San Diego

PUBLICATIONS

Jiang, B.-Y.; Mele, E. J.; Fogler, M. M. "Theory of plasmon reflection by a 1D junction", in submission.

Jiang, B.-Y.*; Ni, G. X.*; Addison, Z.; Shi, J. K.; Liu, X; Zhao, S. Y. F.; Kim, P.; Mele, E. J.; Basov, D. N.; Fogler, M. M. "Plasmon reflections by topological electronic boundaries in bilayer graphene", Nano Lett. 17, 7080 (2017).

Fei, Z.; Ni, G. X.; **Jiang, B.-Y.**; Fogler, M. M.; Basov, D. N. "Nanoplasmonic Phenomena at Electronic Boundaries in Graphene", ACS Photon. (2017)

Jiang, B.-Y.*; Ni, G. X.*; Pan, C.; Fei, Z.; Cheng, B.; Lau, C.; Bockrath, M.; Basov, D.; Fogler, M. "Tunable Plasmonic Reflection by Bound 1D Electron States in a 2D Dirac Metal", Phys. Rev. Lett. 117, 086801 (2016).

Jiang, B.-Y.; Zhang, L. M.; Castro Neto, A. H.; Basov, D. N.; Fogler, M. M. "Generalized spectral method for near-field optical microscopy", J. Appl. Phys. 119, 054305 (2016).

Jiang, B.-Y.; Fogler, M. M. "Electronic response of graphene to linelike charge perturbations". Phys. Rev. B 91, 235422 (2015).

Stinson, H. T.; Wu, J. S.; **Jiang, B.-Y.**; Fei, Z.; Rodin, A. S.; Chapler, B. C.; McLeod, A. S.; Castro Neto, A.; Lee, Y. S.; Fogler, M. M.; Basov, D. N. "Infrared nanospectroscopy and imaging of collective superfluid excitations in anisotropic superconductors", Phys. Rev. B 90, 014502 (2014).

**These authors contributed equally to this work.*

ABSTRACT OF THE DISSERTATION

Plasmonic and Near-Field Phenomena in Low-Dimensional Nanostructures

by

Bor-Yuan Jiang

Doctor of Philosophy in Physics

University of California, San Diego, 2018

Professor Michael M. Fogler, Chair

Plasmonics aims to combine the advantages of nanometer scale electronics with the high operating frequency (terahertz and beyond) of photonics. Control of plasmon propagation can be achieved in a two-dimensional electron gas (2DEG) by tuning the electronic properties of the 1D nanostructures it contains, which act as scatters for plasmons . Plasmonic response of these nanostructures, however, happens on a length scale much smaller than the wavelength of free space electromagnetic radiation and cannot be studied with conventional optical microscopy. Instead, we resolve these nanoscopic phenomena using near-field optical microscopy, which has a spatial resolution of $\sim 20\,\mathrm{nm}$. In this dissertation, we first describe the working principles of near-field optical microscopy, then analyze the plasmonic phenomena we observed around several

1D nanostructures, including a potential well in monolayer graphene, domain walls in bilayer graphene, and a low-conductivity gap in a 2DEG.

In Chapter 1, we give an overview of the basic properties of surface plasmons and graphene, followed by a brief explanation of the operating principles of near-field optical microscopy.

In Chapter 2, we study theoretically the electromagnetic interaction between a subwavelength particle (the 'probe') and a material surface (the 'sample'). The interaction is shown to be governed by a series of resonances corresponding to surface polariton modes localized near the probe. The resonance parameters depend on the dielectric function and geometry of the probe, as well as the surface reflectivity of the material. Calculation of such resonances is carried out for several types of axisymmetric probes: spherical, spheroidal, and pear-shaped. For spheroids an efficient numerical method is developed, capable of handling cases of large or strongly momentum-dependent surface reflectivity. Application of the method to highly resonant materials such as aluminum oxide (by itself or covered with graphene) reveals a rich structure of multi-peak spectra and nonmonotonic approach curves, i.e., the probe-sample distance dependence. These features also strongly depend on the probe shape and optical constants of the model. For less resonant materials such as silicon oxide, the dependence is weak, so that the spheroidal model is reliable. The calculations are done within the quasistatic approximation with radiative damping included perturbatively.

In Chapter 3, we show that surface plasmons of a two-dimensional Dirac metal such as graphene can be reflected by line-like perturbations hosting one-dimensional electron states. The reflection originates from a strong enhancement of the local optical conductivity caused by optical transitions involving these bound states. We propose that the bound states can be systematically created, controlled, and liquidated by an ultranarrow electrostatic gate. Using infrared nanoimaging, we obtain experimental evidence for the locally enhanced conductivity of graphene induced by a carbon nanotube gate, which supports this theoretical concept.

In Chapter 4, we show that topological bound states confined to the domain walls in bilayer graphene are the source of the wall's strong coupling to surface plasmons observed in infrared nanoimaging experiments. These domain walls separate regions of AB and BA interlayer stacking and have attracted attention as novel examples of structural solitons, topological electronic boundaries, and nanoscale plasmonic scatterers. The optical transitions among the topological chiral modes and the band continua enhance the local conductivity, which leads to plasmon reflection by the domain walls. The imaging reveals two kinds of plasmonic standing-wave interference patterns, which we attribute to shear and tensile domain walls. We compute the electronic structure of both wall varieties and show that the tensile wall contains additional confined bands which produce a structure-specific contrast of the local conductivity, in agreement with the experiment. The coupling between the confined modes and the surface plasmon scattering unveiled in this work is expected to be common to other topological electronic boundaries found in van der Waals materials. This coupling provides a qualitatively new pathway toward controlling plasmons in nanostructures.

In Chapter 5, we present a comprehensive study of the reflection of normally incident plasmon waves from a low-conductivity 1D junction in a 2D conductive sheet. Rigorous analytical results are derived in the limits of wide and narrow junctions. Two types of phenomena determine the reflectance, the cavity resonances within the junction and the capacitive coupling between the leads. The resonances give rise to alternating strong and weak reflection but are vulnerable to plasmonic damping. The capacitive coupling, which is immune to damping, induces a near perfect plasmon reflection in junctions narrower than $1/10$ of the plasmon wavelength. Our results are important for 2D plasmonic circuits utilizing slot antennas, split gates or nanowire gates. They are also relevant for the implementation of nanoscale terahertz detectors, where optimal light absorption coincides with the maximal junction reflectance.

Chapter 1

Overview and Introduction

After decades of development, electronics – the transfer and manipulation of information via electrons – is approaching its limits, both in its physical dimensions (nanometer scale) and in its bandwidth (GHz frequency). On the other hand, photonic devices operate at a much higher bandwidth than their electronic counterparts but have a much larger size (micrometer scale), making it difficult to integrate them directly into electronic circuitry. Plasmonics bridges the gap between electronics and photonics by employing surface plasmons, which are charge density oscillations in a metal at an interface with an insulator. These waves oscillate at THz frequencies and beyond but have a wavelength orders of magnitudes smaller than light of the same frequency in free space, potentially facilitating nanoscale information transfer at a much larger bandwidth than current electronics.[Ozb06]

A promising plasmonic material is graphene, a two-dimensional (2D) crystal of carbon atoms in a honeycomb lattice. The advantages of graphene include high plasmonic confinement, wide range of operating frequencies, long plasmon lifetimes, and tunability of the plasmon wavelength.[GPN12, FRA$^+$12] Long plasmon propagation distances and high quality factors have been demonstrated on graphene in recent experiments,[WLG$^+$14, NWG$^+$16] showing that graphene-based plasmonic circuitry is indeed feasible. Before a plasmonic circuit can be realized,

however, one must find a method to control the flow of plasmons in the circuit. The ideal control element would be a switch that can turn the flow on or off, i.e., it can be tuned to either completely transmit or completely reflect plasmons. To this end, we study the plasmonic response of several natural and artificial 1D nanostructures, whose properties can be electrically tuned. In general, objects with dimensions much smaller than the wavelength, such as the 1D nanostructures we study, can only weakly reflect the wave. However, several of the structures we studied exhibited surprisingly strong plasmon reflection that can be tuned, making them potentially useful for nanoplasmonics.

1.1 Surface plasmons

Surface plasmons are waves in the charge density at the interface of a conductor and an insulator, e.g., a metal sheet and vacuum. A longitudinal, or transverse magnetic (TM) surface plasmon wave can be imagined as alternating stripes of positive and negative charges on the surface of the metal that propagate in the direction perpendicular to the stripes, Fig. 1.1(a). These waves have a wavelength that is much smaller than the free-space wavelength of light at the same frequency, so they can be described by electrostatics without having to consider retardation effects. The entire excitation, including the charge density oscillations and the electromagnetic fields they produce outside the surface, is called a surface plasmon polariton, which we abbreviate as plasmon below.

For a 2DEG that has an optical conductivity which is local, $\sigma(r) = \sigma$, the electric and magnetic fields of a surface plasmon wave, propagating in the x-direction, are given by

$$E_x = e^{i(q_p x - \omega t)} e^{-q_p |z|}, \quad E_z = iE_x, \quad H_y = \frac{2\pi}{c}\sigma E_x. \tag{1.1}$$

These fields oscillate sinusoidally in the 2DEG plane while decaying exponentially away from the plane, see Fig. 1.1(b). The wave is characterized by a single quantity, the plasmon momentum

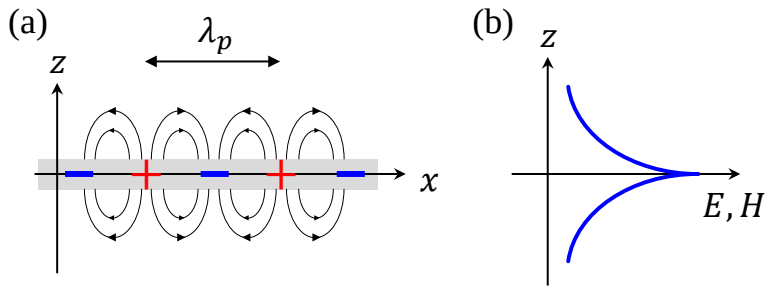

Figure 1.1: (a) Schematic electric field distribution (black curves) of a surface plasmon on a 2DEG (gray). (b) Field distribution in the out of plane direction.

q_p, which can be parametrized as

$$q_p = \frac{2\pi}{\lambda_p}(1 + i\gamma), \tag{1.2}$$

where λ_p is the plasmon wavelength and γ is the damping factor that indicates how fast the plasmon decays. As can be seen from the $e^{-q_p|z|}$ factor in Eq. (1.1), the fields are confined to the 2DEG on the scale of λ_p in the out of plane direction, so a smaller λ_p corresponds to better field confinement.

The plasmon momentum q_p emerges from the dielectric function ε of the 2DEG, which can be simply found through Maxwell's equations (see Section 5.8.1),

$$\varepsilon(q) = 1 - \frac{q}{q_p}. \tag{1.3}$$

The zero of the dielectric function ε correspond to the momentum $q = q_p$ where free-standing waves (i.e., the surface plasmons) are present. The value of the plasmon momentum

$$q_p = \frac{i\kappa\omega}{2\pi\sigma} \tag{1.4}$$

is determined by the dielectric constant κ of the material encasing the 2DEG, the frequency ω, and the conductivity σ of the 2DEG. Comparing with Eq. (1.2), one can see that the plasmon wavelength λ_p is primarily determined by $\mathrm{Im}\,\sigma$, while the damping factor γ is determined by $\mathrm{Re}\,\sigma$.

For a 2DEG with a simple Drude-like conductivity,

$$\sigma_D = \frac{iD}{\pi(\omega + i\nu)}, \tag{1.5}$$

the wavelength λ_p is directly proportional to the Drude weight D, which in turn is proportional to the carrier density of the 2DEG. On the other hand, the damping γ is determined by the phenomenological damping rate ν of the 2DEG, whose value is related to impurities, electron-phonon, and electron-electron interactions in the 2DEG.[WLG+14]

1.2 Plasmonic properties of graphene

One of the properties that make graphene an attractive plasmonic platform is the tunability of its plasmon wavelength through electrostatic gating. A typical gate consists of graphene on top of a metal or semiconductor block separated by a dielectric spacer, with a bias voltage applied between graphene and the metal. By tuning the bias, the chemical potential and hence the plasmon wavelength of graphene can be adjusted. This can be seen from the Drude-like conductivity of graphene,

$$\sigma = \frac{ie^2}{\pi\hbar^2}\frac{\mu}{\omega + i\nu}, \tag{1.6}$$

valid at low momentum and low frequency $\hbar v_F q, \hbar\omega \ll \mu$, where the conductivity σ is directly proportional to the chemical potential μ and inversely proportional to the plasmon momentum q_p, $\sigma \propto \mu \propto q_p^{-1}$. This allows global control of the plasmonic property of the entire graphene sheet. Using patterned electrical gates, split gates, or nanowire gates, one can even make different regions of graphene have different plasmon wavelengths. Plasmons can then be made to scatter between the regions in a controlled fashion, making graphene incredibly versatile in terms of possible plasmonic applications. In this dissertation we study the physics of plasmonic interaction with the simplest pattern – a 1D gate. In Chapters 3 and 5 we study 1D nanowire gates that can

strongly reflect plasmons, due to either bound electron states in the potential well created by the gate, or to the capacitive coupling of either side of a low-conductivity gap.

1.3 Near-field optical microscopy

The surface plasmons and the nanostructures on the 2DEG are both deeply-subwavelength objects and cannot be resolved using conventional optical microscopy. Further, there is a large momentum mismatch, $q_p \gg \omega/c$, between the surface plasmons and light in vacuum. This means that the plasmons cannot be directly excited by shining light on a featureless 2DEG, and the plasmons launched by a nanostructure on a 2DEG cannot be optically detected. To efficiently couple photons to surface plasmons, an optical antenna must used. An optical antenna is a metallic object of dimensions comparable to the wavelength of free space radiation, but also contains sharp protrusions on the scale of the wavelength of surface plasmons. This allows the antenna to couple to electromagnetic waves on both length scales and convert photons into plasmons or vice versa. Additionally, charges on the antenna will be highly concentrated at the protrusions due to the "lightning-rod" effect of metals,[MKG+14] further enhancing the coupling of photons to surface plasmons.

In infrared (IR) scatter-type scanning near-field optical microscopy (s-SNOM), the antenna is the metallic tip in an atomic force microscope (AFM). The shaft of the tip resembles an inverted pyramid (see Fig. 1.2) with a height on the order of tens of μm, while the apex is very sharp with a radius of curvature on the order of a few tens of nm. When IR light is shone on the tip it couples to the tip by polarizing the shaft. Meanwhile the sharp apex produces localized fields that can excite surface plasmons, if the tip were close to a 2DEG. Conversely, the strength of the surface plasmon wave propagating in the vicinity of the apex significantly alters the apex charge concentration and subsequently the dipole moment of the entire tip, so that the light radiated by the tip reflects the amplitude of the plasmon wave at the tip location. Thus, in s-SNOM, the AFM

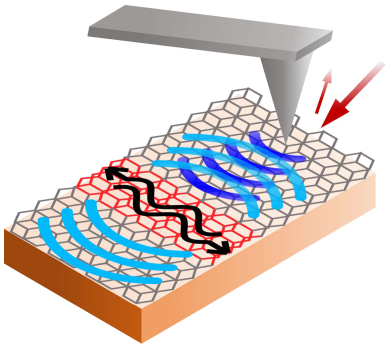

Figure 1.2: Schematic diagram of the s-SNOM. The tip resembles an inverted pyramid attached to a cantilever. Surface plasmons are launched by the tip apex under an external electromagnetic field. They are reflected by 1D nanostructures on the 2DEG, and the amplitude of the combined waves are detected by measuring the backscattered light from the tip.

tip acts as both a launcher and a detector of surface plasmons. By scanning the tip across the sample and measuring the backscattered light at each location, one obtains an image of the local optical property of the sample, at a spatial resolution on the order of the apex radius ($\sim 20\,\text{nm}$).

In our studies of the 1D nanostructures, the plasmons launched by the tip propagate radially away and are subsequently reflected by the nanostructures. The intensity of the total electric field underneath the tip is determined by the interference of the launched and reflected plasmon waves. The amplitude of this interference term oscillates as a function of the distance between the tip and the reflector, with the period equal to one-half of the plasmon wavelength. By analyzing this interference pattern the plasmon reflection coefficient of the nanostructure can be deduced.

Chapter 2

Generalized spectral method for near-field optical microscopy

2.1 Introduction

The problem of electromagnetic interaction between a material surface and a small external particle is fundamental to numerous physical phenomena and spectroscopic techniques, including surface-enhanced Raman scattering, surface fluorescence, adsorbed molecules spectroscopy, and near-field microscopy. From the point of view of electromagnetic theory, it is a special kind of scattering problem where the scatterer resides in a uniform half-space, e.g., vacuum, while the effect of the other half-space — the sample — is represented by the surface reflectivity $r_\alpha(q,\omega)$. The reflectivity may depend on the in-plane momentum q, frequency ω, and polarization $\alpha = \mathrm{P}$ or S. Far-field optics describes the regime $q < \omega/c$. Momenta $q \gg \omega/c$, which correspond to in-plane distances $\Delta\rho$ much smaller than the diameter c/ω of Wheeler's radian sphere, [Whe59] are the domain of near-field optics.

This work is motivated by recent advancements of the scattering-type near-field optical microscopy [KH04, ABJR12] (s-SNOM), which has become one of the leading tools

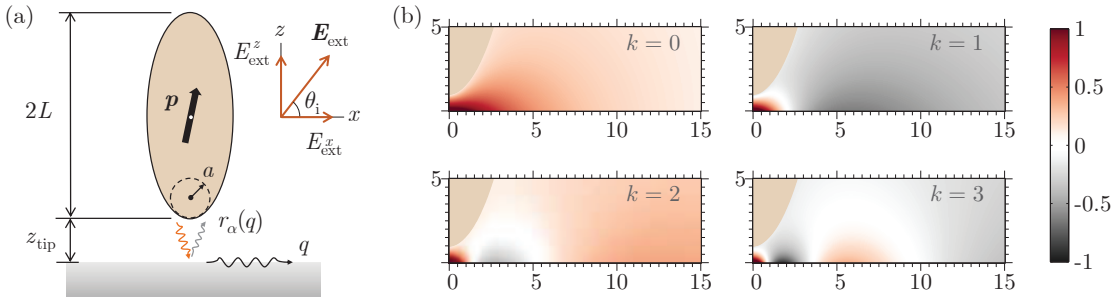

Figure 2.1: (Color online) (a) Schematics of an s-SNOM experiment in which a polarizable probe is used to examine a sample characterized by the surface reflectivity $r_\alpha(q)$. The external electric field $\boldsymbol{E}_{\rm ext}$ incident on the system creates evanescent waves inside the probe-sample gap. This modifies the dipole moment $\boldsymbol{p}$ of the probe, which is detectable by its far-field radiation. (b) The real-space potential distribution for the first four eigenmodes of the probe polarizability $\chi^\perp$ computed numerically for a spheroidal probe of half-length $L = 25a$. The axes are the x- and z-coordinates in units of a, the curvature radius of the apex of the probe. The probe's location is represented by the uniformly shaded beige area in the upper left corner of each panel.

for measuring optical response of diverse materials on spatial scales as short as 5–20 nm. Thanks to technical improvements and the development of tunable and broad-band infrared sources, [AGK09, FAB+11, AK11a, HSW+11] the s-SNOM has provided insights into properties of complex oxides, [ML06, QBC+07, ZAH+07, QBA+09, FQB+09, JBW+10, LNK+10] organic monolayers, [NKP+09] graphene, and other two-dimensional crystals. [FAB+11, FRA+12, CBAG+12, DFM+14]

The schematics of an s-SNOM experiment is shown in Fig. 2.1(a). A sharp elongated probe is brought into close proximity of a sample and is illuminated by an external electromagnetic wave with electric field $\boldsymbol{E}_{\rm ext}e^{-i\omega t}$. Its interaction with the probe creates scattered waves $e^{i\boldsymbol{q}\boldsymbol{\rho}+ik^z z-i\omega t}$, $\boldsymbol{\rho} = (x,y)$, with arbitrary in-plane momentum $\boldsymbol{q}$, including large-q evanescent waves, $k^z = \sqrt{(\omega/c)^2 - q^2} \simeq iq$. Multiple reflections of these waves inside the probe-sample nanogap cause small but important changes in the total radiating dipole moment $\boldsymbol{p}e^{-i\omega t}$ of the probe. These changes are detected by measuring the far-field scattering signal as a function of the probe

coordinates. This signal is proportional to the probe polarizabilities,

$$\chi^{\perp} \equiv p^z / E^z_{\rm ext}\,, \quad \chi^{\parallel} \equiv p^x / E^x_{\rm ext}\,, \tag{2.1}$$

which have the dimension of volume.

The goal of this paper is to study the properties of functions $\chi^{\perp}$ and $\chi^{\parallel}$. For simplicity, we consider only axisymmetric probes. We are especially interested in probes of large aspect ratio. In the experiment, strongly elongated probes are used because of high longitudinal polarizability $\chi^{\perp}$, which promotes an efficient coupling between evanescent and far-field radiation modes — the "antenna" effect — making the detection of the near-field component possible.

We assume that the length of the probe is much smaller than the diameter of the radian sphere c/ω, so that the scattering problem can be treated within the quasistatic approximation. The probe shape we examine the most is a prolate spheroid. At first glance both of these assumptions are unrealistic because actual probes are not spheroidal and their length (typically, tens of μm) can often exceed c/ω for ω in infrared or optical frequency domain. Yet this model was previously found to yield quantitative agreement with the s-SNOM experimental data for many materials. This apparent agreement can be expected in cases where the surface reflectivity $r_\alpha(q,\omega)$ of the sample is not too large, and the aspect ratio of the probe does not vary greatly from one experiment to the next. Under such conditions the gross features of the s-SNOM scattering amplitude should indeed have only a modest dependence on the exact shape of the probe and other experimental parameters. However, fine details of the scattering amplitude are shape-dependent even in this case [MKG$^+$14] and they may be discerned as the instrumental resolution improves. Furthermore, for samples with high reflectivity, even the gross features become sensitive to the shape and size of the probe. To demonstrate these trends in this paper we study the longitudinal and the transverse polarizabilities in great detail. We will ignore the S-polarization reflectivity $r_{\rm S}(q,\omega)$ because for most materials it becomes very small at $q \gg \omega/c$. Hence, χ^{ν} are functionals of the

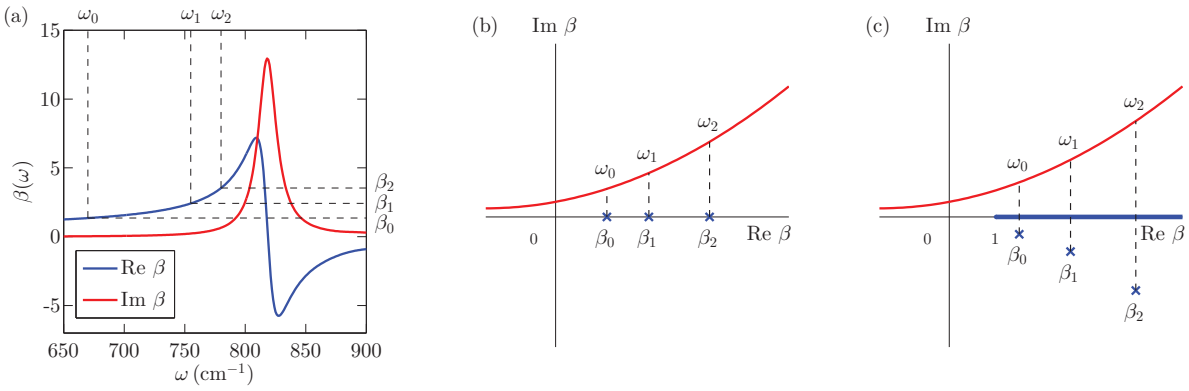

Figure 2.2: (Color online) (a) Near-field reflectivity $\beta(\omega)$ of bulk Al_2O_3 discussed in Sec. 2.6. Whenever the condition $\mathrm{Re}\,\beta(\omega) = \beta_k^{\nu}$ is met, a local maximum appears in $\mathrm{Im}\,\chi^{\nu}$. The frequencies of three such resonances are indicated by the dashed lines. (b) In the complex plane of β, the poles β_k^{ν} lie on the positive real axis, while real materials trace curves in the upper half plane, shown in red. (c) A full electrodynamic treatment predicts that the poles shift into the lower half-plane and an additional nonanalyticity in the form of a branch cut $[1,\infty)$ appears.

remaining reflectivity function $r_{\mathrm{P}}(q,\omega)$ and the probe-sample distance z_{tip}. We show that such functionals can be quite complicated, especially for strongly momentum-dependent reflectivity typical of layered and/or ultrathin materials. Therefore, it is good to start with a simpler case of a *bulk* medium with a q-independent reflectivity

$$\beta(\omega) \equiv r_{\mathrm{P}}(q,\omega)\,, \tag{2.2}$$

so that for a fixed z_{tip} and ω, the probe polarizabilities are functions of a single parameter β.

It should be clarified that while the absolute reflectivity may not exceed unity for the radiative modes $q < \omega/c$ because of energy conservation, for the evanescent ones $q > \omega/c$ it is allowed do so. Large β's are indicative of weakly damped surface modes in a material, e.g., surface phonons in dielectrics or surface plasmons in metals. We use the umbrella term "surface polaritons" for all such modes. The energy loss due to evanescent modes is governed not by $|\beta|$ but by $\mathrm{Im}\,\beta$ which must be nonnegative at $q > \omega/c$. (To compute the losses $\mathrm{Im}\,\beta$ needs to be integrated over q with a weight that depends on the probe-sample distance. [FW84])

In the limit of vanishingly small dissipation, $\mathrm{Im}\,\beta(\omega)$ tends to a δ-function peak at the mode frequency. In practice, $\mathrm{Im}\,\beta$ (and consequently $|\beta|$) as high as 10–20 is possible for well-ordered crystalline solids, e.g., aluminum oxide Al_2O_3 possessing sharp phonon modes [Fig. 2.2(a)]. Therefore, a robust theoretical formalism must be capable of computing functions $\chi^\nu(\beta)$ in the entire upper complex half-plane. To meet this requirement such a formalism must correctly reproduce the analytic properties of functions $\chi^\nu(\beta)$. We adopt a version of the generalized spectral method (GSM) in which the total field outside the probe and sample is decomposed into eigenfunctions of an auxiliary homogeneous problem, and the role of eigenvalues are played by the reflectivity β, the so-called β-method in the terminology of Ref. [AKSV99]. (Similar formalism is also known in the theory of conductivity of heterogeneous media. [Ber78, Ber79]) Following Refs. [AKSV99, Ber78, Ber79], one can show, for the quasistatic case, that for any probe-sample distance $z_{\mathrm{tip}} > 0$ functions $\chi^\nu(\beta)$ are meromorphic. In other words, they admit the series representations

$$\chi^\nu(\beta) = \sum_{k=0}^{\infty} \frac{R_k^\nu}{\beta_k^\nu - \beta}, \quad \nu = \perp \text{ or } \parallel, \tag{2.3}$$

where the sequence of poles β_k^ν has no accumulation points, and so, no upper limit. Additionally, we will show that if the probe is made of an ideal conductor and no other sources of dissipation are present, then the poles $\beta_k^\nu > 1$ and the residues $R_k^\nu > 0$ are real. If the dielectric constant $\varepsilon_{\mathrm{tip}}$ of the probe is considered fixed, R_k^ν and β_k^ν depend only on the geometric factors: the probe shape, size, and its distance z_{tip} to the surface. All these results comply with the general theory of the β-method developed in Ref. [AKSV99].

The poles β_k^ν grow exponentially with k but their ratios with R_k^ν are bounded and satisfy the sum rule

$$\sum_{k=0}^{\infty} \frac{R_k^\nu}{\beta_k^\nu} = \chi_0^\nu. \tag{2.4}$$

Here $\chi_0^\nu \equiv \chi^\nu(\beta = 0)$ is the polarizability of an isolated probe, which does not depend on z_{tip}. These properties ensure convergence of the series (2.3) at any $\beta \neq \beta_k^\nu$. On the other hand, if a

material-specific $\beta(\omega)$ approaches any of β_k^ν, a resonant peak in χ^ν and ultimately, in the near-field signal, would be observed.

The divergence of χ^ν at a given pole implies that a nonzero dipole, i.e., free oscillations may exist in the absence of any external field. Physical intuition about this regime is aided by the method of images, according to which real charges Q_i on the probe interact with their virtual images $-\beta Q_i$ inside the sample and for $\beta > 1$ achieve a runaway positive feedback. However, one must keep in mind that these eigenmodes arise only in the auxiliary problem where the sample is substituted by a fictitious material of reflectivity β_k^ν. The divergence never actually happens in real materials due to their inherent dissipation, which enters in the form of a positive imaginary part in β as shown in Fig. 2.2(a) and 2.2(b). The resonances are further damped due to shifting of the poles β_k^ν to the lower complex half-plane when radiative corrections are considered [Fig. 2.2(c)], as discussed in more detail in Sec. 2.8. For a generic probe that ends in a rounded tip, the amplitude of the eigenmodes is the greatest near the tip, as illustrated in Fig. 2.1(b) for a spheroidal probe. Overall, this physical picture of tip-localized eigenmodes is an elegant and economical approach to understanding the mechanism of probe-sample coupling.

The main objective of the present work is to elucidate the analytical properties of the coefficients β_k^ν and R_k^ν. We focus on the practically interesting case where the probe length L is much larger than the curvature radius a of the probe tip. We show that for such strongly elongated probes three regimes can be distinguished. The first is the short-distance limit $z_{\text{tip}} \ll a$ where the behavior of β_k^ν is universal. We show that it can be derived from the known exact solutions for spherical particles (Sec. 2.2). The second is the long-distance limit, $z_{\text{tip}} \gg L$, where the probe acts as a point-dipole and the functional form of the resonance parameters is again universal. The remaining third regime $a < z_{\text{tip}} < L$ is the most nontrivial one where β_k^ν and R_k^ν depend on the probe shape.

For all the probe geometries we study the poles β_k^ν grows exponentially with k, and so for moderate values of β it is permissible to truncate the series in Eq. (2.3) after one or a few

leading terms. This truncation is effectively done in simplified models [HK00, TKH04, ATdA$^+$08, Oce07, AK11a] of the probe-sample coupling, see Sec. 2.9. However, this simplification may lead to qualitatively and quantitatively wrong results at small $z_{\rm tip}$ and/or for large β. The latter characterize highly polar materials such as SiO_2 [ZAF$^+$12] (a commonly used substrate) and the already mentioned Al_2O_3 (an important reference material of infrared optics).

Besides addressing analytical properties of the probe polarizabilities, we also discuss methods for their numerical computation. For the simplest case of a momentum-independent reflectivity, the calculation can be made virtually instantaneous with the help of Eq. (2.3) once the first few β_k^ν and R_k^ν are computed and stored. For specific case of a spheroidal probe, this calculation can be further accelerated using the spheroidal harmonics basis instead of the standard boundary element method (BEM). Since the number of relevant poles and residues is relatively small, for further convenience, they can be fitted to analytical forms, see an example for $L = 25a$ spheroidal probe in Section 2.11. The speed becomes a crucial consideration if the calculations have to be done repeatedly. An important example is extracting optical constants of the sample from near-field spectroscopy data by curve-fitting algorithms. [MKG$^+$14] One may anticipate to find a considerable speed-up if this inverse problem were treated using the GSM. The acceleration occurs because the unknown physical parameter $\beta = \beta(\omega)$ of the sample and the geometric parameters β_k^ν and R_k^ν of the probe stand clearly separated. The GSM also applies for momentum-dependent $r_{\rm P}(q,\omega)$, e.g., for layered samples; however, in the current implementation the speed-up compared to the BEM is less significant.

The remainder of the article is organized as follows. In Sec. 2.2 we analyze the universal aspects of the short- and the long-distance regimes. In Sec. 2.3 the spheroidal probe model is considered. The equations for the poles and residues are presented and the results of their numerical solution for the case of a q-independent $r_{\rm P}$ are discussed. In Sec. 2.4 we explore the effects due to a weakly q-dependent surface reflectivity. In Sec. 2.5 we discuss caveats in the simulation of the s-SNOM experiment. In Sec. 2.6 we apply our numerical method to computing

the near-field response of bulk Al_2O_3, a strongly polar material. In Sec. 2.7 we perform the calculation for the same Al_2O_3 substrate but covered with graphene, which is a system with a strongly q-dependent reflectivity. In Sec. 2.8 we discuss the effects of the probe shape and retardation on these calculations. We also do a similar comparison for SiO_2, a less polar material. In Sec. 2.9 we discuss prior theoretical work and close with concluding remarks. Technical details of the derivations and the source code of our computer program are available as the online Supplemental Material for this article.

2.2 Probe-sample interaction in short- and long-distance limits

We start with a qualitative analysis of the short-distance regime defined by the condition $z_{\text{tip}} \ll a$. In this limit the structure of the localized polariton modes can be understood intuitively by analogy [RS81] to electromagnetic modes in an open cavity. The probe-sample gap can be approximated by a cavity with height $z(\rho) \simeq z_{\text{tip}} + (\rho^2/2a)$ gradually increasing as a function of the radial position ρ. For simplicity, let us assume the surface reflectivity of the probe is equal to unity, as for an ideal conductor. To have free oscillations exist in such a cavity the surface reflectivity β of the sample must exceed unity, compensating for the exponential decay of the evanescent waves. The condition of the self-sustained oscillations is $\beta \exp\bigl(2ik^z(\rho)z(\rho)\bigr) = 1$. Accordingly, the local radial momentum $q(\rho) \simeq -ik^z(\rho) = \log\beta\,/\,2z(\rho)$. Imposing the quasiclassical Bohr-Sommerfeld quantization condition $\int_0^\infty d\rho q(\rho) = \pi[k + O(1)]$ for mode number k, we obtain

$$\log\beta_k \simeq [k + O(1)]\sqrt{\frac{8z_{\text{tip}}}{a}}, \quad z_{\text{tip}} \ll a\,. \tag{2.5}$$

The mode is localized at distances $\rho \lesssim \sqrt{z_{\text{tip}}a}$. The validity of this qualitative analysis is supported by the exact results for spherical particles. For the $\nu = \parallel$ part, the following compact formulas for

the poles and residues are available [RS81, AM82, AM83]:

$$\beta_k^{\parallel}(\alpha) = e^{(2k+3)\alpha}, \tag{2.6}$$

$$R_k^{\parallel}(\alpha) = 4(k+1)(k+2)a^3 \sinh^3\alpha, \tag{2.7}$$

where

$$\alpha = \operatorname{arccosh}\left(\frac{z_{\text{tip}}}{a}+1\right). \tag{2.8}$$

It is easy to check that Eqs. (2.5) and (2.6) agree in the limit of small α. (Dependence of $\beta_k^{\perp}$ on α is qualitatively similar; however, the residues scale as $R_k^{\perp} \sim ka^3\alpha^2$ at small α, Section 2.12) It is reasonable to think that the behavior of $\beta_k^{\nu}(\alpha)$ at $\alpha \ll 1$ should be common for any shape ending in a rounded tip. As long as the modes are localized at $\rho \ll a$, they should be affected weakly by the rest of the probe. This hypothesis is supported by numerical calculations presented later in this article.

Consider next the long-distance limit $z_{\text{tip}} \gg L$. In this case the probe-sample interaction can be analyzed using the multipole expansion. For the lowest resonance $k = 0$ it is sufficient to include only the dipole term. The dipole moment of the probe is given by $p^{\nu} = \chi_0^{\nu} E_{\text{tot}}^{\nu}$, where $E_{\text{tot}}^{\nu} = E_{\text{ext}}^{\nu} + E_{\text{ind}}^{\nu}$ is the total field at the probe position and E_{ind}^{ν} is the field induced by the image dipole. In particular, $E_{\text{ind}}^{\perp} = \beta p^{\perp}/4z_{\text{tip}}^3$ and $E_{\text{ind}}^{\parallel} = \beta p^{\parallel}/8z_{\text{tip}}^3$. Solving these equations for p^{ν} and casting the result for $\chi^{\nu} = p^{\nu}/E_{\text{ext}}^{\nu}$ in the form (2.3), we get

$$\beta_0^{\perp} \simeq 4z_{\text{tip}}^3/\chi_0^{\perp}, \qquad R_0^{\perp} \simeq 4z_{\text{tip}}^3, \tag{2.9a}$$

$$\beta_0^{\parallel} \simeq 8z_{\text{tip}}^3/\chi_0^{\parallel}, \qquad R_0^{\parallel} \simeq 8z_{\text{tip}}^3. \tag{2.9b}$$

For the sphere $\chi_0^{\nu} = a^3$, so that the last pair of equations agrees with the exact result (2.6) and (2.7). The $k > 0$ resonances are dominated by higher-order multipoles. The principal dependence

of the poles and residues on α is expected to be the same as for the sphere, i.e.,

$$\beta_k^\nu \sim \left(\frac{z_{\rm tip}}{a}\right)^{2k+3},\ R_k^\nu \sim \frac{2}{c^\nu}(k+1)(k+2)z_{\rm tip}^3 \text{ if } z_{\rm tip} \gg L, \tag{2.10}$$

where $c^\perp = 1$ and $c^\parallel = 1/2$. The forms for R_k^ν are verified numerically in a later section. Equations (2.9a)–(2.10) imply that in the large $z_{\rm tip}$ limit the sum rule (2.4) is saturated by the $k = 0$ mode alone.

The case of a q-dependent reflectivity can be treated similarly. Thus, for $k = 0$ one finds [ATdA+08]

$$\chi^\nu(\omega, z_{\rm tip}) = \frac{\chi_0^\nu}{1 - \chi_0^\nu g^\nu(\omega, z_{\rm tip})}, \tag{2.11}$$

$$g^\nu(\omega, z_{\rm tip}) = c^\nu \int_0^\infty r_{\rm P}(q,\omega) e^{-2qz_{\rm tip}} q^2 dq. \tag{2.12}$$

Note that the integral in Eq. (2.12) is dominated by the in-plane momenta $q \sim 1/z_{\rm tip}$, which we assume to be well outside the light cone, $q \gg \omega/c$. At $z_{\rm tip} > c/\omega$ this condition no longer holds and one has to include retardation effects, see Sec. 2.8.

In summary, in this section we presented arguments that the limiting case formulas (2.5) and (2.9a)–(2.12) apply to perfectly conducting probes of arbitrary shapes. For the sphere $L = a$ and for probes of modest aspect ratio $L \gtrsim a$, these formulas match by the order of magnitude at $z_{\rm tip} \sim a$. However, for strongly elongated probes $L \gg a$ an additional intermediate regime $a \ll z_{\rm tip} \ll L$ exists which requires further study. The simplest example of such a shape is the prolate spheroid and we discuss it in the next section.

2.3 Intermediate distances: Spheroidal probe

Unlike the problem of a sphere, that of a spheroidal probe cannot be solved analytically. However, we can take advantage of the separation of variables in prolate spheroidal coordinates

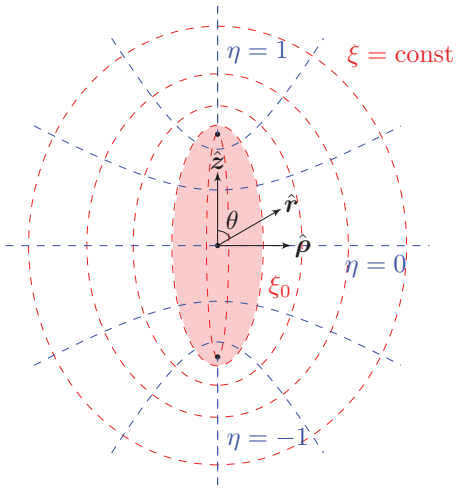

Figure 2.3: The prolate spheroidal coordinate system. Contours of constant ξ (η) are confocal spheroids (hyperboloids). The unit vector $\hat{\boldsymbol{r}}$ and polar angle θ of spherical polar coordinates and unit vectors $\hat{\boldsymbol{\rho}}$ and $\hat{\mathbf{z}}$ of cylindrical polar coordinates are also shown for reference.

(Fig. 2.3), which enables a more efficient numerical solution. [BV87] In this coordinate system the spheroid is a surface of constant $\xi = L/F \equiv \xi_0$. The focal length F, the major semi-axis L, the minor semi-axis W, and the curvature radius a of the spheroid apex are related by

$$F = \sqrt{L^2 - W^2}, \quad a = W^2/L. \tag{2.13}$$

This implies $\xi_0 = [1 - (a/L)]^{-1/2}$. We assume that the major axis of the spheroid is along the z-axis. If the distance between the spheroid and the sample is z_{tip}, the sample surface is at $z = -L - z_{\text{tip}}$.

We consider the quasistatic limit where the scalar potential has the harmonic time dependence $\propto e^{-i\omega t}$. Its spatial part must obey the Laplace equation in the domain outside both the spheroid and the sample. Therefore, it can be expanded into spheroidal harmonics, which are products of the generalized Legendre functions of the first and second kind $P_l^m(x)$ and $Q_l^m(x)$. Here $m = 0, \pm 1, \pm 2, \ldots$ is the z-axis angular momentum and l must be greater or equal to $|m|$. As shown in Section 2.10, the expansion coefficients $A^m{}_l$ can be related to the charge distribution

on the spheroid. For example, ${A^m}_0$ is proportional to the total oscillating charge of the spheroid $\propto e^{-i\omega t}$. For a passive probe, ${A^m}_0 = 0$. The $l = 1$ terms determine the components of the dipole moment induced on the probe:

$$p_z = -\frac{1}{3}F^3 {A^0}_1\,, \quad p_x - ip_y = \frac{2}{3}F^3 {A^1}_1\,. \tag{2.14}$$

For each m the set of coefficients ${A^m}_l$ satisfies the infinite-order system of linear equations

$$\sum_{l'=1}^{\infty} (\Lambda^m_{ll'} - H_{ll'}) {A^m}_{l'} = {b^m}_l\,, \tag{2.15}$$

where $\Lambda^m_{ll'}$ and $H_{ll'}$ are defined by Eqs. (2.21) and (2.24) below. According to Eq. (2.14), to find $\boldsymbol{p}$ we need to consider only $m = 0$ and $m = 1$. The requisite coefficients ${b^m}_l$ on the right-hand side of Eq. (2.15) are given by

$${b^0}_1 = -\frac{4}{3}E^z\,, \tag{2.16}$$

$${b^1}_1 = \frac{4}{3}(E_x - iE_y)\,, \quad {b^{-1}}_1 = \frac{1}{3}(E_x + iE_y)\,. \tag{2.17}$$

If the external field $\boldsymbol{E}_{\rm ext} = E_x\hat{\boldsymbol{x}} + E_y\hat{\boldsymbol{y}} + E_z\hat{\boldsymbol{z}}$ is uniform, all other ${b^m}_l$ vanish. Once we solve the system (2.15) for $m = 0$, we can find the transverse polarizability from

$$\chi^{\perp} = \frac{p^z}{E^z} = \frac{4}{9}F^3\,\frac{{A^0}_1}{{b^0}_1}\,. \tag{2.18}$$

In turn, the solution for $m = 1$ would give us ${A^1}_1$ and

$$\chi^{\parallel} = \frac{p_x - ip_y}{E_x - iE_y} = \frac{8}{9}F^3\,\frac{{A^1}_1}{{b^1}_1}\,. \tag{2.19}$$

Equation (2.15) we wish to solve can be cast in a matrix form

$$(\mathbf{\Lambda}^m - \mathbf{H})\,\mathbf{A}^m = \mathbf{b}^m\,. \tag{2.20}$$

Matrix $\mathbf{\Lambda}^m$ is diagonal, $\Lambda^m{}_{ll'} = \Lambda^m_l \delta_{ll'}$, where

$$\Lambda^m_l = \frac{(-1)^m}{2l+1}\frac{4}{\varepsilon_{\rm tip}-1}\left[\varepsilon_{\rm tip}\frac{Q^m_l(\xi_0)}{P^m_l(\xi_0)} - \frac{\frac{d}{d\xi_0}Q^m_l(\xi_0)}{\frac{d}{d\xi_0}P^m_l(\xi_0)}\right] \tag{2.21}$$

and $\varepsilon_{\rm tip}$ is again the dielectric constant of the spheroid. If the probe is made of an ideal conductor, $\varepsilon_{\rm tip} \to \infty$, then Eq. (2.21) simplifies to

$$\Lambda^m_l = (-1)^m \frac{4}{2l+1}\frac{Q^m_l(\xi_0)}{P^m_l(\xi_0)}\,. \tag{2.22}$$

All these Λ^m_l are actually positive numbers because the factor $(-1)^m$ is compensated by the same factor in the definition of $Q^m_l(\xi_0)$. The behavior of Λ^m_l at large l is approximately exponential, as can be deduced from the asymptotic formula

$$(-1)^m \frac{Q^m_l(\xi_0)}{P^m_l(\xi_0)} \simeq \pi e^{-(2l+1)\alpha_0}\,, \quad \alpha_0 \equiv \operatorname{arccosh}\xi_0\,. \tag{2.23}$$

In Sec. 2.6 we also consider the case where $\varepsilon_{\rm tip}$ is a finite positive number. In this case the decay of Λ^m_l at large l is also exponential but with a different factor in front.

The elements of matrix $\mathbf{H}$ in Eq. (2.20) are given by

$$H_{ll'} \equiv 2\pi \int_0^\infty r_{\rm P}(q,\omega) I_{l+\frac{1}{2}}(qF) I_{l'+\frac{1}{2}}(qF) e^{-2qz_{\rm p}} \frac{dq}{q} \tag{2.24}$$

where $I_\nu(z)$ are the modified Bessel functions of the first kind and

$$z_{\rm p} \equiv z_{\rm tip} + L\,. \tag{2.25}$$

As mentioned in Sec. 2.1, the reflectivity $r_{\mathrm{P}}(q,\omega)$ may have strong peaks at the dispersion curves $\omega(q)$ of the surface polaritons of the sample. In practice, $r_{\mathrm{P}}(q,\omega)$ is always finite, so that the integrand in Eq. (2.24) is well-behaved and exponentially decreasing. A fast method of computing $H_{ll'}$ numerically is explained in Supplemental Material. In the remainder of this section we will assume that $r_{\mathrm{P}}(q,\omega)$ is q-independent. We will show that the polarizabilies of the spheroidal probe are meromorphic functions as stated in Sec. 2.1. We will also present our analytical and numerical results concerning the behavior of their poles and residues.

If $r_{\mathrm{P}}(q,\omega)=\beta=\mathrm{const}$, then matrix $\mathbf{H}$ factorizes $\mathbf{H}=\beta\bar{\mathbf{H}}$ and Eq. (2.20) becomes

$$\left(\mathbf{\Lambda}^m-\beta\bar{\mathbf{H}}\right)\mathbf{A}^m=\mathbf{b}^m\,. \tag{2.26}$$

A particular case of this equation for $z_{\mathrm{tip}}=0$ was previously derived in Ref. [BV87]. In general, Eq. (2.26) implies that $\mathbf{A}^m$ as a function of β has poles β_k^{ν} that are the solutions of the eigenvalue problem

$$\left(\mathbf{\Lambda}^m-\beta_k^{\nu}\bar{\mathbf{H}}\right)\mathbf{u}_k=0\,. \tag{2.27}$$

The substitution $\mathbf{u}_k=(\mathbf{\Lambda}^m)^{-1/2}\mathbf{v}_k$ transforms it to

$$\mathbf{v}_k=\beta_k^{\nu}\mathbf{M}\mathbf{v}_k\,,\quad \mathbf{M}=(\mathbf{\Lambda}^m)^{-1/2}\bar{\mathbf{H}}\,(\mathbf{\Lambda}^m)^{-1/2}\,. \tag{2.28}$$

Since all Λ_l^m are assumed to be positive, matrix $\mathbf{M}$ is real and symmetric, and so its eigenvalues $(\beta_k^{\nu})^{-1}$ are real and its eigenvectors $\mathbf{v}_k$ can be chosen to be orthonormal. Assuming $\mathbf{v}_k$ form a complete basis, the solution $\mathbf{A}^m$ of Eq. (2.26) can be sought as a linear combination of the corresponding $\mathbf{u}_k$. Taking into account Eqs. (2.18) and(2.19), we arrive at Eq. (2.3) with

$$\frac{R_k^{\nu}}{\beta_k^{\nu}}=\chi_0^{\nu}\,|(\mathbf{v}_k)_0|^2\,,\quad \chi_0^{\nu}=\frac{4}{9}\frac{m+1}{\Lambda_1^m}F^3\,, \tag{2.29}$$

where, once again, $m=0$ for $\nu=\perp$, $m=1$ for $\nu=\parallel$, and $(\mathbf{v}_k)_0$ is the first component of vector

$\mathbf{v}_k$. The completeness of the basis entails $\sum_k |(\mathbf{v}_k)_0|^2 = 1$, leading to the sum rule (2.4). The explicit formulas for χ_0^ν that follow from Eqs. (2.22) and (2.29) are

$$\chi_0^\perp = \frac{L^3}{3\xi_0^3}\left[\frac{1}{2}\ln\left(\frac{\xi_0+1}{\xi_0-1}\right) - \frac{1}{\xi_0}\right]^{-1} = \frac{V}{4\pi L^\perp}\,, \tag{2.30a}$$

$$\chi_0^\parallel = \frac{2L^3}{3\xi_0^3}\left[\frac{\xi_0}{\xi_0^2-1} - \frac{1}{2}\ln\left(\frac{\xi_0+1}{\xi_0-1}\right)\right]^{-1} = \frac{V}{4\pi L^\parallel}\,, \tag{2.30b}$$

where $V = (4\pi/3)L^2 a$ is the volume of the spheroid and L^ν are the depolarization factors of the spheroid [BH04]

$$L^\perp = (\xi_0^2-1)\left[\frac{1}{2}\xi_0\ln\left(\frac{\xi_0+1}{\xi_0-1}\right) - 1\right], \tag{2.31}$$

$$L^\parallel = \frac{1-L^\perp}{2}\,. \tag{2.32}$$

These formulas should be familiar from classical electrostatics or from the theory of light scattering by small particles. [BH04] For strongly elongated spheroid $L \gg a$, $\xi_0 \simeq 1$, they yield

$$\chi_0^\perp \simeq \frac{2}{3}\frac{L^3}{\ln(4L/a)}\,, \tag{2.33a}$$

$$\chi_0^\parallel \simeq \frac{2}{3}L^2 a\,. \tag{2.33b}$$

In Sec. 2.1 we stated that the sequence β_k^ν may not have accumulation points. For the present case of a spheroidal probe this can be proven directly from the properties of matrix $\mathbf{M}$. The first step is to show that the matrix elements of $\bar{\mathbf{H}}$ obey the asymptotic bound

$$\ln \bar{H}_{ll'} < -(l+l'+1)\,\mathrm{arccosh}\left(\frac{z_\mathrm{p}}{F}\right) \tag{2.34}$$

at large l and l'. This can be established using the saddle-point integration in Eq. (2.24). Together with Eqs. (2.22) and (2.23), this bound ensures that at $z_\mathrm{tip} > 0$ the high-order matrix elements of

$\mathbf{M}$ decay exponentially,

$$\ln M_{ll'} < -(l+l'+1)\left[\operatorname{arccosh}\left(\cosh\alpha + \frac{z_{\rm tip}}{F}\right) - \alpha\right].$$

Here α [Eq. (2.8)] parametrizes the probe-sample distance $z_{\rm tip}$. Hence, the double series $\sum_{ll'} M_{ll'}^2 = \operatorname{tr}\mathbf{M}^2$ is convergent. Considering the identity

$$\sum_{k=0}^{\infty} (\beta_k^{\rm v})^{-2} = \operatorname{tr}\mathbf{M}^2 < \infty \tag{2.35}$$

we see that the accumulation points are ruled out. On the contrary, $\operatorname{tr}\mathbf{M}^2$ diverges at $z_{\rm tip} = 0$ and one accumulation point does exist: $\beta = 1$. For the sphere this can be found directly from Eq. (2.7) by setting $\alpha = 0$.

In the spherical limit $\xi_0 \to \infty$ an analytical solution of our equations exists although it is not obvious. We deduced the form of this solution from the method of images, Section 2.12. At finite ξ_0 we resorted to solving the problem numerically. As already mentioned, due to an exponential growth of $\beta_k^{\rm v}$ with k, only a first few of such poles are usually needed for evaluating the polarizabilities in question $\chi^{\rm v}$. To compute such $\beta_k^{\rm v}$ and the corresponding $R_k^{\rm v}$ we used the following procedure. Given L/a and α, we would generate an $N \times N$ matrix made of the first N rows and columns of the full infinite matrix $\mathbf{M}$. We would diagonalize this finite-size matrix by standard library routines (using MATLAB). The obtained eigenvalues approximate the first N poles $\beta_k^{\rm v}$. We would gradually increase the matrix size until the poles we are interested in would show no variation as a function of N within the desired accuracy. The larger L/a and the smaller α, the higher N was needed. We found this procedure workable as long as N did not exceed about 500. As a rule, the higher eigenvalues of larger matrices would either fail to reach the accuracy or would show an α-dependence inconsistent with physical principles. This behavior stems most likely from roundoff errors. In principle, one can combat them by utilizing higher-precision arithmetic but we did not pursue this route. For $L = 25a$ the computation of the first nine poles

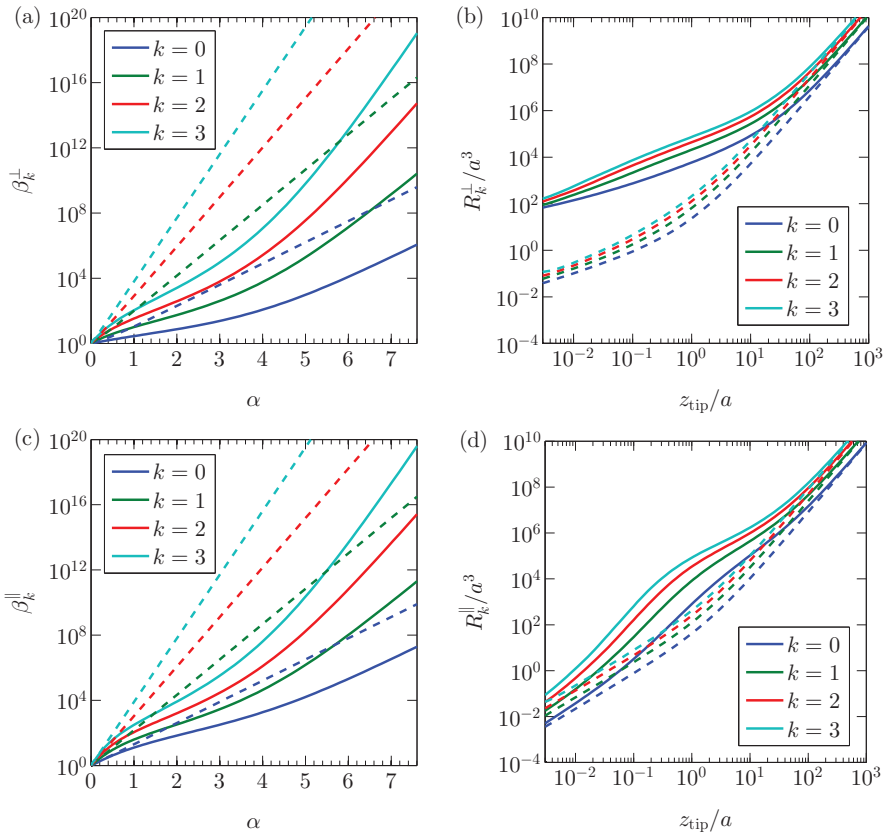

Figure 2.4: (Color online) (a) The first four poles β_k^{ν} of the polarizability χ^{ν} for perfectly conducting spheroids. The probe-sample distance is parameterized by $\alpha = \operatorname{arccosh}(1 + z_{\mathrm{tip}}/a)$ [Eq. (2.8)]. The solid lines are for a strongly elongated spheroid $L = 25a$, the dashed lines are for a nearly spherical one $L = 1.01a$. The external field is in the z-direction, $\nu = \perp$. (b) The corresponding residues $R_k^{\perp}$ divided by a^3. Poles for different shapes converge at small z_{tip}, while residues converge at large z_{tip}. (c), (d) Similar plots for the external field in the x-y plane, $\nu = \parallel$.

with at least two-digit accuracy was possible for $\alpha > 0.08$, i.e., $z_{\mathrm{tip}} > 0.003a$. The residues R_k^{ν} were obtained from the eigenvectors of the truncated matrix $\mathbf{M}$ using Eqs. (2.29)–(2.30b). In the interval $0 < \alpha < 0.08$ we used the linear interpolation between $\beta_k^{\nu}(\alpha = 0.08)$ and $\beta_k^{\nu}(\alpha = 0) = 1$.

The results of these calculations are presented in Fig. 2.4 for the first four modes, $k = 0$ to 3. The solid lines in panels (a) and (c) show $\beta_k^{\perp}$ and $\beta_k^{\parallel}$, respectively, as a function of α. The corresponding quantities for a sphere are shown by the dashed lines. The residues R_k^{ν}/a^3 are plotted in panels (b) and (d). The first nine pole-residue pairs of the spheroid for $\nu = \perp$ have also been fitted with an error of 5% or smaller to a combination of elementary functions in the

range $0.003a < z_{\text{tip}} < 10a$. The fitting formulas and their coefficients are cataloged in Section 2.11. The residue $R_8^{\perp}$ behaves differently from the others because it was constrained to satisfy the sum rule (2.4). Using these formulas one can find the response $\chi^{\perp}$ with negligible computational cost for any $\beta(\omega)$ as long as its value is not extremely large. Note that although these results are for perfectly conducting spheroids $\varepsilon_{\text{tip}} = \infty$, calculations for arbitrary finite ε_{tip} can be done in the same way except one has to use Eq. (2.21) instead of Eq. (2.22).

Let us now compare the obtained dependence of β_k^{ν} on z_{tip} with the limiting asymptotic behavior predicted in Sec. 2.2. First, at $z_{\text{tip}} \ll a$, the poles of the spheroid approach that of a sphere, as expected, see Fig. 2.4(a) and (c). The other limit is $z_{\text{tip}} \gg L$, where the point-dipole formulas (2.9a)–(2.9b) should apply. In Fig. 2.4 it is seen that the lowest eigenvalue of both shapes indeed have the correct behavior. The intermediate regime $a \ll z_{\text{tip}} \ll L$ is the most nontrivial one. We argue that in this regime function $\beta_0^{\perp}(z_{\text{tip}})$ behaves as

$$\beta_0^{\perp}(z_{\text{tip}}) = c \ln(z_{\text{tip}}/a), \quad a \ll z_{\text{tip}} \ll L, \tag{2.36}$$

with some coefficient $c \sim 1$ independent of L. To arrive at this formula we first find bounds on $\beta_0^{\perp}$ using the following theorem. Consider two perfectly conducting probes of different sizes. If the surface of one probe can be inscribed into the other, then the first probe must have a larger β_0^{ν}. This statement is physically natural because self-sustained oscillations around the smaller body require a larger compensation from the surface reflectivity, cf. Sec. 2.2. It can also be proven mathematically from the variational principle. [AKSV99, MF81] To place bounds on $\beta_0^{\perp}$ of the spheroid we can consider two other probes, a larger one and a smaller one. We get

$$\beta_0^{\text{cone},\nu} < \beta_0^{\nu} < \beta_0^{\text{ss},\nu}, \tag{2.37}$$

where $\beta_0^{\text{cone},\nu}$ is the lowest pole of a cone with a vertex touching the sample and enveloping the spheroid; $\beta_0^{\text{ss},\nu}$ is the lowest pole of a spheroid of shorter length $L = z_{\text{tip}}$. It can be

shown [PWM$^+$01, PRL$^+$05, Fog13] that $\beta_0^{\text{cone},\perp} \simeq (1/\pi)\ln(z_{\text{tip}}/a)$. As for the smaller spheroid, the point-dipole formula should apply by order of magnitude, $\beta_0^{\text{ss},\perp} \sim 6\ln(z_{\text{tip}}/a)$, cf. Eqs. (2.9a) and (2.33a). Since the functional form of these bounds coincides with Eq. (2.36) up to numerical coefficients, we argue that $\beta_0^{\perp}(z_{\text{tip}})$ should obey the same equation as well. The graph shown in Fig. 2.4(a) is consistent with this prediction. However, due to numerical limitations L/z_{tip} and z_{tip}/a could not be very large in our simulation and we could obtain only a crude estimate $1 < c < 3$ of the coefficient c. The poles $\beta_k^{\parallel}$ of the in-plane polarizability, which are plotted in Fig. 2.4(c) as a function of α, also show crossovers among three regimes (short, long, and intermediate distances) and can be understood in a similar way.

The behavior of the residues R_k^{ν} is more difficult to analyze. At large distances $z_{\text{tip}} \gg L$, the residues of the spheroid approach those of the sphere [Eq. (2.10)]. At small distances, where the poles behave as $\ln\beta_k^{\nu} \sim (2k+3)\alpha$, the polarizability is determined by a large number $\sim 1/\alpha$ of terms in the pole-residue series. The sum rule (2.4) implies that the sum of these dominant residues must be of the order of χ_0^{ν} for each shape. Indeed, the residues of the sphere, which have the form $R_k^{\perp} \propto ka^3\alpha^2$ and $R_k^{\parallel} \propto (k+1)(k+2)a^3\alpha^3$ (see Section 2.12), obey this requirement. The residues of the spheroid are always larger than those of the sphere, consistent with the higher χ_0^{ν}. The intermediate-distance behavior of R_k defies an obvious characterization. It is intriguing that at small distances only the residues are affected by the aspect ratio of the probe, while at large distances only the poles are altered.

Information about the probe-sample coupling complementary to the properties of the poles and residues can be obtained by examining the potential distribution of the polariton modes in real space. The examples for the $\perp$ modes are depicted in Fig. 2.1(b). The potential is strongly peaked near the tip of the spheroid, demonstrating the localized nature of near-field coupling. Note that the number of times the potential changes sign along x is equal to k.

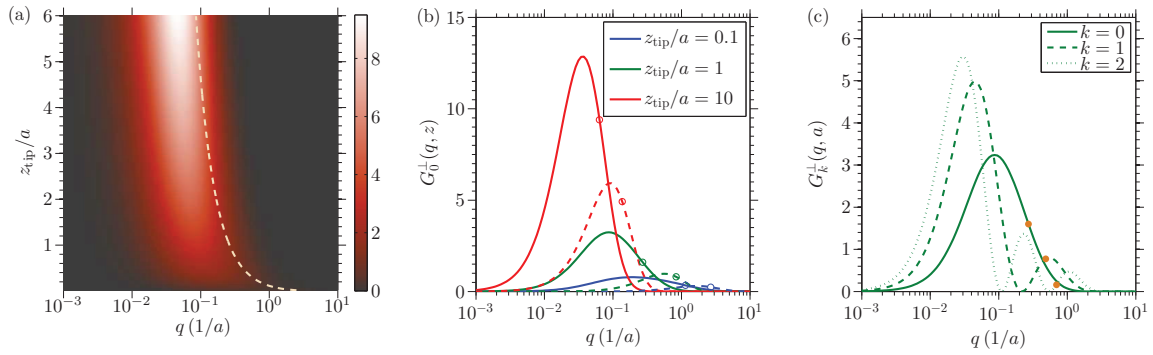

Figure 2.5: (Color online) (a) The weight function $G_0^\perp(q, z_{\rm tip})$ for the spheroid with $L = 25a$ and $\varepsilon_{\rm tip} = \infty$. The dashed line shows $\bar{q}_0^\perp(z_{\rm tip})$. (b) $G_0^\perp(q, z_{\rm tip})$ for several $z_{\rm tip}$, with circles indicating $\bar{q}_0^\perp$. The solid lines are for the spheroid, the dashed lines are for the sphere. The spheroid is more sensitive to small q compared to the sphere, while both shapes are more sensitive to large q as $z_{\rm tip}$ decreases. (c) The first three $G_k^\perp(q)$ for $z_{\rm tip} = a$, with solid circles indicating $\bar{q}_k^\perp$. The number of nodes in $G_k^\perp(q)$ is equal to k, while $\bar{q}_k^\perp$ increases with k. The logarithmic scaling of the horizontal axes is used to show the small-q structure more clearly.

2.4 Momentum-dependence of the probe-sample coupling

A simple physical picture of the s-SNOM that served as an important insight in the early days of the field and still remains popular today is the notion that the probe couples predominantly to momenta $q \sim 1/a$. Accordingly, the s-SNOM signal is collected from a very small region of size $\sim a$ directly underneath the tip. Modern applications of s-SNOM to two-dimensional and layered systems require going beyond this oversimplified picture because the q-dependence of the reflectivity $r_{\rm P}(\omega, q)$ of such systems can be very sharp due to presence of dispersive collective modes (Sec. 2.7). Recall that for a momentum-independent reflectivity β [Eq. (2.2)], the poles and residues of the polariton eigenmodes are determined solely by the permittivity and geometry of the probe. Unfortunately, for a q-dependent reflectivity such a clean separation of the probe and sample properties in the eigenproblem is not possible. While one can still define the eigenmodes by suitably modifying Eq. (2.3), the corresponding poles and residues will be, in general, complicated functionals of $r_{\rm P}(\omega, q)$. However, if the q-dependence of the reflectivity is weak, it can be treated as a perturbation, and the sample-independent resonant modes are retained.

As we show in this Section, in this perturbative case one can precisely define the probe-sample coupling as a function of q and the 'dominant' momentum as a function of z_{tip}.

Consider a small q-dependent correction to the reflectivity:

$$r_{\text{P}}(q) = \beta + \delta r_{\text{P}}(q)\,. \tag{2.38}$$

The kth pole β_k^{v} of the probe-sample eigenmodes is a functional of r_{P}. The key question is how this pole is affected by the nonlocal correction to r_{P}. The answer can be written in terms of $-G_k^{\text{v}}(q, z_{\text{tip}})$, the first variational derivative of $\beta_k^{\text{v}}[r_{\text{P}}(q)]$ with respect to r_{P}:

$$\delta\beta_k^{\text{v}}(z_{\text{tip}}) = -\int_0^\infty G_k^{\text{v}}(q, z_{\text{tip}})\delta r_{\text{P}}(q)dq\,. \tag{2.39}$$

This is the desired relation to the leading order in δr_{P}. A few general properties of function G_k^{v} at $q < 1/z_{\text{tip}}$ can be established. First, this function decays exponentially at large q:

$$G_k^{\text{v}}(q, z_{\text{tip}}) \sim e^{-2qz_{\text{tip}}}. \tag{2.40}$$

This is so because the probe-sample interaction is mediated by multiple reflections of evanescent waves (Sec. 2.1) and the shortest distance such waves have to travel is $2z_{\text{tip}}$. Next, it is easy to see that G_k^{v} is normalized:

$$\int_0^\infty G_k^{\text{v}}(q, z_{\text{tip}})dq = 1\,. \tag{2.41}$$

Using a variation principle one can also show that for a perfectly conducting probe $G_k^{\text{v}}(q, z_{\text{tip}})$ is nonnegative. Therefore, functions $G_k^{\text{v}}(q, z_{\text{tip}})$ can be considered *weight* functions for the perturbation $\delta r_{\text{P}}(q)$. To put it another way, this set of functions quantifies the momentum dependence of the probe-sample coupling. Below we show that the properties of these functions paint a much more nuanced physical picture than the naive idea that the coupling is maximized at a single momentum $q \sim 1/a$. However, if one insists on characterizing the entire distribution of

relevant momenta by a single number, the logical candidates are the average momenta

$$\bar{q}_k^{\nu} = \int_0^{\infty} G_k^{\nu}(q, z_{\text{tip}}) q dq \,. \tag{2.42}$$

The idea is that unless $G_k^{\nu}(q, z_{\text{tip}})$ has a complicated structure or a slow decay, $\bar{q}_k^{\nu}$ should play the role of a characteristic momentum that determines kth polariton pole β_k^{ν}. Accordingly, we may expect that $1/\bar{q}_k^{\nu}$ should give an improved estimate of the spatial resolution of the probe in the context of near-field imaging by s-SNOM. Interestingly, $\bar{q}_k^{\nu}$ can be found by differentiating $\beta_k^{\nu}(z_{\text{tip}})$:

$$\bar{q}_k^{\nu}(z_{\text{tip}}) = \frac{1}{2} \frac{\partial}{\partial z_{\text{tip}}} \log \beta_k^{\nu} \,. \tag{2.43}$$

To obtain this formula consider first a sample with a q-independent reflectivity β and let the probe-sample separation be $z_{\text{tip}} = z + dz$. This system is equivalent to another one: the probe separated by $z_{\text{tip}} = z$ from a fictitious *two-component* medium composed of a vacuum layer of thickness dz plus the original sample. The surface reflectivity of such a two-component medium is q-dependent, $r_{\text{P}}(q) = \beta e^{-2qdz}$, so that it has the form (2.38) with $\delta r_{\text{P}}(q) = -2q\beta dz$. Evidently, such a $\delta r_{\text{P}}(q)$ shifts the resonant pole from $\beta = \beta_k^{\nu}(z)$ to $\beta = \beta_k^{\nu}(z + \delta z)$, i.e., causes a differential change $\delta\beta_k^{\nu} = (\partial\beta_k^{\nu}/\partial z)dz$. Substituting these relations into Eq. 2.39, we get Eq. (2.43). Note that as β_k^{ν} rises more steeply with z_{tip} for larger k, $\bar{q}_k^{\nu}$ increases with k.

An equivalent description of the effect of a q-dependent perturbation is that it induces a correction to the surface reflectivity. The effective reflectivity β^{eff} is different for each k and ν,

$$\beta_k^{\nu,\text{eff}} \equiv \beta - \delta\beta_k^{\nu} = \int_0^{\infty} G_k^{\nu}(q, z_{\text{tip}}) r_{\text{P}}(q) dq \,. \tag{2.44}$$

The corresponding polarizability χ^{ν} is given by

$$\chi^{\nu} = \sum_{k=0}^{\infty} \frac{R_k^{\nu}}{\beta_k^{\nu} - \beta_k^{\nu,\text{eff}}} \,. \tag{2.45}$$

In the following we focus on function $G_0^{\nu}(q, z_{\rm tip})$ because $k = 0$ is the dominant resonance at all but very small $z_{\rm tip}$. Actually, the large-distance limit of this function has the universal form

$$G_0^{\nu}(q, z_{\rm tip}) \simeq 4 z_{\rm tip}^3 q^2 e^{-2q z_{\rm tip}}, \quad z_{\rm tip} \gg L, \tag{2.46}$$

same for both ν. Equation (2.46) follows from Eqs. (2.11) and (2.39) and is consistent with the surmised large-q behavior (2.40). As one can see, Eq. (2.46) gives $G_0^{\nu}(q, z_{\rm tip})$ that is normalized, nonnegative, and has a single maximum at $q = 1/z_{\rm tip}$. The average momentum is $\bar{q}_0^{\nu} \simeq 3/(2z_{\rm tip})$.

In the intermediate-distance regime functions $G_k^{\nu}(q, z_{\rm tip})$ are not expected to be universal. The specific example we treat in detail is again the conducting spheroidal probe. Combining Eq. (2.43) with the results of Secs. 2.2 and 2.3, for the strongly elongated spheroid we obtain the following:

$$\frac{1}{\bar{q}_0^{\perp}(z_{\rm tip})} \sim \begin{cases} (a z_{\rm tip})^{1/2}, & z_{\rm tip} \ll a, \\ 2 z_{\rm tip} \log\left(\dfrac{z_{\rm tip}}{a}\right), & a \ll z_{\rm tip} \ll \tilde{L}, \\ \dfrac{2 z_{\rm tip}}{3}, & z_{\rm tip} \gg L. \end{cases} \tag{2.47}$$

Since the left-hand side has the physical meaning of the spatial resolution of the probe, we expect it to monotonically decrease as $z_{\rm tip}$ decreases. Therefore, the length scale $\tilde{L}$ appearing on the second line of Eq. (2.47) should be of the order of $L/3\log(L/a)$. The presence of a large logarithmic factor $\log(z_{\rm tip}/a)$ in the intermediate-distance regime $a \ll z_{\rm tip} \ll \tilde{L}$ indicates that function $G_0^{\perp}(q, z_{\rm tip})$ has a considerable weight at q parametrically smaller than $1/z_{\rm tip}$. In other words, a strongly elongated spheroidal probe senses electric fields beyond its immediate vicinity $\rho < z_{\rm tip}$. (A similar point was made previously in Ref. [ZAF$^+$12].) As L/a decreases, $\tilde{L}$ comes close to a, and this intermediate regime disappears. For example, the sphere acts essentially as a local probe.

The calculation of $G_k^{\nu}(q, z_{\rm tip})$ for the spheroid can be done as follows. Applying the

first-order perturbation theory to the linear system (2.26), one finds

$$G_k^{\nu}(q, z_{\rm tip}) = \frac{\mathbf{u}_k^{\dagger} \mathbf{H}' \mathbf{u}_k}{\mathbf{u}_k^{\dagger} \bar{\mathbf{H}} \mathbf{u}_k}\,, \tag{2.48}$$

where $\mathbf{H}'$ is the matrix with elements

$$H'_{ll'} = \frac{2\pi}{q} I_{l+\frac{1}{2}}(qF) I_{l'+\frac{1}{2}}(qF) e^{-2qz_{\rm p}}\,. \tag{2.49}$$

Once the eigenvectors $\mathbf{u}_k$ are found, e.g., as described in Sec. 2.3, function $G_k^{\nu}(q, z_{\rm tip})$ can be readily computed.

Our numerical investigation of $G_k^{\nu}(q, z_{\rm tip})$ was limited mainly to $k = 0$ and $\nu = \perp$. We observed that the eigenvector components approximately followed the geometric series $(\mathbf{u}_0)_j \sim t^j$. The quotient t is somewhat larger than unity for small $z_{\rm tip}$. As $z_{\rm tip}$ increases, t becomes less than unity, so that the first component $(\mathbf{u}_0)_0$ dominates. Neglecting all other components and expressing the modified Bessel function $I_{3/2}(z)$ in terms of elementary functions, we obtain the analytical approximation from Eqs. (2.48) and (2.49):

$$G_0^{\nu}(q, z_{\rm tip}) = \frac{c_0}{q^4} \left(qF \cosh qF - \sinh qF\right)^2 e^{-2qz_{\rm p}}\,, \tag{2.50}$$

where c_0 is a normalization constant. At $z_{\rm tip} \gg L$ we can focus on the range of momenta less than $1/L$ because at larger q this function is already exponentially small. For such q the bracketed expression on the right-hand side can be replaced by $(Fq)^6/9$ and $z_{\rm p} = z_{\rm tip} + L$ by $z_{\rm tip}$, which yields the asymptotic form (2.46).

To examine small and intermediate distances we used the direct numerical evaluation of $\mathbf{u}_0$ and $G_0^{\nu}(q, z_{\rm tip})$. As in Sec. 2.3, we considered two aspect ratios: $L/a = 25$ and $L/a = 1$. Only $\nu = \perp$ part was studied. The results for $L/a = 25$ are shown using the false color scale in Fig. 2.5(a). It can be seen that as $z_{\rm tip}$ decreases, both $\bar{q}_0^{\perp}(z_{\rm tip})$ and the position of the maximum of

$G_0^{\perp}(q, z_{\text{tip}})$ as a function of q shift toward larger values. This implies that the probe becomes more sensitive to finer spatial features of the sample, as discussed above. The line plot of $G_0^{\perp}(q, z_{\text{tip}})$ for several z_{tip} presented in Fig. 2.5(b) depicts the same trend. The average momentum $\bar{q}_0^{\perp}$ and the position of the $G_0^{\perp}(q)$ maximum are of the same order of magnitude except at very short distances where $\bar{q}_0^{\perp}$ increases more rapidly as z_{tip} decreases. Note that Eq. (2.47) predicts that $\bar{q}_0^{\perp}$ diverges at $z_{\text{tip}} = 0$. From Fig. 2.5(b) we also see that for the same z_{tip} the maximum of $G_0^{\perp}(q, z_{\text{tip}})$ is found at q smaller by a factor of 3–10 for the spheroid compared to the sphere. This confirms that the spheroid is much more sensitive to small in-plane momenta than the sphere, i.e., the response of a strongly elongated spheroid is affected by a relatively wide range of lengthscales.

For $k > 0$, $G_k^{\perp}(q, z_{\text{tip}})$ has nodes as a function q at fixed z_{tip}. The number of nodes is equal to k, see Fig. 2.5(c). Apparently, at such q near-field coupling between oscillatory charge distributions on the probe and the sample exactly vanishes. Therefore, small perturbations at such discrete q do not affect the kth resonant mode. Finally, although $\bar{q}_k^{\perp}$ increase with k for the reasons explained above, the maxima of $G_k^{\perp}$ show the opposite trend, which is presently not understood.

2.5 From near-field polarizabilities to far-field observables

In order to apply our theory to simulation of s-SNOM experiments, we need to include a few more ingredients in our calculation. The first one is the so-called far-field factor (FFF) $F^{\nu}(\omega)$. This factor accounts for the fact that the probe is illuminated not only by the incident wave but also by its reflection from the sample. In experiment P-polarized incident field is usually used, to take advantage of the high transverse polarizability of the probe. Assuming the sample surface is flat, uniform, and its linear dimensions are much longer than the radian sphere diameter c/ω, the reflection of the incident wave is described by the coefficient $r_{\text{P}}(q_s, \omega)$, where

$$q_s = \frac{\omega}{c} \sin\theta \tag{2.51}$$

is the in-plane photon momentum and θ is the angle of incidence. Hence, the ratio of ν-component of the electric field at the surface to that of the incident wave is $1 \pm r_{\mathrm{P}}(q_s, \omega)$ for $\nu = \perp$ and $\parallel$, respectively. The FFF also takes into account that the field scattered by the probe reaches the detector in two waves: directly and after reflection from the sample surface. Usually, the backscattered field is measured. It has the in-plane momentum $-q_s$ and therefore the same reflection coefficient $r_{\mathrm{P}}(-q_s, \omega) = r_{\mathrm{P}}(q_s, \omega)$ as the incident wave. The total FFFs for this setup are given by

$$F^{\perp}(\omega) = [1 + r_{\mathrm{P}}(q_s, \omega)]^2 \sin^2\theta\,, \tag{2.52a}$$

$$F^{\parallel}(\omega) = [1 - r_{\mathrm{P}}(q_s, \omega)]^2 \cos^2\theta\,. \tag{2.52b}$$

The trigonometric factors on the right-hand side take care of conversion between the total electric field E_{ext} of the waves and their $\perp$, $\parallel$ components. Note that our assumption of the plane-wave illumination is not entirely realistic. In experiment, a focused Gaussian beam is typically used, in which case the FFFs are effectively averaged out over a range of angles θ. Numerical apertures ~ 0.4 are common. We must also stress that Eq. (2.52) should be modified if the system studied by s-SNOM is nonuniform on scales shorter than c/ω. Typical examples include a small sample residing on some substrate [ZAF$^+$12] or measurements done close to a boundary of two different materials.

Another point we have to discuss is signal demodulation. In the experiment the probe is made to oscillate mechanically, which causes periodic variation of the probe-sample distance:

$$z_{\mathrm{tip}}(\varphi) = z_0 + \Delta z\,(1 - \cos\varphi)\,, \varphi \equiv \Omega t\,. \tag{2.53}$$

The oscillation amplitude is typically $\Delta z = 20$–$90\,\mathrm{nm}$, comparable to the radius of curvature $a \sim 30\,\mathrm{nm}$ of the probe. The minimal approach distance $z_0 \geq 0$ can be equal to zero if the probe taps the sample. The tapping frequency Ω is many orders of magnitude smaller than the laser frequency

ω, and so the motion of the tip does not affect the electromagnetic response. Effectively, the experiment consists of measuring the scattered signal for many static configurations with different z_{tip}. The nth Fourier harmonic of the backscattered field is referred to as the demodulated signal s_n. (Here we define s_n as a complex number but in experimental literature it is common to discuss the amplitude and the phase of s_n separately.) The primary purpose of demodulation is to suppress the far-field background signal created by reflections from the body of the tip, the cantilever, *etc*. This background is large but depends on z_{tip} very weakly (linearly) and thus contributes predominantly to the $n = 1$ harmonic. Unfortunately, demodulation strongly diminishes the signal amplitude, making it more susceptible to experimental noise. In practice, $n = 2$ or 3 usually gives the best approximation of the true near-field signal. [KH04, KH09, ABJR12]

The demodulated signal is related to the polarizabilities $\chi^{\nu}(\omega, z_{\text{tip}})$ we have been discussing in previous sections by

$$s_n^{\nu}(\omega) = \text{const} \times \chi_n^{\nu}(\omega) F^{\nu}(\omega)\,, \tag{2.54}$$

where $\chi_n^{\nu}(\omega)$ is the nth Fourier harmonic of χ^{ν}:

$$\chi_n^{\nu}(\omega) = \int_0^{\pi} \frac{d\varphi}{\pi}\, \chi^{\nu}\big(\omega, z_{\text{tip}}(\varphi)\big) \cos n\varphi\,. \tag{2.55}$$

One more element of the experimental protocol is normalization. What is typically reported is $s_n^{\nu}(\omega)$ normalized against a certain reference material, e.g., Si or Au:

$$\tilde{s}_n^{\nu}(\omega) = s_n^{\nu}(\omega) / s_n^{\nu,\,\text{ref}}(\omega)\,. \tag{2.56}$$

The normalization eliminates a number of physically uninteresting or poorly known factors, such as the constant in Eq. (2.54) that are related to the optical setup of the experiment. The FFFs may also be canceled if both the studied and the reference objects in the experiment are positioned

nearby, so that the data for the two are taken at points no farther apart than the diameter c/ω of the radian sphere.

The last point we wish to draw attention to is that the absolute value of the minimum probe-sample distance z_0 [Eq. (2.53)] cannot be determined very accurately. Therefore, experimentalists have to measure the so-called approach curve, which is the s-SNOM response as a function of z_0 at a fixed frequency. They then identify the point $z_0 = 0$ as a point where a qualitative change in behavior in s_2 or s_3 appears. The logic behind this procedure is that once the probe makes the mechanical contact with the sample, its oscillations become reduced in amplitude, marking an unambiguous change. A potential flaw of this argument is that sharp changes in s_n's may be generated by a rapid variation of electromagnetic coupling between the probe sample at short separation even *before* making mechanical contact. We will discuss this issue in more detail in Sec. 2.6.

2.6 Case of local reflectivity: aluminum oxide

In this and the following Sections we discuss the implications of our theory for near-field response of real materials. We choose bulk α-Al_2O_3, also known as sapphire or corundum, as our first example of highly resonant material with a momentum-independent reflectivity β [Fig. 2.2(a)]. Another material with these properties, silicon carbide, has been a subject of a recent s-SNOM study co-authored by two of the present authors.[MKG+14] Modeling results based on the BEM showing good agreement with the data were also reported in that work. Realistic probe shapes and retardation effects have been taken into account in order to achieve that. The latter was necessary since the probe length $2L \sim 20\,\mu$m in the experiments was in fact larger than the diameter of the radian sphere $c/\omega \sim 11\,\mu$m. Here we do not aim for a perfect agreement with a particular experiment but instead wish to illustrate how the general theory of multiple eigenmodes formulated in the preceding Sections can generate novel features in far-field observables. We

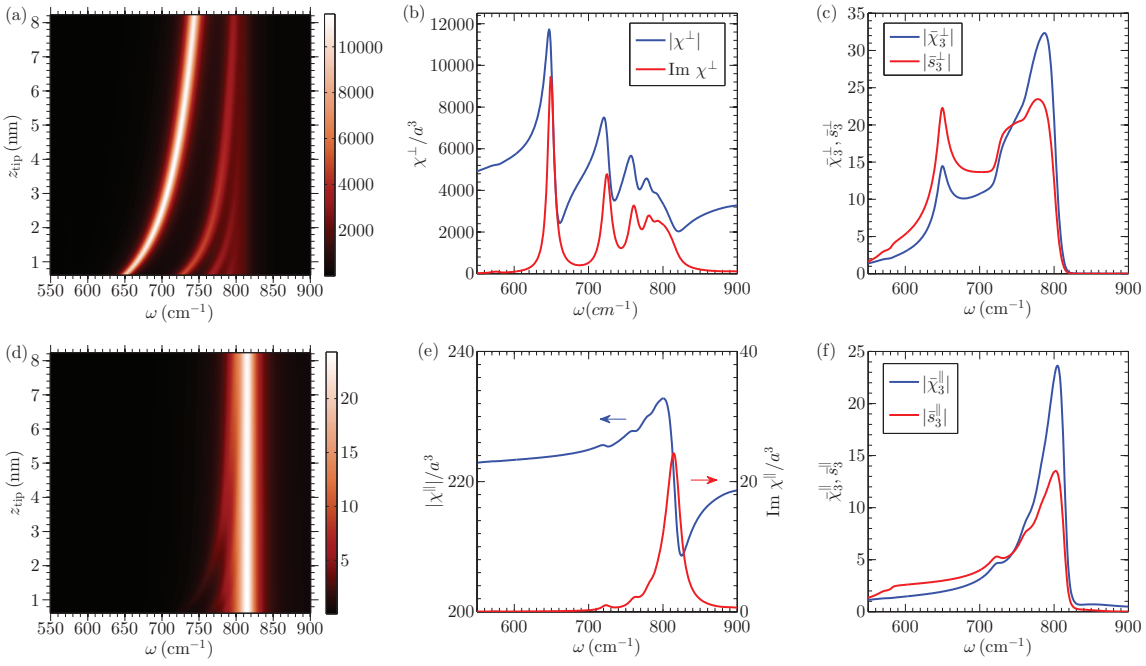

Figure 2.6: (Color online) Response of a perfectly conducting spheroidal probe with $L = 25a$ on bulk Al_2O_3 sample. (a) The false color plot of $\mathrm{Im}\,\chi^{\perp}(\omega, z_{\rm tip})/a^3$. The bright curves correspond to the resonant modes, with $k = 0$ mode having the lowest frequency. (b) The polarizability $\chi^{\perp}$ (absolute value and imaginary part) at $z_{\rm tip} = 0.02a = 0.6\,$nm, the smallest distance in panel (a). (c) The absolute value of the demodulated polarizability $|\bar{\chi}_3|$ and scattering signal $|\bar{s}_3|$ for the tapping amplitude $\Delta z = 50\,$nm and $z_0 = 0.6\,$nm. The origin of the three peaks is discussed in the text. (d)–(f) The counterparts of panels (a)–(c) for the parallel component, $\nu = \parallel$. The plots again reveal multiple resonances. However, the overall magnitude of the polarizability is greatly reduced, $\chi^{\parallel} \sim 10^{-2}\chi^{\perp}$, and the resonances are more strongly bunched near the surface phonon frequency $\omega_{\rm SP} = 818\,\mathrm{cm}^{-1}$.

study mostly probes of an idealized spheroidal shape but examine some other shapes as well. We stay within the quasistatic approximation but we will comment on retardation effects in Sec. 2.8.

We use the following momentum-independent model for the reflection coefficient of the uniaxial Al_2O_3 crystal,

$$\beta(\omega) = \frac{\varepsilon_{\rm eff} - 1}{\varepsilon_{\rm eff} + 1}, \quad \varepsilon_{\rm eff}(\omega) = \sqrt{\varepsilon_o \varepsilon_e}, \tag{2.57}$$

where ε_ρ for $\rho = o$ (ordinary) and e (extraordinary) axes is given by

$$\varepsilon_\rho(\omega) = \varepsilon_{\infty,\rho} \prod_j \frac{\omega_{j\mathrm{LO},\rho}^2 - \omega^2 - i\gamma_{j\mathrm{LO},\rho}\omega}{\omega_{j\mathrm{TO},\rho}^2 - \omega^2 - i\gamma_{j\mathrm{TO},\rho}\omega}. \tag{2.58}$$

The optical constants of Al_2O_3 reported in the literature [HNA94, ZPM13] have slight variations, presumably because of different crystal purity and processing. In our calculations we adopt the results of Ref. [ZPM13] at room temperature, reproduced in Section 2.15. (For simplicity, the weak oscillator at $\omega_{\mathrm{TO},o} = 634\,\mathrm{cm}^{-1}$ is neglected.) Due to smallness of the optical phonon linewidths γ_ρ in this material, the near-field reflectivity of Al_2O_3 can be as high as $\beta \sim 10$.

We start by studying the behavior of the probe polarizabilities χ^ν as a function of frequency ω. In the mid-infrared range, the reflection coefficient β of Al_2O_3 has a single peak centered at the surface-phonon frequency $\omega_{\mathrm{SP}} = 818\,\mathrm{cm}^{-1}$, depicted in Fig. 2.2(a). As ω approaches ω_{SP} from below, $\mathrm{Re}\,\beta(\omega)$ steeply rises. Equation (2.3) implies that whenever $\mathrm{Re}\,\beta$ is equal to a pole β_k^ν, $\mathrm{Im}\,\chi^\nu$ has a local maximum as long as the damping $\mathrm{Im}\,\beta(\omega)$ is not too large. The positions of three such underdamped resonances are indicated schematically in Fig. 2.2(a). Thus, a *single* surface mode ω_{SP} of Al_2O_3 may produce *multiple* modes of the coupled probe-sample system. These localized eigenmodes (resonances) have been discussed at length in the preceding Sections. For example, they are depicted in Fig. 2.1(b) for the case of a spheroidal probe. Note that all the resonances are *red-shifted* from the frequency ω_{SP}. Since $\mathrm{Im}\,\beta$ increases as ω approaches ω_{SP}, higher-order resonances are progressively more broad.

The scenario above is described in terms of constant β_k^ν. However, the poles are functions of z_{tip}, and so the frequency of each resonance shifts with z_{tip}. This is clearly seen in a false color plot of $\mathrm{Im}\,\chi^\perp(\omega, z_{\mathrm{tip}})$ [Fig. 2.6(a)], where each mode creates a bright curve. All the curves are red-shifted from ω_{SP} but converge to it at large z_{tip}. The smallest $z_{\mathrm{tip}} = 0.02a$ in Fig. 2.6(a) is limited by the accuracy of our numerical calculation. Based on our analytical results we expect that at smaller z_{tip} the resonance curves are shaped as parabolas that approach $\omega_{\mathrm{TO}} = 576\,\mathrm{cm}^{-1}$

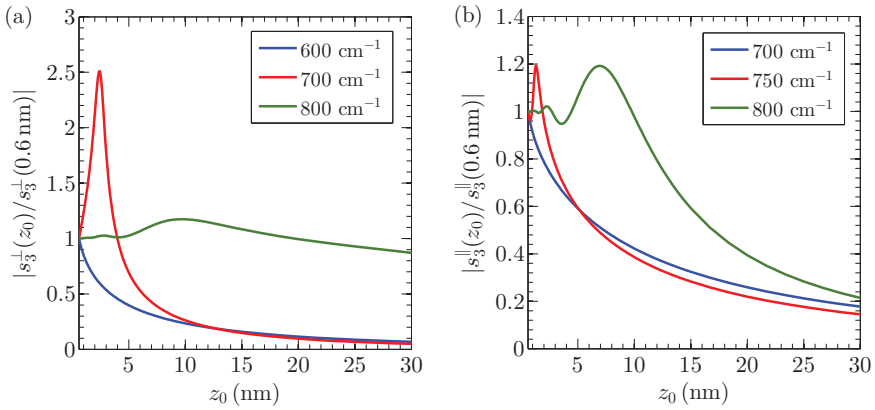

Figure 2.7: (Color online) Approach curves of $|s_3^{\nu}|$ for bulk Al_2O_3, normalized to the value at $z_0 = 0.6$ nm, for several characteristic frequencies. The lowest frequencies in both (a) and (b) are such that no resonance curves are crossed during the probe tapping motion. The approach curves are monotonic. For the middle pair of frequencies one crossing (of the $k = 0$ resonance) does occur. At such crossing each approach curve has a peak. The last pair corresponds to the frequencies where $|\bar{s}_3^{\nu}|$ is close to the maximum value in the spectral range studied. The approach curves have several peaks because of multiple resonance crossings.

where $\mathrm{Re}\,\beta = 1$, cf. Eqs. (2.5), (2.57), and (2.58). A horizontal line cut through Fig. 2.6(a) taken at $z_{\mathrm{tip}} = 0.6\,\mathrm{nm}$ is plotted in Fig. 2.6(b) along with the absolute value of $\chi^{\perp}$. The strongest peak in this plot corresponds to the $k = 0$ mode. The multiple weaker peaks at higher frequencies are produced by $k > 0$ modes.

Next we consider the effects of demodulation on the s-SNOM signal, which can be understood as follows. As the probe oscillates, it spends most time at the minimum and maximum distances from the surface. One therefore expects peaks in χ_n^{ν} at frequencies near those of $\chi^{\nu}(z_0,\omega)$ and $\chi^{\nu}(z_0+2\Delta z,\omega)$. This gives two frequencies per each resonant mode. Actually, the number of observable peaks is smaller. Indeed, from Figs. 2.6(a) and 2.6(d) one can see that all the resonance curves modes should merge together at $z = z_0 + 2\Delta z$ for typical $\Delta z \sim 50\,\mathrm{nm}$. Hence, all the modes should produce a single common peak in the demodulated signal from such z. Furthermore, while the peaks of $\chi^{\nu}(z_0,\omega)$ are distinct, only a few strongest of them can survive the smearing effect of the demodulation. These expectations are supported by Fig. 2.6(c), where we plot the normalized quantities $\bar{\chi}_3(\omega,z_0) \equiv \chi_3/\chi_3^{\mathrm{ref}}$ and $\bar{s}_3$ for $\nu = \perp$, assuming tapping

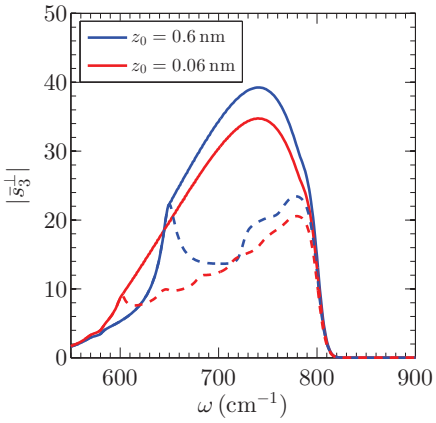

Figure 2.8: (Color online) Comparison of spectra of the quantity $|\bar{s}_3^{\perp}|$ using two different experimental protocols, at two different minimum approach distance z_0. The value at each frequency is taken either from the maximum of the $|s_3^{\perp}|$ approach curve (solid) or from a fixed z_0 (dashed).

amplitude $\Delta z = 50$ nm, $z_0 = 0.6$ nm, and Si as the reference material. In Fig. 2.6(c) we see only three peaks. The peak at 650 cm^{-1} in $|\bar{s}_3|$ is produced by the dominant $k = 0$ mode. It has the same frequency as the $k = 0$ peak in Fig. 2.6(b). The second peak near $725\,\mathrm{cm}^{-1}$ in $|\bar{s}_3|$ (which looks more like a shoulder in $\bar{\chi}_3$) is produced by the $k = 1$ mode at the $z_{\mathrm{tip}} = z_0$ point. The remaining third peak at $787\,\mathrm{cm}^{-1}$ is produced collectively by all the modes. A similar correspondence between the resonance curves of the polarizability function and the peaks in the demodulated signal is found in the $\nu = \parallel$ component, cf. Figs. 2.6(d)–(f). However, the lower $k = 1$ peak is now very weak and is considerably blurred by the demodulation, Fig. 2.6(f). Should we have considered a model with smaller dissipation, this and other high-order peaks would have been more clearly distinguishable in $|\bar{s}_3|$. Note that although the normalized and demodulated signal strength is comparable for the two ν components, the polarizability for $\nu = \parallel$ is orders of magnitude smaller so its contribution can be safely ignored.

The discussion above pertain to horizontal cuts of $\chi^{\nu}(z_{\mathrm{tip}}, \omega)$. Taking a fixed-frequency (vertical) cut through Fig. 2.6(a), and performing the demodulation for a range of minimum distances z_0, one obtains the $\nu = \perp$ approach curve for the scattering signal. An intriguing result

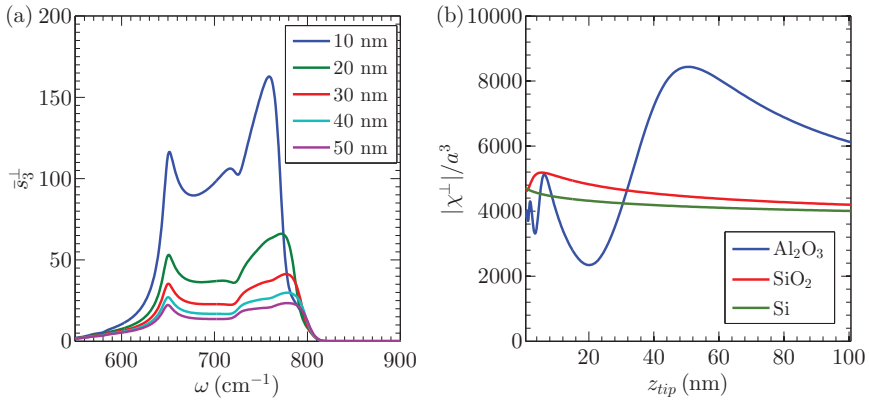

Figure 2.9: (Color online) (a) Spectra of $|\bar{s}_3^\perp|$ for $z_0 = 0.6\,\text{nm}$ and five different tapping amplitudes. The magnitude of $\bar{s}_3^\perp$ increases rapidly with decreasing Δz. (b) The $\chi^\perp$ approach curves for Al_2O_3, SiO_2 and Si, taken at frequencies corresponding to the largest peak in $|s_3^\perp|$ (790 cm^{-1} for Al_2O_3, 1120 cm^{-1} for SiO_2, and an arbitrary ω for the frequency-independent case of Si). For Al_2O_3 sample, $|\chi^\perp|$ shows multiple oscillations; for SiO_2 sample, it has a single maximum at small z_{tip}; for Si, the approach curve decays monotonically with z_{tip}. As Δz decreases, the approach curves become increasingly different.

of this analysis is the possibility of a nonmonotonic dependence of the approach curve on z_0. The nonmonotonicity is due to the crossing of the resonance curves of $\chi^\perp$ by the vertical line cut. Such crossings are found between ω_{TO} where $\text{Re}\,\beta = 1$ and ω_{SP} where $\text{Re}\,\beta$ reaches its maximum. Near the low-frequency end of this interval, the $k = 0$ mode should be again dominant. It is expected to produce a peak in the approach curve, which would follow the same trajectory as the $k = 0$ curve in Fig. 2.6(a), moving to larger z_0 as ω increases. Higher order modes should appear at frequencies closer to ω_{SP} and produce weaker peaks at smaller z_0. The amalgamation of these peaks give rise to the nonmonotonicity of the approach curve.

We show in Fig. 2.7(a) the s_3 approach curves for $\nu = \perp$ for three frequencies. All the curves are normalized to their value at their left ends, $z_0 = 0.6\,\text{nm}$. The approach curve for $\omega = 600\,\text{cm}^{-1}$ decays monotonically with increasing z_0 because the cut at such ω does not cross any of the resonances. In the approach curve for $700\,\text{cm}^{-1}$, a strong peak is seen at around 2 nm due to the crossing of the $k = 0$ resonance. The last approach curve, for $800\,\text{cm}^{-1}$ contains a series of oscillations at small z_0 and a broad hump at large z_0, due to the multiple resonance

crossings. The approach curves for $\nu = \parallel$ plotted in Fig. 2.7(b) exhibit the same general trends as those for $\nu = \perp$.

The striking multi-peak spectra and anomalous nonmonotonic approach curves we described above stem from the large r_P of Al_2O_3 and are not found in less resonant materials, see Sec. 2.8 and Ref. [MKG+14]. This rich structure is also quite sensitive to the choice of z_0. If this parameter is too large, the peaks in the spectrum of the scattering signal merge together at $\omega = \omega_{SP}$. If z_0 is too small, the resonance curves become very flat at $\omega < \omega_{SP}$, so the corresponding peaks are smeared by demodulation and dwarfed by the ω_{SP} peak. Hence, there exists an optimal value of z_0 that allows one to resolve multiple peaks most clearly. For our Al_2O_3 model this value is actually not too far from $z_0 = 0.6\,\text{nm}$ used in Fig. 2.6. For example, the s_3 spectrum for a smaller $z_0 = 0.06\,\text{nm}$ is shown in Fig. 2.8 (dashed lines), where the $k = 0$ peak is much less pronounced while more higher order peaks become distinguishable and form small steps. For even smaller z_0 the steps are further smoothed, eventually leaving only one peak near ω_{SP}.

In addition to the value of z_0, many other experimental parameters and procedures can significantly alter the resultant spectrum. For instance, the experimental determination of z_0 based solely on the s-SNOM approach curve can be inaccurate due to its possible nonmonotonicity, as discussed in the previous Section. It is generally incorrect to ascribe $z_0 = 0$ to the probe position at which the near-field signal has the highest amplitude. Such a protocol effectively yields a frequency-dependent z_0. The difference from the spectra taken for a truly constant z_0 can be drastic, as illustrated in Fig. 2.8. Conversely, the strong sensitivity of the near-field signal to the value of probe-sample distance may perhaps be used for a more accurate measurement of z_0 (although this may require knowing the curvature radius a and perhaps other details of the probe shape).

The tapping amplitude Δz is another parameter that affects the spectrum. When Δz is small, the demodulation at nth order is roughly equivalent to taking the nth order derivative of $\chi^\nu(z_{tip})$. Therefore, a material with a sharply varying approach curve yields a stronger demodulated signal

than the material with a smoothly varying one. In our case the signal of Al_2O_3 is normalized against Si, whose polarizability decays monotonically with z_{tip} [Fig. 2.9(b)]. As Δz decreases, the polarizability of Al_2O_3 become increasingly oscillatory, while that of Si remains smooth. This results in the increased contrast of the demodulated signal for the two materials for smaller Δz [Fig. 2.9(a)].

Other than these controllable parameters, the scattering signal is also dependent on the dielectric function of the probe itself. The calculation in the preceding discussion is done for a perfectly conducting probe, $\varepsilon_{\text{tip}} = \infty$. In practice, near-field probes often have a Si core and a layer of metallic coating whose thickness $\sim 20\,\text{nm}$ can be smaller than the skin depth, i.e., the electric field penetration length of the metal. In this case, it may be more appropriate to set $\varepsilon_{\text{tip}} = \varepsilon_{\text{Si}} \approx 11.7$ in Eq. (2.21). Repeating the calculations, we find that while qualitative features in the signal are retained, there are major quantitative differences (Fig. 2.10).

The discussion above shows that the rich structure of the s-SNOM signal found for the case of Al_2O_3 sample is susceptible to many experimental parameters. (Retardation effects, discussed later in Sec. 2.8, introduce further significant dependence on the probe geometry.) This presents a serious challenge to realistic modeling of s-SNOM experiments. On the other hand, these strong dependences arise only for highly crystalline material with low dissipation. For other, less resonant materials, the modeling can be quite robust, as discussed in Sec. 2.8.

2.7 Nonlocal reflection function

The example material of the previous Section is a bulk crystal with a local (momentum independent) reflectivity function. However, in many other systems studied through s-SNOM, including thin films, graphene, and multi-layered systems reflection is inherently nonlocal. Thus, it is imperative to study how the q-dependence of the reflectivity affects the probe-sample interaction. As mentioned in Sec. 2.4, a general description of such interaction is challenging

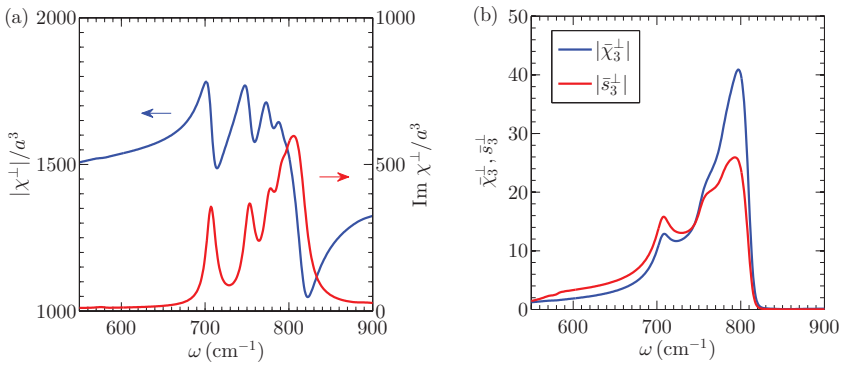

Figure 2.10: (Color online) (a) $\chi^\perp$ and (b) $|\bar{\chi}_3^\perp|$ and $|\bar{s}_3^\perp|$ spectra of Al_2O_3 for the case of a Si probe. All other parameters are the same as in Fig. 2.6(b) and (c). The spectra retain the same structure as for a metallic probe ($\varepsilon_{\mathrm{tip}} = \infty$).

because the series representation of the polarizability

$$\chi = \sum_k \frac{\mathcal{R}_k}{\lambda_k} \tag{2.59}$$

has generalized eigenvalues λ_k and residues $\mathcal{R}_k$ that are now complicated functionals of r_{P} [λ_k^{-1} is the k-th eigenvalue of the matrix $(\mathbf{\Lambda}^m)^{-1/2}\mathbf{H}(\mathbf{\Lambda}^m)^{-1/2}$, cf. Eq. (2.20)]. Still, we can attempt to analyze these expressions using the simple perturbation theory developed in Sec. 2.4, in which λ_k are computed from the poles of the q-independent theory, with corrections obtained by integrating the weighting functions over the momentum. As shown below, this scheme produces qualitative agreement with the calculated s-SNOM response for graphene on bulk Al_2O_3.

The Al_2O_3/graphene system has two collective modes (the upper and the lower one) that emerge from hybridization of the surface phonon of Al_2O_3, originally at $\omega_{\mathrm{SP}} \approx 750\,\mathrm{cm}^{-1}$ with the plasmon of graphene, $\omega(q) \propto \sqrt{\mu v_F q}$. (Coupling of substrate phonons to graphene plasmons has been probed by s-SNOM experiments with graphene/SiO_2 systems. [FAB+11, FRA+12] This and related work is reviewed in Ref. [BFL+14].) The modes share the optical weight and exhibit a level-repulsion that causes both to be dispersive. Both features depend on the chemical potential μ of graphene. Below we focus on the upper mixed mode and study its s-SNOM response for a

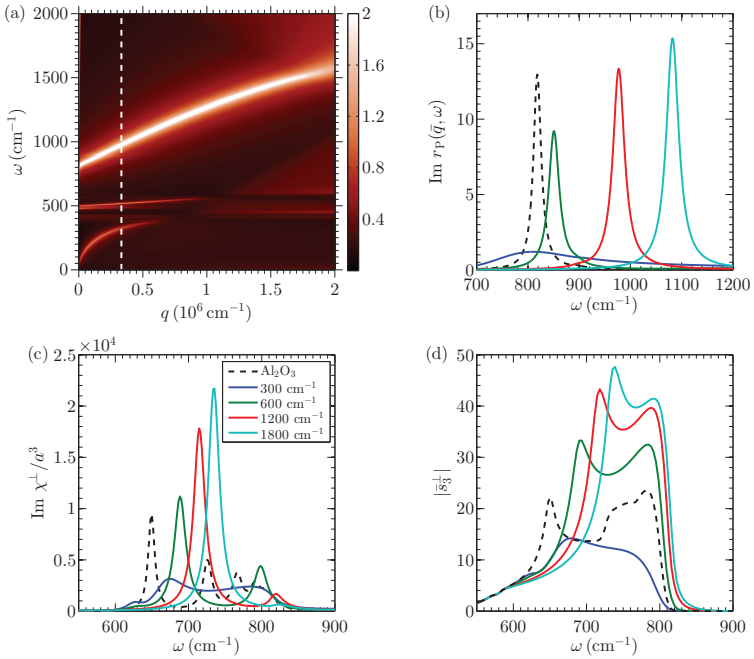

Figure 2.11: (Color online) (a) Collective mode dispersion of graphene/Al_2O_3 system. The mode repulsion between the graphene plasmon and the Al_2O_3 surface phonons are evident. The false color stands for $\mathrm{Im}\,r_\mathrm{P}(q,\omega)$, which is a measure of power dissipation. [FW84] This quantity is additionally raised to power 0.35 to reduce the contrast. The vertical dashed line marks $q = 1/a$. The faint curve just below $\omega = 500\,\mathrm{cm}^{-1}$ is a weak surface phonon [ZPM13] that we do not discuss. The chemical potential of graphene is $\mu = 1200\,\mathrm{cm}^{-1}$. (b) The solid curves are constant momentum $\bar{q} = 1/a$ line cuts through maps like (a) for several μ. The particular case of (a) is shown by the red curve (second solid curve from the right). The dashed curve is the same quantity computed for bulk Al_2O_3 without graphene. (c) $\mathrm{Im}\,\chi^\perp(\omega)$ and (d) $\bar{s}_3^\perp(\omega)$ computed using the q-dependent $r_\mathrm{P}(q,\omega)$ at $z_\mathrm{tip} = 0.02a$ and $z_0 = 0.02a$, respectively. Graphene chemical potentials μ for (b)–(d) are indicated in the legend of panel (c).

range of μ, and compare the results with the perturbation theory method. To proceed, we need the formula for the reflectivity of the composite system. This formula is well-known (see, e.g., Ref. [FAB$^+$11])

$$r_\mathrm{P}(q,\omega) = \frac{\dfrac{\varepsilon_1}{k_1^z} - \dfrac{\varepsilon_0}{k_0^z} + \dfrac{4\pi\sigma}{\omega}}{\dfrac{\varepsilon_1}{k_1^z} + \dfrac{\varepsilon_0}{k_0^z} + \dfrac{4\pi\sigma}{\omega}}\,. \tag{2.60}$$

Here $\varepsilon_1 = \varepsilon_\mathrm{eff}$ [Eq. (2.57)] is the permittivity of the lower half-space (Al_2O_3), $\varepsilon_0 = 1$ is that of the upper half-space (vacuum), $k_j^z = \sqrt{\varepsilon_j \frac{\omega^2}{c^2} - q^2}$ is the z-component of the wave vector in medium

$j = 0, 1$, and $\sigma = \sigma(q, \omega + i\tau^{-1})$ is the conductivity of graphene, which we calculate within the random phase approximation [WSSG06, HDS07] with a finite relaxation time $\tau^{-1} = 25\,\mathrm{cm}^{-1}$. For $q \gg \omega/c$, one finds $k_j^z \simeq iq$ and Eq. (2.60) reduces to

$$r_{\mathrm{P}}(q,\omega) = \frac{\varepsilon_1 - 1 + 4\pi q \dfrac{i\sigma}{\omega}}{\varepsilon_1 + 1 + 4\pi q \dfrac{i\sigma}{\omega}}, \tag{2.61}$$

which can be compared to Eq. (2.57). A convenient way to visualize the dispersion of the collective modes is to plot the imaginary part of $r_{\mathrm{P}}(q,\omega)$, which represents the power dissipation in the system, [FW84] as a false-color map. An example for $\mu = 1200\,\mathrm{cm}^{-1}$ is shown in Fig. 2.11(a). In the low-q regime ($\hbar v_F q \ll \hbar\omega \ll \mu$),[WSSG06, HDS07] the lower bright curve is mainly the plasmon with dispersion $\omega \propto \sqrt{\mu v_F q}$, while the upper bright curve represents the dispersionless Al_2O_3 surface phonon. (The additional bright curve around $\omega = 500\,\mathrm{cm}^{-1}$ is a weaker Al_2O_3 surface phonon, which we do not discuss.) An increase in μ leads to a steeper dispersion of the plasmon, which causes both hybrid modes to go up in frequency. Decreasing μ has the opposite effect. Additionally, if μ drops below $\hbar\omega_{\mathrm{SP}}/2 \approx 380\,\mathrm{cm}^{-1}$, the upper mode falls into the interband transition region of graphene, which results in strong damping of the surface phonon. As we will see below, this causes the $\mu = 300\,\mathrm{cm}^{-1}$ curve to look qualitatively different from the rest in Fig. 2.11(b). Let us now discuss how the collective modes manifest themselves in the s-SNOM response.

In the simplistic picture of the s-SNOM response, the probe-sample interaction is dominated by a single momentum $\bar{q} = 1/a$. If this assumption were accurate, we could set $r_{\mathrm{P}}(\bar{q},\omega)$ as $\beta(\omega)$ and calculate the response using the set of poles and residues established previously. We would then see peaks in the response generated by the upper hybrid mode. However, this crude approximation leads to higher peak frequencies than the calculation using the full $r_{\mathrm{P}}(q,\omega)$, as seen in Figs. 2.11(b), 2.11(c), and 2.11(d). Indeed, we have shown in Sec. 2.4 that when the q-dependence in reflection is treated as a perturbation, each mode has its own range of sensitive

momenta due to the inherent length scales in its potential distribution. The distributions change with an additional length scale — the tip-sample distance z_{tip}, so that the momentum weighting functions are dependent on z_{tip} as well, $G_k = G_k(q, z_{\text{tip}})$. For each mode, these functions provide a means to average over momentum and find an effective q-independent sample reflection $\beta_k^{\text{eff}}(\omega)$, cf. Eq. (2.44), so that we can again apply the established pole-residue decomposition. Strictly speaking, the perturbative method cannot be applied here as the mixed mode may be strongly q-dependent. Even so, we find a very reasonable agreement with the computed signal in the range of graphene chemical potentials $\mu = 600$–$1800\,\text{cm}^{-1}$ that we study. We first consider peak frequencies in $\text{Im}\,\chi^{\perp}$, which can be predicted by invoking the resonance condition $\text{Re}\,\beta_k^{\text{eff}} = \beta_k$. For the lowest mode $k = 0$ and $z_{\text{tip}} = 0.02\,a$, there is a systematic overestimate of the peak position by 20–$30\,\text{cm}^{-1}$ for $\mu = 600$–$1800\,\text{cm}^{-1}$. The discrepancy is larger for higher μ at which the q-dependence of the upper hybrid mode is stronger. This discrepancy is due in part to the well-known general tendency of the first-order perturbation theories to overestimate the lowest eigenvalues. Next, for the $k = 1$ mode, the resonance condition is satisfied only for $\mu = 600\,\text{cm}^{-1}$ at $\omega = 797$ cm^{-1} and μ=1200 cm^{-1} at $\omega = 823$ cm^{-1}, which agree well with the smaller peaks in $\text{Im}\,\chi^{\perp}$. At these frequencies $\text{Im}\,\beta_1^{\text{eff}}$ are larger than the $k = 0$ case and the peaks have smaller magnitudes. For $\mu = 1800$ cm^{-1}, the resonance condition is not met and the very small peak at $\omega = 827$ cm^{-1} in $\text{Im}\,\chi^{\perp}$ corresponds to where $\text{Re}\,\beta_1^{\text{eff}}$ is largest and thus closest to β_1. Finally, for $k > 1$, β_k is larger than $\text{Re}\,\beta_k^{\text{eff}}$ for all frequencies and no peaks in $\text{Im}\,\chi^{\perp}$ are found. Seeing qualitative agreement in the polarizability, we proceed to analyzing the demodulated signal.

As inferred in Sec. 2.6, the demodulated signal is strongest near the peaks in $\chi^{\perp}(z_0, \omega)$ and $\chi^{\perp}(z_0 + 2\Delta z, \omega)$, where each peak is attributed to a resonant mode. For the dominant $k = 0$ mode, we find a set of corresponding peaks in $s_3^{\perp}(z_0, \omega)$ at the same frequencies as those in $\chi^{\perp}(z_0, \omega)$, as shown in Fig. 2.11(c) and Fig. 2.11(d). For the other set of peaks in the s_3 spectra we must consider how the situation is changed at $z_0 + 2\Delta z$. At such distances z_{tip} itself becomes the primary length scale and the sensitivity function G_k is shifted toward smaller momentum, where

the upper mode has a flatter dispersion and its frequency is close to ω_{SP} of the bulk Al_2O_3 crystal. Therefore, this set of peaks should all appear near ω_{SP}, which is indeed the case. Repeating this procedure for the $k = 1$ mode, we find that the peaks it contributes are inseparable from the set of higher frequency peaks produced by the $k = 0$ mode as both have frequencies very close to ω_{SP}. Its contributions, however, alter the heights of these peaks. For instance, the $k = 1$ peak is strongest in $\chi^{\perp}(z_0, \omega)$ for $\mu = 600\,\mathrm{cm}^{-1}$ (among the four we used), so the high frequency peak in $s_3^{\perp}$ for this chemical potential has the largest relative magnitude with respect to the low frequency peak. Thus we conclude the demodulated s-SNOM signal can be qualitatively explained by the perturbative method, albeit with inaccuracy in the lower peak frequency. However, as we argued in Sec. 2.6, the lower frequency peak in the demodulated signal is mainly an artifact of the finite z_0 we are forced to use. If z_0 were truly zero, only the peak near ω_{SP} would survive.

2.8 Model-dependent effects

The spheroid model differs from real s-SNOM probes in two important ways: i) the real probe resembles an inverted pyramid, ii) at infrared wavelengths, the length $\sim 10\mu$m of the probe exceeds several times the diameter c/ω of the radian sphere. In previous literature it was assumed that these differences can all be neglected as the probe-sample interaction is focused around the apex of the probe [Fig. 2.1(b)], while contribution from the rest of the probe is canceled out during the process of demodulation and normalization. Hence, the exact shape of the probe is unimportant and the only relevant physical quantity is the apex radius of curvature a. Further, since the characteristic length scale a is well within the radian sphere, a quasistatic description should suffice. This simplistic argument is backed by previous agreement between the spheroid model and experiment. [DFM$^+$14, FRA$^+$12, ZAF$^+$12] However, we have shown that different probe shapes exhibit universal behavior only when z_{tip}/a is of the order of a few percent (cf. Fig. 2.4(a)). This range is much smaller than typical tapping amplitudes, so the

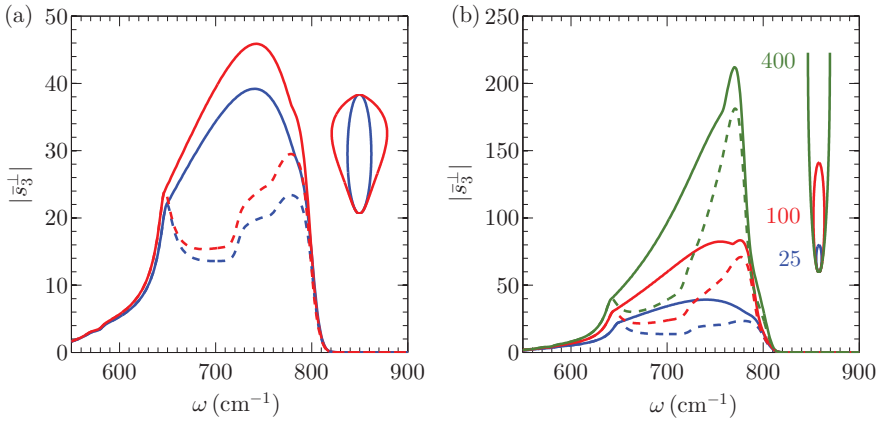

Figure 2.12: (Color online) (a) The s-SNOM signal $s_3^{\perp}$ computed for Al_2O_3 samples. The inset shows the probe shapes used (spheroidal, with $L = 25a$, and a pear-shaped). The two types of probes produce qualitatively similar but quantitatively different results. (b) Spheroids of longer length have drastically increased signal strength. The inset depicts the probe shape and the values of L/a used. Note that this quasistatic calculation neglects radiative damping and antenna resonances, see Sec. 2.8. If included, such effects are expected to greatly reduce $s_3^{\perp}$. In all cases $\Delta z = 50\,\mathrm{nm}$ and $a = 30\,\mathrm{nm}$. The value of s_3 is taken either from the maximum of the approach curves at each frequency (solid lines) or at the closest approach distance $z_0 = 0.6\,\mathrm{nm}$ (dashed lines).

majority of the s-SNOM response lies outside the universality regime and should indeed be probe shape dependent. Additionally, recent experiment and modeling have shown that a quasistatic formalism with *ad hoc* probe shapes is insufficient for highly resonant materials such as on silicon carbide. [MKG+14]

In this Section we re-examine these issues by examining two materials, the highly resonant Al_2O_3 and the dissipative SiO_2, and study the probe shape dependence of their response as well as electrodynamic corrections. We find that for dissipative materials shape dependence is weak and retardation effects are of less importance, so the spheroid model describes the s-SNOM experiment reasonably well. This explains the success of our model in reproducing the response of various materials in experiment. On the other hand, we find the response of resonant materials to be highly dependent on the probe shape and less well described within the quasistatic approximation. For such materials a full electrodynamic treatment with the exact probe shape may be required.

Common numerical methods suitable for electrodynamic treatment of light scattering by a spheroid near a surface include T-matrix method [WD98, DEW99] and BEM. [MKG$^+$14] For the case of a sphere near a surface, the calculation of necessary matrix elements can be done efficiently using recursion technique similar to what we use here.[BV86]

We consider the probe shape dependence and the retardation effects separately. To study the former, we simulated the s-SNOM signal of Al_2O_3 samples obtained with spheroidal probes of different length. We also calculated (using BEM) the results for pear-shaped probes that may better mimic the inverted pyramids. As shown in Fig. 2.12(a), the signal for a pear-shaped probe is qualitatively similar to that for the spheroid of the same length, but there are quantitative differences. For spheroids, we find that the signal strongly increases and the peak frequencies steadily decrease as the length of the probe increases at a fixed apex radius, as shown in Fig. 2.12(b). These features can be explained by the scale invariance of the problem. It implies that an increase in probe length is equivalent to a simultaneous decrease in tapping amplitude and the apex radius. The decrease in radius produces changes in both the poles and residues. The former explains the shift in peak frequencies. The latter is mostly canceled out by normalization. In turn, the decrease in tapping amplitude leads to a larger contrast between the sample and the reference as discussed in Sec. 2.6 [see Fig. 2.9(b)], so the signal strength is dramatically increased.

The strong probe-shape dependence found above seem to suggest that theoretical modeling of the s-SNOM experiments must always be done using the actual shape to be reliable. In fact, such a sensitivity to the probe shape pertains only to the highly-resonant, i.e., large β materials. In Al_2O_3 this parameter reaches the maximum value of $|\beta| \approx 12$, Fig. 2.2(a). For comparison, in Fig. 2.13(b) and Fig. 2.13(c), we show that the pear-shaped probe and the spheroid produced almost identical signals for amorphous SiO_2, a material with $|\beta| \leq 1.5$. (For experimental studies of this material see, e.g., Refs. [AK11a, ZAF$^+$12].) In this case, a factor of 16 increase in the probe length leads to only a doubled signal strength, compared to a nearly tenfold increase for

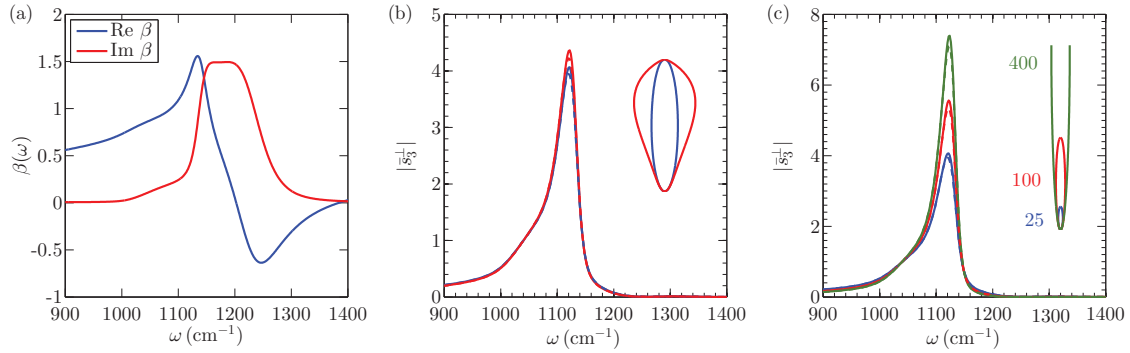

Figure 2.13: (Color online) (a) The reflection coefficient of SiO_2 [KN94] has a larger imaginary part than Al_2O_3 due to its inherent dissipation, leading to a weaker shape dependence in the s-SNOM signal. (b) The signal of the pear-shaped probe is very close to that produced by the spheroid. (c) Increasing the probe length leads to a much smaller increase in the signal strength. The overall shape of the spectrum is also preserved. All geometric parameters are the same as in Fig. 2.12.

Al_2O_3 seen in Fig. 2.12(b).

The results above are obtained within the quasistatic approximation. In reality, a probe half-length of $200a$ already exceeds the diameter c/ω of the radian sphere and one has to consider retardation effects. Naively, contributions from such effects should be eliminated by demodulation, as they pertain to a length scale much larger than the tapping amplitude. However, we show that one contribution — the radiative damping — survives demodulation. The radiative damping has an effect similar to a finite $\mathrm{Im}\,\beta$, i.e., the dissipation in the sample. Hence, for dissipative materials one can neglect radiative damping and still find reasonable agreement with experiment, while doing so for highly resonant materials may lead to qualitatively wrong results. Let us illustrate these statements using the simplest model for the probe — the point dipole. The electrodynamic interaction between the dipole and the sample with the dielectric constant ε is given [FW84] by a modified version of Eq. (2.12),

$$g^{\nu} = c^{\nu} \int_0^{\infty} \frac{iq^3}{k_0^z(q)} \frac{\varepsilon k_0^z(q) - k_1^z(q)}{\varepsilon k_0^z(q) + k_1^z(q)} e^{-2qz_{\mathrm{tip}}} dq\,, \tag{2.62}$$

where the second fraction in the integrand is the full form of the reflectivity $r_{\mathrm{P}}(q,\omega)$. [It is

obtained from Eq. (2.60) by setting σ to zero.]

Suppose ω and z_{tip} are fixed, then the above integral defines g^{ν} as a function of ε, which is generally a complex number. Alternatively, g^{ν} is a function of $\beta = (\varepsilon - 1)/(\varepsilon + 1)$. The integration domain Eq. (2.62) includes momenta q both inside and outside the light-cone. The radiative damping effect arises from the integration over former, i.e., the momenta $q < k_0^z$. This part of the integral yields a negative imaginary contribution to g^{ν}, which shifts the pole of χ^{ν} [Eq. (2.11)] to the lower complex half-plane of β. The real parts of the poles also change but this is less conceptually important, see below. Consider now the remaining part of the integral, over momenta $q > k_0^z$. It is easy to see that if $\varepsilon = -q^2/(q^2 - \omega^2/c^2)$, then

$$\varepsilon k_0^z(q) + k_1^z(q) = 0\,, \tag{2.63}$$

so that there is a pole on the integration path. As a result, functions g^{ν} and χ^{ν} have branch cuts at $\varepsilon \in (-\infty, -1]$ in the complex ε plane or equivalently at $\beta \in [1, \infty)$ in the complex β plane. These additional features are shown schematically in Fig. 2.2(c). The physical origin of both the poles and the branch cut is quite clear. The discrete poles has been discussed at length in this article. They correspond to the polariton modes localized near the tip, Fig. 2.1(b). In turn, the branch cut corresponds to the continuum of *delocalized* surface polaritons that exist without the probe. Indeed, Eq. (2.63), is the well-known equation for the spectrum of such excitations. [FW81]

Of the two features, the branch cut is not expected to affect the signal as the small-momentum contribution is greatly diminished by demodulation. Demodulation should also make less important the change in the real parts of the poles, because these real parts vary greatly with z_{tip} on account of the tapping motion of the probe. However, the shift of the discrete poles away from the real axis is a qualitative change and its effects remain after demodulation. Our next objective is therefore to find this shift for the case of the spheroidal probe.

A free standing spheroid has an effective polarizability given by

$$\chi_{0,\mathrm{eff}} = \frac{\chi_0}{1 - i\frac{2}{3}(\frac{\omega}{c})^3\chi_0} \tag{2.64}$$

to the lowest order in ω/c when radiative correction is considered. [WGL82, Mor09] Modifying $\mathbf{\Lambda}$ accordingly [specifically Λ_1^m, cf. Eq. (2.29)], it is easily shown that this formula applies to our geometry as well. Namely, the s-SNOM polarizability corrected for the radiative damping is given by

$$\chi^\nu_{\mathrm{rad}} = \frac{\chi^\nu}{1 - i\frac{2}{3}(\frac{\omega}{c})^3\chi^\nu}, \quad \chi^\nu = \sum_{k=0}^{\infty} \frac{R_k^\nu}{\beta_k^\nu - \beta}. \tag{2.65}$$

Viewed in the complex β plane, this correction is equivalent to the shift of the poles β_k into the lower half-plane by $-i(2/3)(\omega/c)^3 R_k$ (to the leading order in ω/c). Therefore, both the radiative damping and the intrinsic dissipation in the sample play a similar role: they increase the distance from the poles to the curve traced by the surface reflectivity β of the sample as ω varies [Fig. 2.2(c)]. For a dissipative material, the curve begins far from the poles, and so further increase in the distance produces little change. Conversely, for highly resonant materials the $\beta(\omega)$ curve passes close to the real axis, and so radiative damping may obscure or eliminate the fine features of the signals, such as multiple resonant peaks discussed in Sec. 2.6. It is worth noting however that while it may be important for s-SNOM in infrared or visible domains, the radiative damping should be rather weak in the (experimentally more challenging) terahertz range, where typical s-SNOM probes would fit well inside the radian sphere.

Finally, a class of retardation effects we have not addressed here are antenna resonances arising when the length of the probe exceeds several times the diameter of the radian sphere. They give rise to additional peaks in the s-SNOM signal as a function of ω. For most materials such resonances are removed once the s-SNOM signal is normalized to a reference sample; however, for strongly resonant materials such as SiC and presumably also Al_2O_3 we studied here, the cancellation is not complete. [MKG+14]

2.9 Discussion and Conclusion

Further progress in the s-SNOM and related areas of near-field microscopy requires a quantitatively reliable procedure for determining the fundamental response function $r_{\mathrm{P}}(q,\omega)$ from the amplitude and phase of the s-SNOM scattering data, from which one can proceed to the next step of inferring the optical constants of the studied sample. Typically, materials with a higher absolute value of $r_{\mathrm{P}}(q,\omega)$ produce a higher amplitude s-SNOM signal. However, the peaks in the s-SNOM signal are often red-shifted with respect to those in $|r_{\mathrm{P}}(q,\omega)|$ or $\mathrm{Im}\, r_{\mathrm{P}}(q,\omega)$.

Given additional information about the system, these inverse problems can be tackled by fitting the experimental data to the solution of the direct problem with a trial form of $r_{\mathrm{P}}(q,\omega)$ as the input. [MKG$^+$14] Unfortunately, the direct problem is also difficult to solve. The three-dimensional nature of this problem and the presence of widely different length scales make realistic simulations [PJALR03, EVK06, EVK09] of s-SNOM experiments very computationally intensive. This led to popularity of simple *ad hoc* approximations known as the point-dipole [HK00, TKH04, ATdA$^+$08] and the finite-dipole model, [Oce07, AK11a, AK11b, HET12] in which the actual charge distribution induced on the probe is approximated by a point-like image dipole or a combination thereof with additional point charges.

The point-dipole model [KH04] postulates that Eqs. (2.9a)–(2.12) that are rigorous in the asymptotic long-distance limit $z_{\mathrm{tip}} \gg L$ remain qualitatively correct at much shorter z_{tip} if the input physical parameters are suitably renormalized. Thus, the bare polarizabilities χ_0^{ν} become the adjustable parameters of the model. It is customary to assume that the in-plane polarizability $\chi_0^{\parallel}$ is negligible compared the out-of-plane one, which is taken to be

$$\chi_0^{\perp} = a^3, \tag{2.66}$$

where a is of the order of the curvature radius of the tip. Another adjustable parameter [RGEH05,

FAB$^+$11] $b \lesssim 1$ specifies the position of the effective dipole inside the probe:

$$z_{\rm p} = ba + z_{\rm tip}\,. \tag{2.67}$$

Clearly, the point-dipole model accounts only for the sharp tip and ignores the body of the probe, as $\chi^{\rm v}$ for the point-dipole in Eq. (2.11) is much smaller than $\chi_0^{\rm v}$ for a tip with $L \gg a$. If the point-dipole model were literally correct, the radiating dipole of the probe in typical s-SNOM experiments would be so small that no measurable signal would be observed.

The finite-dipole model improves upon the point-dipole one by including the missing antenna-like enhancement approximately. It assumes that the electric field of a spheroidal probe of length $2L$ is equivalent to that of several point charges of total zero charge that are positioned inside the spheroid near both of its ends. For small $z_{\rm tip}/L$, this model [COH07, AK11b] yields the following functional form of the probe polarizability:

$$\chi^{\rm fdp} = {\rm const} + \frac{R_0^{\rm fdp}}{\beta_0^{\rm fdp} - \beta}\,, \quad \beta_0^{\rm fdp} \approx 1.4 + O\left(\frac{z_{\rm tip}^3}{L^3}\right)\,, \tag{2.68}$$

where $R_0^{\rm fdp} \propto aL^2$. The finite-dipole model was shown to give a good qualitative agreement with s-SNOM data obtained for quartz, amorphous SiO_2, and SiC samples once parameters $R_0^{\rm fdp}$ and $\beta_0^{\rm fdp}$ are suitably adjusted. [AK11a] Thus, the best fit to the data was achieved choosing the length $2L = 600\,{\rm nm}$ of the probe, which is about one third of the diameter $c/\omega \approx 1700\,{\rm nm}$ of the radian sphere. Interestingly, this is approximately the value of $2L$ in the quasistatic calculation for which one obtains, in the case of SiO_2 sample, the same result for s_3 as one gets from the full electrodynamic calculation for a probe of a realistic (much longer) length. [MKG$^+$14]

Agreement with the data notwithstanding, from the theory point of view Eq. (2.68) is unsatisfactory on at least three counts. First, $R_0^{\rm fdp}$ does not follow the correct scaling $L^3/\ln L$ as a function of L, thus underestimating the probe polarizability. Second, the constant term in Eq. (2.68) violates the general requirement that $\chi \to 0$ as $\beta \to \infty$, corresponding to the case when

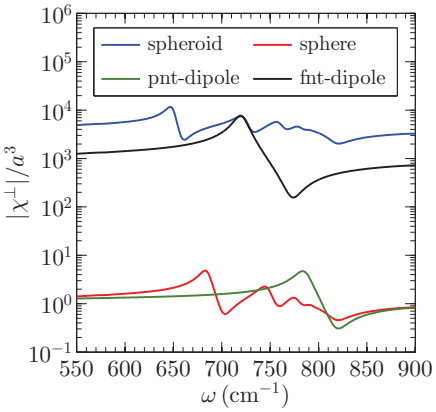

Figure 2.14: (Color online) The spectrum of the probe polarizability $|\chi^{\perp}|$ for Al_2O_3 sample according to four different models. The point- and the finite-dipole models each predict a single peak in $|\chi^{\perp}|$. The calculations for spherical and spheroidal probes reveal multiple peaks. The sphere and the point-dipole models produce $\chi^{\perp}/a^3 \sim 1$. The $L = 25a$ spheroidal probe yields $\chi^{\perp} \sim 10^3$–10^4; the finite-dipole of the same L gives about an order of magnitude lower $\chi^{\perp}$. These dramatic differences in both the form and the absolute magnitude of $\chi^{\perp}$ can however be significantly reduced in the usually reported $\bar{s}_3^{\perp}$, the normalized demodulated signal.

the applied field is screened completely by the induced charges in the sample. Third, β_0^{fdp} goes to ~ 1.4 when $z_{\mathrm{tip}} = 0$. Instead, all smooth probe shapes must behave as a sphere at $z_{\mathrm{tip}} \ll a$, and therefore yield $\beta_0 = 1$ at $z_{\mathrm{tip}} = 0$. The fact that finite-dipole model violates these general requirements suggest its limited usability. Figure 2.14 is an illustration of how widely different the predictions of the four discussed s-SNOM models can be for the case of Al_2O_3. Additional examples of similarly large differences for SiO_2 and SiC samples can be found in previous works of the present authors and their collaborators. [ZAF+12, MKG+14] All these examples compel us to conclude that the prior success of the point- and finite-dipole models in fitting experimental data has to be due to insufficient range of the data, multitude of adjustable parameters, and also the demodulation and normalization procedures that mask the errors in both the functional form and the magnitude of the calculated signal.

Another way to explain the difference between the earlier *ad hoc* models and our GSM is as follows. For the case of a sample with a local reflectivity β, the exact scattering problem of a dielectric probe near a surface reduces to a generalized eigenproblem, [AKSV99] that has an

infinite number of eigenmodes, as we discussed in Sec. 2.1. In contrast, both the point- and the finite-dipole models attempt to approximate the *infinite* number of eigenmodes by a *single* one.

Since the real-space potential distribution of the eigenmodes [Fig. 2.1(b)] depends on the shape and size of the probe and probe-sample distance but not on β, we can describe interaction of the probe with an arbitrary sample efficiently using the precalculated basis of such eigenmodes. This allows one to use our GSM approach to model s-SNOM response for a wide range of materials. However, calculations for realistic probe shapes are not always practical. In search of a broadly applicable yet simple model, we have chosen the prolate spheroid to be our probe shape, as it captures the essential features of the actual probes — a sharp apex and a strongly elongated shaft. We quantified the eigenmodes of the probe-sample system in the form of poles and residues of the polarizability functions χ^{ν}, Section 2.11, allowing an expedient, in fact, instantaneous calculation of the s-SNOM reponse. The point-dipole, finite-dipole and other *ad hoc* models no longer have the advantage of computational speed and should now be considered obsolete.

Recent work [MKG$^+$14] has shown that in the strong-coupling regime of the probe-sample interaction a fully electrodynamic treatment using the BEM and realistic probe shape is necessary in order to reproduce the measurements. This regime is realized experimentally [AK11a, MKG$^+$14] when using samples of SiC, a material for which $|\beta|$ can be as high as 15. The same considerations apply for Al_2O_3 for which $|\beta|$ can reach 12, see Fig. 2.2(a). Our GSM theory gives analytical insight into near-field response of such materials. We have shown that due to simultaneous excitation of multiple eigenmodes, novel features of the s-SNOM signal such as multi-peaked spectra and nonmonotonic approach curves can appear. These features are however very sensitive to experimental parameters such as tapping amplitude, minimum approach distance, and even the data collection protocol. Retardation effects, especially radiative damping can also qualitatively alter the signal and must be considered. In order to observe the predicted anomalous approach curves and multi-peak spectra, it may be necessary to make efforts to minimize the radiative damping, which requires working with shorter probes or at lower frequencies. In contrast,

in the weak- and moderate-coupling regimes, which are relevant for the vast majority of samples, the lowest-order eigenmode is dominant. Hence, the approach curves should be monotonic in z_{tip}, while the spectra should be mostly insensitive to experimental details and retardation effects. This is the regime where our spheroidal probe model can be used with the greatest confidence.

Our GSM theory also applies to a more complicated problem where the sample reflectivity is nonlocal, i.e., momentum-dependent. Here the salient advantages of our method are two-fold. First, in the case of a weak nonlocality, our GSM provides a mapping of the nonlocal problem to a local one. Thereby the sample-independent eigenmode decomposition is retained, providing an intuitive interpretation of the scattering signal. Second, our numerical algorithm (see Supplementary online materials) is much more efficient than the standard BEM because the number of necessary matrix element calculations scales linearly instead of quadratically with the matrix size. It will be worthwhile to compare the actual computational speed of our algorithm with that of a recently developed and significantly more efficient BEM that utilizes pre-calculated matrix elements. [MKG$^+$14]

We hope that the improved physical understanding of near-field probe-sample coupling enabled by the generalized spectral method advanced in this work as well as the numerical procedures we developed for its implementation can be of use for modeling and analysis of future s-SNOM and other near-field experiments.

2.10 The electrostatic problem of a spheroidal probe

The electric field created by a spheroidal object is most conveniently described in the prolate spheroidal coordinates (ξ,η,ϕ) where the origin of the coordinate system is located at the center of the probe, as shown in Fig. 3. The relationships to the cylindrical polar coordinates (ρ,ϕ,z) are

$$z = F\,\xi\eta\,, \quad \rho = F\sqrt{\xi^2-1}\sqrt{1-\eta^2}\,. \tag{2.69}$$

In the spheroidal coordinate each spatial position is specified by $\xi \in [1,\infty)$, $\eta \in [-1,1]$, and ϕ is the usual azimuthal angle. Contours of constant ξ are a series of concentric spheroids centered at the origin, with the major axis along the z direction and common foci at $z = \pm F$. For each such spheroid, ξ is equal to the ratio of its major semi-axis and focal length. We consider the case when the surface of the probe coincides with one of the spheroidal surfaces $\xi = \xi_0 = L/F$, where L is the half-length or major semi-axis of the probe. Related quantities such as the minor semi-radius W of the probe or the radius of curvature a at the apex are given by $W = \sqrt{L^2 - F^2}$ and $a = W^2/L$.

It is well known that Laplace's equation $\nabla^2\Phi = 0$ has separable solutions in the prolate spheroidal coordinates. In particular, we are interested in solutions outside a spheroidal probe that decay at large ξ. Their most general form is written in terms of a linear combination of spheroidal harmonics as follows:

$$\Phi_{\text{sphd}}(\xi,\eta,\phi) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l} B_l^m \mathsf{P}_l^m(\eta) e^{im\phi} \frac{P_l^m(\xi_<) Q_l^m(\xi_>)}{P_l^m(\xi_0)}, \tag{2.70}$$

where B_l^m are coefficients to be determined from boundary condition, $\mathsf{P}_l^m(\zeta)$ is the associated Legendre polynomial defined on the interval $[-1,1]$ and $P_l^m(\zeta)$ and $Q_l^m(\zeta)$ are the associated Legendre function of the first kind and second kind (See, e.g. Ref. [GR00]) with

$$\xi_< \equiv \min(\xi,\xi_0), \quad \xi_> \equiv \max(\xi,\xi_0). \tag{2.71}$$

With the above definition of $\xi_{>,<}$, Eq. (2.70) covers both inside and outside the surface of the probe at ξ_0.

For the geometry considered in this paper [Fig. 1(a)], the total potential Φ can be written as

$$\Phi = \Phi_0 + \Phi_{\text{plane}} + \Phi_{\text{sphd}}, \tag{2.72}$$

where

$$\Phi_0(\boldsymbol{r}) = -\boldsymbol{E}_0\boldsymbol{r} \tag{2.73}$$

is the potential of the external uniform field, and Φ_{plane} is the potential due to charges in the sample, which can be decomposed into evanescent plane wave as

$$\Phi_{\text{plane}}(\boldsymbol{r}) = \int B(\boldsymbol{q}) e^{-qz} e^{i\boldsymbol{q}\boldsymbol{\rho}} \frac{d^2q}{4\pi^2}\,, \tag{2.74}$$

where the position vector $\boldsymbol{r} = \boldsymbol{\rho} + z\hat{\boldsymbol{z}}$ is broken up into its cylindrical polar coordinate components. We determine $B(\boldsymbol{q})$, as well as B_l^m from boundary conditions.

To do that we quote two well-known mathematical results: the decompositions of evanescent plane waves in terms of spheroidal harmonics and vice versa. The first reads

$$e^{i\boldsymbol{q}\boldsymbol{\rho}-qz} = \sum_{l=0}^{\infty}\sum_{m=-l}^{l} \frac{2l+1}{2}(-)^l i^m \frac{(l-m)!}{(l+m)!}\sqrt{\frac{2\pi}{qF}} I_{l+\frac{1}{2}}(qF) P_l^m(\xi)\mathsf{P}_l^m(\eta) e^{im(\phi-\phi_q)}\,, \tag{2.75}$$

with ϕ and ϕ_q being the azimuthal angles of $\boldsymbol{\rho}$ and $\boldsymbol{q}$, respectively. The reverse is:

$$Q_l^m(\xi)\mathsf{P}_l^m(\eta) = (-)^l i^m \frac{(l+m)!}{(l-m)!}\int \frac{\pi F}{q}\sqrt{\frac{2\pi}{qF}} I_{l+\frac{1}{2}}(qF) e^{i\boldsymbol{q}\boldsymbol{\rho}+qz} e^{-im(\phi-\phi_q)} \frac{\mathrm{d}^2q}{4\pi^2}\,, \tag{2.76}$$

where $I_\nu(z)$ is the modified Bessel function of the first kind. [MF81] These two relations follow easily from addition theorems of general Legendre functions such as those in Ref. [GR00] and [Hob29].

Near the sample surface $z = -z_{\text{p}}$, the boundary condition is

$$\tilde{\Phi} = B(\boldsymbol{q}) e^{-qz_{\text{p}}} + \tilde{\Phi}_{\text{sphd}}(\boldsymbol{q}, z_{\text{p}}) \propto e^{qz_{\text{p}}} - r_{\text{P}}(q) e^{-qz_{\text{p}}}\,, \tag{2.77}$$

where we use the notation

$$\tilde{f}(\boldsymbol{q}, z) = \int f(\boldsymbol{r}) e^{-i\boldsymbol{q}\boldsymbol{\rho}} d^2\rho \tag{2.78}$$

for a partial Fourier transformation. Eq. (2.77) implies:

$$B(\boldsymbol{q}) = -r_{\mathrm{P}} e^{-qz_{\mathrm{p}}} \tilde{\Phi}_{\mathrm{sphd}}(\boldsymbol{q}, z_{\mathrm{p}}) \,. \tag{2.79}$$

The other boundary condition for a uniform spheroidal probe with dielectric constant $\varepsilon_{\mathrm{tip}}$ is:

$$\left.\frac{\partial \Phi}{\partial \xi}\right|_{\xi \to \xi_0^+} = \varepsilon_{\mathrm{tip}} \left.\frac{\partial \Phi}{\partial \xi}\right|_{\xi \to \xi_0^-} . \tag{2.80}$$

Boundary conditions in Eqs. (2.79) and (2.80) with the decompositions in Eqs. (2.75) and (2.76) allow one to compute the unknown coefficients B_l^m. The result can be summarized by first defining an infinite matrix:

$$H_{ll'} \equiv 2\pi \int_0^\infty r_{\mathrm{P}}(q) I_{l'+\frac{1}{2}}(qF) I_{l+\frac{1}{2}}(qF) e^{-2qz_{\mathrm{p}}} \frac{dq}{q} \,, \tag{2.81}$$

whose elements are integrals of $r_{\mathrm{P}}(q)$. Then for each integer m a quantity related to B_l^m, a column vector defined by:

$$A^m{}_l \equiv (-)^{l+m} \frac{(l+m)!}{(l-m)!} \frac{B_l^m}{F} \,, \tag{2.82}$$

is the solution to the linear system of equations

$$\sum_{l'=1}^{\infty} (\Lambda^m - H)_{ll'} A^m{}_{l'} = b^m{}_l \,. \tag{2.83}$$

The elements of the diagonal matrix $\Lambda^m{}_{ll'} = \Lambda_l^m \delta_{ll'}$ are defined by

$$\Lambda_l^m = \frac{(-1)^m}{2l+1} \frac{4}{\varepsilon_{\mathrm{tip}} - 1} \left[\varepsilon_{\mathrm{tip}} \frac{Q_l^m(\xi_0)}{P_l^m(\xi_0)} - \frac{\frac{d}{d\xi_0} Q_l^m(\xi_0)}{\frac{d}{d\xi_0} P_l^m(\xi_0)} \right] \tag{2.84}$$

and the numbers on the right-hand side of the equations are given by

$$b^m{}_l = \frac{4}{3}\frac{(1+m)!}{(1-m)!}C^m\delta_{l1}\,, \tag{2.85}$$

where $C^0 = -E_z$ and $C^{\pm 1} = (E_x \mp iE_y)/2$. Thus, the form of matrix $\mathbf{\Lambda}$ is determined completely by the geometry of the probe (in terms of ξ_0) and its dielectric constant $\varepsilon_{\rm tip}$, while $\mathbf{H}$ describes the interaction between the sample reflection function $r_{\rm P}$ and the momentum selectivity of the modes. The column vector $\mathbf{b}$ describes the uniform external field.

The coefficients B_l^m can be obtained directly from Eq. (2.82) after one solves for $A^m{}_l$ from Eq. (2.83). But for the purpose of determining the induced probe dipole moment, only $|m| \le 1$ cases are important. By examining the asymptotic behavior of the electrostatic potential $\Phi(\boldsymbol{r})$, one obtains the total dipole moment of the spheroid probe. Its Cartesian components are related to the components of $A^m{}_l$ by

$$\begin{aligned} p_{{\rm sp},0} &= p_{{\rm sp},z} = -\frac{F^3}{3}A^0{}_1\,, \\ p_{{\rm sp},1} &= \frac{p_{{\rm sp},x} - ip_{{\rm sp},y}}{2} = \frac{F^3}{3}A^1{}_1\,. \end{aligned} \tag{2.86}$$

If the probe is made of an ideal conductor, $\varepsilon_{\rm tip} \to \infty$, then Eq. (2.84) simplifies to

$$\Lambda_l^m = (-1)^m \frac{4}{2l+1}\frac{Q_l^m(\xi_0)}{P_l^m(\xi_0)}\,, \tag{2.87}$$

which are all positive numbers. If the dielectric function of the probe is real and negative, then Λ_l^m can be negative, too. It can also be zero, which corresponds to a plasmon (or phonon) polariton resonance of the probe. The resonances occur at discrete $\varepsilon_{\rm tip}$ values

$$\varepsilon_{{\rm tip},l}^m = \frac{\frac{d}{d\xi_0}\ln Q_l^m(\xi_0)}{\frac{d}{d\xi_0}\ln P_l^m(\xi_0)}\,, \quad l = 1,2,\ldots, \tag{2.88}$$

see, e.g., Refs. [CKC$^+$14, SGRBF15] (where, in fact, a more general case of anisotropic ε_{tip} is treated). For each m, the sequence $\varepsilon^m_{\text{tip},l}$ asymptotically approaches -1 as $l \to \infty$. The smallest, i.e., the most negative value in each sequence is the starting one. It can be alternatively written as

$$\varepsilon^m_{\text{tip},1} = 1 - \frac{1}{L^\nu}\,, \tag{2.89}$$

where $\nu = \perp$ for $m = 0$, $\nu = \parallel$ for $m = 1$, and L^ν are the depolarization factors of the spheroid [BH04]

$$L^\perp = (\xi_0^2 - 1)\left[\frac{1}{2}\xi_0 \ln\left(\frac{\xi_0 + 1}{\xi_0 - 1}\right) - 1\right], \tag{2.90}$$

$$L^\parallel = \frac{1 - L^\perp}{2}\,. \tag{2.91}$$

For prolate spheroids, these depolarization factors obey the inequalities $0 < L^\perp < L^\parallel < \frac{1}{2}$, and so $\varepsilon^0_{\text{tip},1} < \varepsilon^1_{\text{tip},1}$. For example, if $L = 25a$, which we use in our calculations below, then $\xi_0 = \sqrt{25/24}$, $\varepsilon^0_{\text{tip},1} = -16.9$, and $\varepsilon^1_{\text{tip},1} = -1.11$. If the probe is made out of platinum or iridium, which are common materials for AFM tips, its dielectric function can indeed be negative. It is in principle possible to achieve plasmonic resonances in such probes somewhere in the near-infrared or visible spectral range. On the other hand, at mid-infrared frequencies, for which we do calculations in this paper, the dielectric functions of such metals are in the range of hundreds or thousands. Such probes are very far from any of the plasmonic resonances and the approximation of the ideal conductor, Eq. (2.87), can be safely used.

Table 2.1: Coefficients of the nine-pole rational fits for $L = 25a$.

k	a_5	a_4	a_3	a_2	a_1	a_0
0	3	−36.399	234.56	−762.76	1783.1	−0.015667
1	5	−25.733	111.01	−93.002	290.11	0
2	7	−33.029	157.55	−118.29	1391.3	0
3	9	−36.251	173.3	−40.018	1879.9	0
4	11	−47.517	221.75	−85.286	2292.3	−0.00017314
5	13	−45.678	233.45	42.551	3435.5	−0.000094847
6	15	−46.254	254.23	223.28	2547	0
7	17	−27.808	235.93	770.95	1961.3	0
8	19	−65.583	251.72	−308.28	402.67	0.000032595

k	b_4	b_3	b_2	b_1	b_0
0	1	−10.345	83.048	−417.42	1522.2
1	1	−3.4964	27.841	11.949	87.231
2	1	−0.9961	−2.6274	149.14	253.32
3	1	−0.12625	−8.2396	185.03	246.31
4	1	−0.3866	−8.4205	180.3	237.17
5	1	1.0728	−20.329	253.41	291.59
6	1	−0.75722	7.941	175.21	185.07
7	1	−1.8957	34.251	163.71	122.51
8	1	1.6624	4.104	−1.1364	23.24

k	c_5	c_4	c_3	c_2	c_1	c_0
0	3.9999	303.23	5141.1	4811.1	282.17	1.4941
1	12.001	916.75	17089	33226	2371.9	12.255
2	24.001	1844.7	35207	90005	36.881	11.908
3	40	3166.2	65393	288144	224067	4417.7
4	60.028	4584.9	87200	196974	5632.8	28.67
5	84.685	6304.9	122561	226585	4364.5	4.237
6	116.26	8216.9	166927	316568	98948	843.4
7	146.83	10354	214160	350367	77249	334.56

k	d_3	d_2	d_1	d_0
0	1	0.77084	0.023552	0.000027594
1	1	1.4472	0.11961	0.0003052
2	1	1.9006	0.012624	0.00025669
3	1	4.1622	2.6905	0.09134
4	1	1.5802	0.066244	0.00048749
5	1	1.2714	0.044389	0.000077993
6	1	1.4088	0.44188	0.0087351
7	1	1.1606	0.28331	0.0034808

2.11 Rational fits of poles and residues

The polarizability χ^ν of the probe has the form

$$\chi^\nu(\beta) = \sum_{k=0}^{\infty} \frac{R_k^\nu}{\beta_k^\nu - \beta}, \quad \nu = \perp \text{ or } \parallel, \tag{2.92}$$

for a sample that has momentum independent reflection, $r_P(q) = \beta$. The poles β_k^ν and residues R_k^ν depend only on the geometry of the probe and not on the sample, so a single set of poles and residues can be used for any β. Following the method described in Sec. 2.10, we calculated and fitted the first nine poles and residues of a spheroidal probe with $L = 25a$ for $\nu = \perp$. Coefficients of the fit are cataloged in Table. 2.1, where the poles are given by

$$\log \beta_k = \sum_{i=0}^{5} a_i \alpha^i \Big/ \sum_{i=0}^{4} b_i \alpha^i, \tag{2.93}$$

and the residues are given by

$$R_k a^{-3} \mathcal{Z}^{-1} = \sum_{i=0}^{5} c_i \mathcal{Z}^i \Big/ \sum_{i=0}^{3} d_i \mathcal{Z}^i, \tag{2.94}$$

with $\mathcal{Z} \equiv z_{\rm tip}/a$, valid in the range $0.003 < \mathcal{Z} < 10$. The fits for the residues apply only to the first eight poles, $k = 0$ through 7. The remaining residue R_8 is constrained to obey the sum rule,

$$\sum_{k=0}^{8} \frac{R_k}{\beta_k} = \chi_0 \,. \tag{2.95}$$

2.12 The spherical probe limit

The spheroidal probe model presented in Appendix 2.10 is quite general and can be a good model for tips of any aspect ratio L/a. Here we explore a particular limit of $F \to 0$ and $\xi \to \infty$ while keeping the product $F\xi \to a$ constant. This corresponds to the problem of a spherical probe of radius a. The derivation in Appendix 2.10 simplifies to that in Sec. 4.1 of Ref. [FW84], and by using the following asymptotic forms of various special functions:

$$\sqrt{\frac{2\pi}{qF}} I_{l+\frac{1}{2}}(qF) \simeq \frac{2^{2l+1} l!}{(2l+1)!} \left(\frac{qF}{2}\right)^l , \tag{2.96a}$$

$$Q_l^m(\xi) \simeq (-)^m \frac{2^l l!(l+m)!}{(2l+1)!} \xi^{-l-1} , \tag{2.96b}$$

$$P_l^m(\xi) \simeq \frac{(2l)!}{2^l l!(l-m)!} \xi^l , \tag{2.96c}$$

one can show that the decompositions Eqs. (2.75) and (2.76) reduce to Eq. (4.9) and (4.10) of Ref. [FW84]. The characteristic equation Eq. (2.83) for ${B^m}_l$ reduces to:

$$\sum_{l'=1}^{\infty} \left\{ \frac{\delta_{ll'}}{\alpha_l a^{2l+1}} - \frac{(l+l')!}{(l+m)!(l'-m)!} \mathcal{F}_{l+l'} \right\} {\bar{B}^m}_{l'} = \frac{\delta_{l1}}{(l-m)!} C^m , \tag{2.97}$$

where

$${\bar{B}^m}_l = (-)^{l+m} (l+m)! \frac{2^l l!}{(2l+1)!} F^{l+1} {B^m}_l \tag{2.98}$$

is similarly related to the induced charge distribution of the probe,

$$\alpha_l = \frac{l(\varepsilon_{\text{tip}} - 1)}{l(\varepsilon_{\text{tip}} + 1) + 1} \tag{2.99}$$

is the multipole polarizability of the probe, and

$$\mathcal{F}_l = \frac{1}{l!} \int_0^\infty r_{\text{P}}(q) q^l e^{-2qd_0} dq \tag{2.100}$$

is the integral that characterizes the interaction between the spherical probe and the sample with $d_0 = a + z_{\text{tip}}$. Eq. (2.97), the characteristic equation for a spherical probe, is derived in Ref. [FW84] as Eq. (4.20). The solution to Eq. (2.97) has some of the same properties as the spheroid case: $\bar{B}^m{}_l = 0$ for all l and $|m| > 1$; $\bar{B}^0_l$ is related to the charge distribution due to the z component of the electric field and $\bar{B}^{\pm 1}_l$ are related to the charge distribution due to the x-y component of the electric field.

For the case of q independent $r_{\text{P}}(q, \omega) = \beta(\omega)$, in which the integrals $\mathcal{F}_l$ reduces to:

$$\mathcal{F}_l = \frac{\beta}{(2d_0)^{l+1}}, \tag{2.101}$$

there is an exact solution to the spherical characteristic equation. Let

$$\alpha = \operatorname{arccosh} \frac{d_0}{a} = \operatorname{arccosh} \left(1 + \frac{z_{\text{tip}}}{a}\right), \tag{2.102}$$

be a dimensionless parameter that characterizes the sphere-to-sample distance relative to its size, and let

$$\sigma_k(\beta; \alpha) = \sum_{m=0}^{\infty} \frac{(2m+1)^k}{e^{(2m+1)\alpha} - \beta}. \tag{2.103}$$

Using the following quantities:

$$p_0 = \chi_0 E_z\,, \quad \chi_0 = a^3\,, \tag{2.104a}$$

$$q_0 = \frac{p_0}{a}\left(\cosh\alpha - \sinh\alpha\frac{\sigma_1}{\sigma_0}\right)\,, \tag{2.104b}$$

$$p_n = p_0\beta^n\left(\frac{\sinh\alpha}{\sinh(n+1)\alpha}\right)^3\,, \tag{2.104c}$$

$$q_n = \frac{\beta^n\sinh\alpha}{\sinh(n+1)\alpha}\left[q_0 - \frac{p_0}{a}\frac{\sinh n\alpha}{\sinh(n+1)\alpha}\right]\,, \tag{2.104d}$$

it can be shown that for a metallic sphere

$$\bar{B}_l^0 = (-)^l\sum_{n=0}^{\infty} q_n(d_0-d_n)^l - p_n l(d_0-d_n)^{l-1}\,, \tag{2.105}$$

where

$$d_0 - d_n = a\frac{\sinh n\alpha}{\sinh(n+1)\alpha}\,. \tag{2.106}$$

The physical meaning of Eq. (2.105) becomes clear when one treats the problem with method of images (Fig. 2.15). Suppose that an external electric field E_z would have induced a bare dipole moment p_0 in the sphere. This would induce an image dipole in the dielectric half-space, which would in turn induce an image dipole and an image charge in the sphere. The position and strength of each successive image dipole and charge can be solved by recursion. Setting the sample surface to $z = 0$, the position of the center of the sphere is at $z = d_0$. At each position $z = d_n$ given by Eq. (2.106) there is a point dipole p_n and a point charge q_n. Charge q_0 is determined by the neutrality condition $\sum_n q_n = 0$, which yields Eq. (2.104b). Summing up all the contributions to the total dipole moment from both the dipoles *and* the point charges inside the

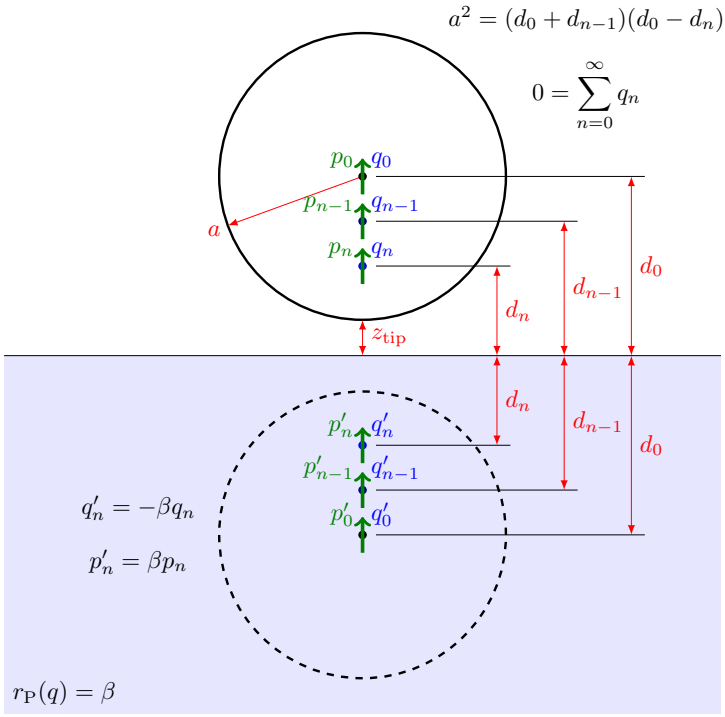

Figure 2.15: The method-of-images solution of the problem of a metallic sphere above a dielectric half-space with the external field normal to the interface. The method involves an infinite series of dipoles p_n and point charges q_n located inside the sphere at distances d_n above the interface. The total charge inside the sphere is zero.

sphere, we get:

$$\begin{aligned}\frac{\chi^{\rm sph,\perp}}{\chi_0} \equiv \frac{p_z^{\rm total}}{p_0} &= \frac{1}{p_0}\sum_{n=0}^{\infty}\left[p_n + q_n(d_n - d_0)\right] \\ &= 2\sinh^3\alpha\left(\sigma_2 - \frac{\sigma_1^2}{\sigma_0}\right).\end{aligned} \tag{2.107}$$

with σ_k given by Eq. (2.103).

The above analysis resulting in Eq. (2.107) is for the case where the electric field is perpendicular to the sample. For the case where the electric field is parallel to the sample, the analysis is simpler in that the positions of the image dipoles and their strength are the same, but no image point charges are present. Therefore, in this polarization:

$$\frac{\chi^{\rm sph,\parallel}}{\chi_0} \equiv \frac{p_{xy}^{\rm total}}{p_0} = \sinh^3\alpha\,(\sigma_2 - \sigma_0)\,. \tag{2.108}$$

Both Eq. (2.107) and (2.108) conforms to our earlier assertion that χ has the form of Eq. (2.92):

$$\chi^{\mathrm{sph},\parallel} = \sum_{k=0}^{\infty} \frac{R_k^{\mathrm{sph},\parallel}}{\beta_k^{\mathrm{sph},\parallel} - \beta}, \quad \chi^{\mathrm{sph},\perp} = \sum_{k=0}^{\infty} \frac{R_k^{\mathrm{sph},\perp}}{\beta_k^{\mathrm{sph},\perp} - \beta}. \tag{2.109}$$

For horizontal electric fields, $\chi^{\mathrm{sph},\parallel}$ is singular whenever σ_2 or σ_0 is, so that

$$\beta_k^{\mathrm{sph},\parallel} = e^{(2k+3)\alpha}. \tag{2.110}$$

The corresponding residues are:

$$R_k^{\mathrm{sph},\parallel} = 4(k+1)(k+2)\chi_0 \sinh^3\alpha. \tag{2.111}$$

For electric fields perpendicular to the sample, the parenthesis in Eq. (2.107) vanishes at each $\beta_k^{\mathrm{sph},\parallel}$, so they are not poles of $\chi^{\mathrm{sph},\perp}$. Instead, $\beta_k^{\mathrm{sph},\perp}$ occur at the zeros of σ_0 which has no simple analytic form. The poles of the two polarizations, however, interleave:

$$\beta_k^{\mathrm{sph},\parallel} < \beta_k^{\mathrm{sph},\perp} \lesssim \beta_{k+1}^{\mathrm{sph},\parallel}. \tag{2.112}$$

2.13 Boundary element method for axisymmetric probes

The spheroidal probe belongs to a large class of probes that has axial symmetry: its cross section is a circle with radius $w(z)$ varying as a function of height. Here we briefly describe the method used to calculate the polarizability χ of such probes. Dividing the interval $z \in [-L, L]$ that the probe occupies into n segments, the j-th segment can be approximated as a ring with radius $w_j \equiv w(z_j)$ and charge Q_j. In the presence of an external field, the charges on rings Q_j satisfies the equation:

$$\sum_{j=1}^{n} \left(L_{ij} + H_{ij}\right) Q_j = -\varphi_i, \tag{2.113}$$

where

$$L_{ij} = \int J_0(qw_i)J_0(qw_j)e^{-q|z_i-z_j|}\mathrm{d}q \tag{2.114}$$

is the quasistatic interaction between the ith and the jth rings,

$$H_{ij} = \int r_\alpha(q)J_0(qw_i)J_0(qw_j)e^{-q(z_i+z_j+2z_{\mathrm{tip}}+2L)}\mathrm{d}q \tag{2.115}$$

is the interaction between the ith ring and the image of the jth ring, and $\varphi_i = \phi_{\mathrm{ext}}(z_i)$ is the external potential at the i-th ring. Defining

$$\mathcal{K}(w,w';z) = \frac{2}{\pi}\frac{K\left[\frac{4ww'}{(w+w')^2+z^2}\right]}{\sqrt{(w+w')^2+z^2}}\,, \tag{2.116}$$

in terms of the complete elliptical integral of the first kind $K(z)$, we can write

$$L_{ij} = \mathcal{K}(w_i, w_j, z_i - z_j)\,, \tag{2.117}$$

while in the case of a q-independent $r_\alpha(q,\omega) = \beta(\omega)$, H_{ij} can be simplified as

$$H_{ij} = \beta\overline{H}_{ij}\,, \quad \overline{H}_{ij} = \mathcal{K}(w_i, w_j, z_i + z_j + 2z_{\mathrm{tip}} + 2L)\,, \tag{2.118}$$

so that Eq. (2.113) takes on the same form as the equation for the spheroid derived in the text,

$$\left(\mathbf{\Lambda}^m - \beta\,\overline{\mathbf{H}}\right)\mathbf{A}^m = \mathbf{b}^m\,. \tag{2.119}$$

Indeed, Eq. (2.119) can be viewed as a special case of Eq. (2.113) expressed in a different basis, and for q-independent $\beta(\omega)$ the polarizability of the probe always has the form of Eq. (2.92). Eq. (2.113) can be easily implemented in a computer. However, for general $r_\alpha(q,\omega)$ the evaluation of each element of H_{ij} as a numerical integral can be slow, making this method computationally

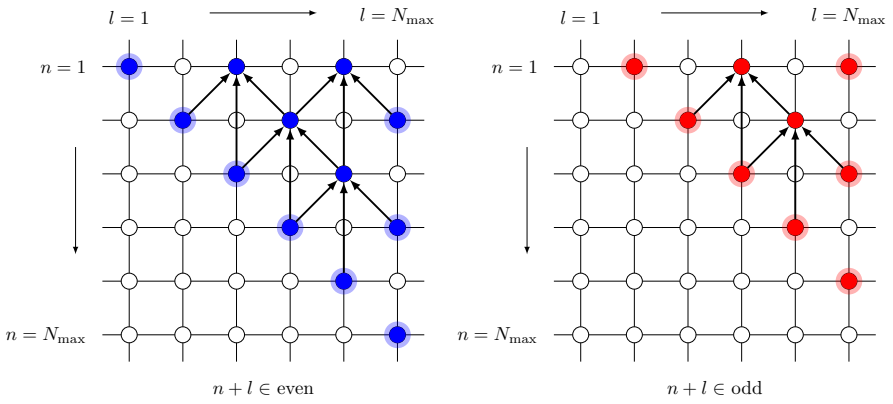

Figure 2.16: Illustration of the algorithm for the recurrence relation of $H_{n,l}$.

demanding when one needs a large matrix size to achieve good accuracy. Nevertheless, it serves well as a benchmark test for comparison with other models. In particular, our spheroidal model yields almost indistinguishable results and is an order of magnitude faster, due to the recursive relation between matrix elements described in Sec. 2.14.

2.14 Numerical implementation for the spheroid model

The integral

$$H_{n,l} \equiv 2\pi \int_0^\infty r_\alpha(q) I_{n+\frac{1}{2}}(qF) I_{l+\frac{1}{2}}(qF) e^{-2qz_p} \frac{dq}{q} \tag{2.120}$$

has no simple analytic form even in the simplest case where the reflection coefficient $r_\alpha(q,\omega)$ does not depend on momentum q. Further, the summation

$$\sum_{l=1}^{\infty} (\Lambda^m - H)_{nl} A^m{}_l = b^m{}_n \,, \tag{2.121}$$

runs through all $l \geq 1$ and one does not expect a closed form solution to the equation. In practice Eq. (2.121) can be solved numerically by truncating the dimension of l at some large $N_{\max}$ and

Table 2.2: Parameters of the optical constant of α-Al_2O_3. The frequency unit is $1\,\mathrm{cm}^{-1}$.

ρ	ε_∞	j	$\omega_{j\mathrm{LO}}$	$\gamma_{j\mathrm{LO}}$	$\omega_{j\mathrm{TO}}$	$\gamma_{j\mathrm{TO}}$
o	3.05	1	908	22.4	569	7.86
		2	482	2.96	439	3.23
		3	387	5.18	384	6.03
e	2.9	1	885	21.6	582	4.17
		2	481	3.21	482	3.42
		3	511	1.42	400	4.68

evaluating each of the $N_{\max}^2$ elements.

Fortunately, not all $N_{\max}^2$ elements of $H_{n,l}$ needs to be evaluated explicitly by integration. First, since $H_{n,l}$ is symmetric one only needs to independently determine the upper triangular elements. Secondly, the four neighboring elements of element $H_{n,l}$ are related to one another via the relation:

$$\frac{H_{n+1,l} - H_{n-1,l}}{2n+1} = \frac{H_{n,l+1} - H_{n,l-1}}{2l+1}\,. \tag{2.122}$$

This relation allows one to compute the remaining one of the four neighbors knowing any three. In fact, by repeated use of this relation all "interior" elements can be determined recursively from the values of the boundary terms as illustrated in Fig. 2.16. A similar construction was employed previously in Ref. [BV87]. Further, as also shown in Fig. 2.16 the upper triangular elements separate into two classes depending on whether the sum $n+l$ is odd or even. Therefore, only $3(N_{\max}-1)$ terms need to be evaluated through integration, consisting of those of the main diagonal $H_{n,n}$, and the next upper diagonal $H_{n,n+1}$, and the last column $H_{n,N_{\max}}$.

2.15 Optical constant of aluminum oxide

The optical constant of α-Al_2O_3 used in our calculations is given by

$$\varepsilon_\rho(\omega) = \varepsilon_{\infty,\rho} \prod_j \frac{\omega_{j\mathrm{LO},\rho}^2 - \omega^2 - i\gamma_{j\mathrm{LO},\rho}\omega}{\omega_{j\mathrm{TO},\rho}^2 - \omega^2 - i\gamma_{j\mathrm{TO},\rho}\omega}, \tag{2.123}$$

where the parameters are cataloged in Table. 2.2. They are adapted from Ref. [ZPM13] for temperature $T = 300\,\mathrm{K}$.

2.16 Acknowledgment

Chapter 2, in full, is a reprint of the material as it appears in Journal of Applied Physics: Jiang, B.-Y.; Zhang, L. M.; Castro Neto, A. H.; Basov, D. N.; Fogler, M. M. "Generalized spectral method for near-field optical microscopy", J. Appl. Phys. 119, 054305 (2016). The dissertation author was the primary investigator and author of this paper.

Chapter 3

Tunable plasmonic reflection by bound 1D electron states in a 2D Dirac metal

3.1 Introduction

Plasmon scattering and plasmon losses in Dirac materials, such as graphene and topological insulators, are problems of interest to both fundamental and applied research. It is an outstanding challenge to understand various kinds of interaction (electron-electron, electron-phonon, electron-photon, electron-disorder) responsible for these complex phenomena [DSAHR11, KUP+12, BFL+14, PCL+14, WLG+14]. At the same time, control of plasmon scattering is critical if this class of materials is to become a new platform for nanophotonics [JBS09, VE11, GPN12, GdA14].

One source of plasmon scattering is long-range inhomogeneity of the electron density, which causes local fluctuations in the plasmon wavelength λ_p. If the inhomogeneities are weak, those of size comparable to the average λ_p are expected to be the dominant scatterers [KDS13, GPNMM13] Surprisingly, recent experiments have revealed that one-dimensional (1D) defects of nominally atomic width can act as effective reflectors for plasmons with wave-

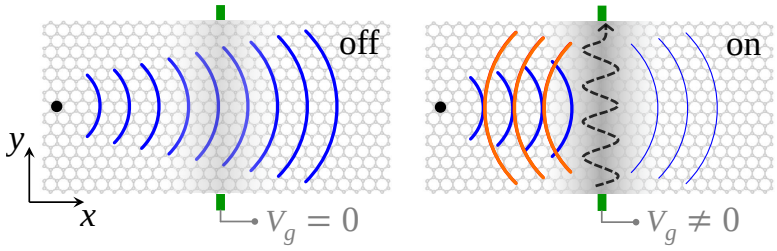

Figure 3.1: (Color online) Schematic of an ultranarrow plasmon reflector. The incident plasmon (blue) can propagate freely unless a local perturbation hosting a 1D electron state (the dashed arrow) causes it to be reflected (orange). The bound state parameters are controlled by voltage V_g of a nanotube gate (green).

lengths as large as a few hundred nm. Strong plasmon reflection was observed near grain boundaries [FRG+13, SCH14], topological stacking faults [JSN+15], as well as nanometer-scale wrinkles and cracks [FRG+13, GPNMM13] in graphene. If this anomalous reflection is indeed an ubiquitous effect largely unrelated to the specific nature of a defect, it calls for a universal explanation. In this Letter we attribute its origin to electron bound states commonly occurring near 1D defects. We show that optical transitions involving the bound states can produce strong dissipation at small distances x from the defect and therefore, alter plasmon dynamics. To support this idea we present a theoretical analysis of an exactly solvable model, which illustrates qualitative and quantitative characteristics of the bound states and predicts how their optical response depends on the tunable parameters of a 1D potential well. We also report an attempt to probe the predicted effects experimentally. Our approach is to employ an ultranarrow electric gate in the form of a carbon nanotube (CNT) to create a precisely tunable 1D barrier in graphene. This device enables a systematic investigation and control of plasmon propagation, including, in principle, an implementation of a plasmon on-off switch (Fig. 3.1). What we find is that the measured real-space profile of the plasmon amplitude (Fig. 3.4) cannot be accounted for by a local change in λ_p alone. Instead, the data are consistent with the presence of an enhanced dissipation in the region next to the CNT. The amount of this dissipation agrees in the order of magnitude with the power absorption due to 1D bound states in our model.

3.2 The model

We assume that the graphene quasiparticles can be described by a 2D Dirac Hamiltonian $H = \hbar v_F (\sigma_z k_x + \sigma_y k_y) + v(x)$, where σ_y, σ_z are the Pauli matrices and $v(x)$ is the total (screened) potential induced by the 1D gate. For simplicity, we assume that $v(x)$ is a square well of width d and depth u although more realistic potentials [Ken02, HP14, CB13, VG03] can also be considered. In the present case the eigenfunctions Ψ are combinations of plane waves and/or exponentials that have to be matched at $x = \pm d/2$, see Section 3.5. The electron momentum k_y along the perturbation (in the y-direction) is conserved, so that the gapless 2D Dirac spectrum is effectively replaced by a 1D one with a gap $\Delta = |\hbar v_F k_y|$. Within the gap electron states localized at the well exist [Fig. 3.2(b)]. The energies $\varepsilon_n(k_y)$ of these bound states, where $n = 1, 2, \ldots$, are the solutions of the transcendental equation [BSAT09]

$$\frac{\tan\sqrt{(E+U)^2 - K_y^2}}{\sqrt{(E+U)^2 - K_y^2}} = \frac{i\sqrt{E^2 - K_y^2}}{K_y^2 - E(E+U)}\,. \tag{3.1}$$

Here $E = \varepsilon_n d/(\hbar v_F)$ is the dimensionless energy and

$$K_y = k_y d\,, \quad U = ud/(\hbar v_F)\,, \tag{3.2}$$

are, respectively, the dimensionless y-momentum and the well depth. The dispersions of the three lowest bound states for $U = 5$ are shown in Fig. 3.2(a).

The response of the system to an optical excitation of frequency ω polarized in the x-direction is described by an effective conductivity $\sigma(x)$ given by the Kubo formula (see Section 3.5), which determines the local current density $j_x(x) = E_x\sigma(x)$ in the approximation that the total electric field E_x due to the optical excitation is uniform. Below we focus on the real part of $\sigma(x)$, which determines local power dissipation. We assume that graphene is doped and consider only frequencies $\hbar\omega < 2|\varepsilon_F|$, for which the optical conductivity of an infinite graphene sheet

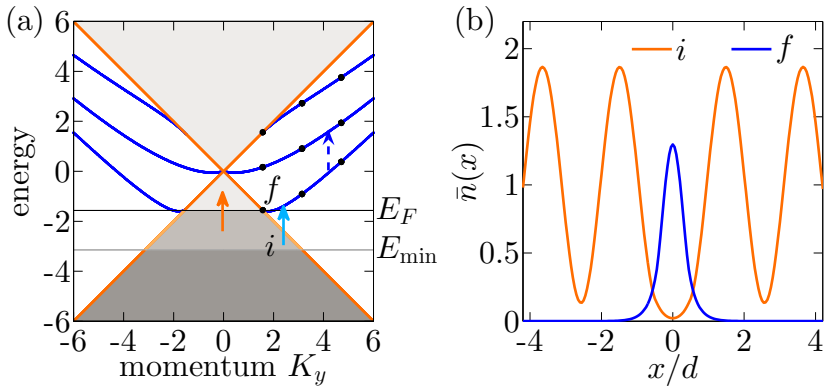

Figure 3.2: (Color online) (a) Dispersion of bound states for a sheet (blue) or a ribbon of width $2d$ (the black dots) for $U = 5$. The light gray are empty states in the continuum. The dark and medium gray are occupied states in the continuum. The last of these, with E between $E_F = \mu d/\hbar v_F$ and $E_{\min} = E_F - \omega d/v_F$, enable optical transitions (the arrows) of frequency ω. Transitions between bound states (the dashed arrow) can occur for some E_F, e.g., $E_F = 0$ at which the state i is filled and the state f is empty. (b) The density distribution $\bar{n} = |\Psi|^2$ of the two states i and f for the transition indicated by the cyan arrow in (a). The state i (blue) is localized in the well, while the state f (orange) is extended. Parameters: $K_y = 2.5$, $\omega d/v_F = \pi/2$.

vanishes (if we neglect disorder, many-body scattering, and thermal broadening [BFL$^+$14]). This implies that in the absence of the perturbation, $U = 0$, we must have $\mathrm{Re}\,\sigma(x) = 0$ at all x. On the other hand, when the potential well is present, a finite $\mathrm{Re}\,\sigma(x)$ exists. There are two types of relevant optical transitions: those that involve the bound states [as either the initial i or the final f states, Fig. 3.2(a)] and those that do not. The contribution of the former to $\mathrm{Re}\,\sigma(x)$ is maximized near the potential well and decays exponentially at $|x| > d/2$ due to the localized nature of the bound states. The contribution of the latter is small, oscillating, and decaying algebraically with x (see Section 3.5). Resolving the detailed real-space features of $\sigma(x)$ in an optical experiment is challenging (see below). A more practical observable is the normalized integrated conductivity:

$$\bar{\sigma} \equiv \frac{1}{d}\int_{-\infty}^{\infty} dx\,\mathrm{Re}\,\sigma(x)\,. \tag{3.3}$$

According to our simulations, transitions that involve the bound states give the dominant contribution to $\bar{\sigma}$. In particular, bound-to-bound state transitions produce numerically large values

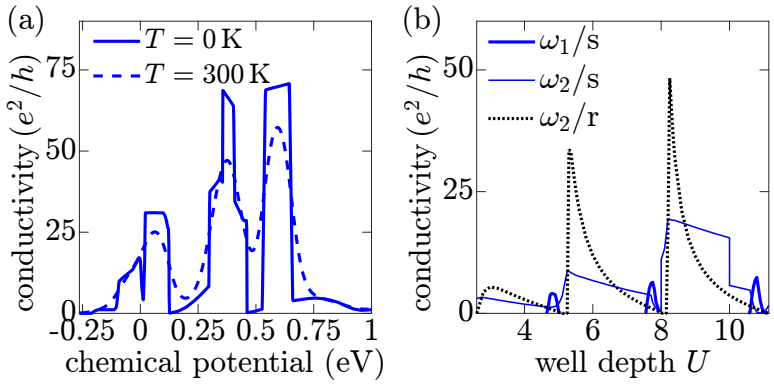

Figure 3.3: (Color online) (a) Integrated conductivity $\bar{\sigma}$ of a graphene sheet at $\omega = 830\,\mathrm{cm}^{-1}$. The sharp changes are caused by blocking/unblocking of the transitions involving bound states as a result of changing occupations of the levels as a function of the graphene chemical potential μ. For example, the plateau at $0.02 < \mu(\mathrm{eV}) < 0.12$ is due to the (blue) dashed-line transition in Fig. 3.2(a). (b) Integrated conductivity $\bar{\sigma}$ of a sheet (s) and a ribbon (r) at $T = 0$ and $K_F = -\pi/2$. Sharp changes at $U = 8$ and 10 for $\omega = \omega_2$ arise from a transition between bound states. Parameters: $d = 10\,\mathrm{nm}$, $\omega_1 = 83\,\mathrm{cm}^{-1}$, $\omega_2 = 830\,\mathrm{cm}^{-1}$.

of $\bar{\sigma}$ expressed in units of e^2/h. Such transitions are possible at discrete k_y where the energy difference between the states of the same momentum matches $\hbar\omega$ provided the lower (higher) state is occupied (empty). If the chemical potential μ is gradually increased, e.g., by electrostatic gating, the state occupations would change, leading to either blocking or unblocking of these transitions. Accordingly, $\bar{\sigma}$ would either sharply drop or jump, see Fig. 3.3(a). These changes persist, albeit blurred, at finite temperatures, see the dashed curve in Fig. 3.3(a).

Sharp drops in $\bar{\sigma}$ also occur when the bound states merge with the continuum and get liquidated (become extended). The drop is abrupt if the optical transitions probe a single k_y or a narrow range of k_y. In principle, this situation can be realized in a graphene ribbon running perpendicular to the linelike perturbation. In such a ribbon the allowed $k_y = m\pi/W + \mathrm{const}$ are discrete, as shown schematically by the dots in Fig. 3.2(b). The coupling to a single bound state can be achieved under the condition $\pi/W > \omega/v_F$, i.e., by using a ribbon of a narrow width W or the excitation of a low frequency ω. In Fig. 3.3(b) we show three numerically calculated traces of $\bar{\sigma}$ as a function of the well depth U for a fixed dimensionless chemical potential $E_F = \mu d/(\hbar v_F) = -\pi/2$. The first trace is computed for a ribbon of width $W = 2d$

probed at the excitation energy $\hbar\omega = |\mu|$. It exhibits pronounced oscillations of $\bar{\sigma}$. In particular, $\bar{\sigma}$ drops to zero when a bound state merges with the continuum. The other two traces correspond to a 2D graphene sheet. Although the sharp drops become blurred, they remain pronounced at a low excitation energy $\hbar\omega_1 = |\mu|/10$ and still evident at $\hbar\omega_2 = |\mu|$.

The enhanced local optical conductivity around the 1D gates described above causes plasmons to be strongly reflected. According to the first-order perturbation theory [FRG+13, GPNMM13] (see Section 3.5), the reflection coefficient r of a normally incident plasmon wave is

$$r_1 \simeq \frac{2\pi i}{\lambda_p} \int_{-\infty}^{\infty} dx \left[\frac{\sigma(x)}{\sigma_\infty} - 1 \right] . \tag{3.4}$$

For arbitrary perturbations, we can use the approximation $|r| \approx \min(|r_1|, 1)$. Using the results of Fig. 3.3(b) we estimate $|r| \approx 0.3$ at the chemical potential of 0.25 eV where the predicted $\bar{\sigma} \approx 5e^2/h$. This roughly corresponds to the regime probed by our experiments (see below). At the chemical potential of 0.3 eV where the calculated local conductivity is much larger, $\bar{\sigma} \approx 40e^2/h$, the reflection coefficient should approach unity, realizing the "reflector on" state in Fig. 3.1.

3.3 Experiment and analysis

To investigate the described above phenomena experimentally we fabricated a nanodevice that contained (bottom to top) a Si/SiO_2 substrate, a 10 nm-thick layer of hexagonal boron nitride (hBN), and a mechanically exfoliated graphene flake. A metallic single-wall CNT was placed between hBN and SiO_2. The local charge density of graphene was tunable by the voltage V_g applied between the CNT and graphene. The average carrier density in graphene $|n| \sim 5 \times 10^{12}\,\mathrm{cm}^{-2}$ was produced by uncontrolled ambient dopants (acceptors) [NWW+15]. To infer the local optical conductivity $\sigma(x)$ we used scattering-type scanning near-field optical microscopy (s-SNOM) [KH09, ABJR12, BFL+14], see Fig. 3.4(a). The s-SNOM utilizes a tip of

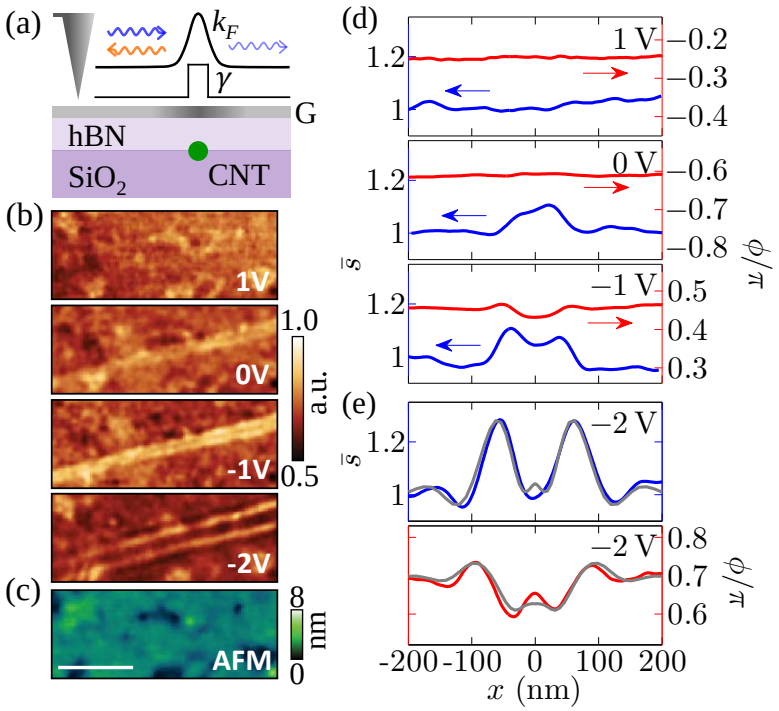

Figure 3.4: (Color online) Measurement of the conductivity $\bar{\sigma}$ by the s-SNOM. (a) A schematic showing graphene (variable intensity gray) gated by a CNT (green) separated from it by a thin hBN layer. The induced perturbation is parameterized by spatially varying k_F and γ. In the experiment, the AFM tip (triangle) is polarized by a focused infrared beam (not shown), which enables it to launch a plasmon (blue). The reflected plasmon (orange) causes an additional tip polarization, resulting in a modified optical signal backscattered by the tip and detected in the far field. (b) The s-SNOM amplitude images of the region next to the CNT for $V_g = +1 \ldots -2\,\mathrm{V}$ and $\omega = 890\,\mathrm{cm}^{-1}$. The twin fringes (bright lines) intensify and separate as $|V_g|$ increases. (c) The AFM topography image of the same region. Scale bar: $1\,\mu$m. (d)-(e) The s-SNOM amplitude ($\bar{s}$) and phase (ϕ) along the line perpendicular to the CNT; $\bar{s}$ is normalized to $x = -200\,$nm point. The best theoretical fits (gray) for $V_g = -2\,\mathrm{V}$ are included in (e).

an atomic force microscope (AFM) with a radus 25 nm as an optical antenna that couples incident infrared light to graphene plasmons. The backscattered light is analyzed to extract the amplitude $\bar{s}$ and the phase ϕ of the genuine near-field signal, Fig. 3.4(b,d,e). Crudely speaking, this signal is proportional to the electric field inside the tip-sample nanogap. The variation of this field with the tip position is caused by the standing-wave patterns of surface plasmons [FRA$^+$12, CBAG$^+$12]. These standing waves are due to the interference of the plasmon waves launched by the tip with the waves reflected by the charge inhomogeneity induced by the CNT. The spacing of the interference fringes is equal to one half of the plasmon wavelength λ_p. The latter is given by $\lambda_p = \mathrm{Re}\,(2\pi/q_p)$, where $q_p(x) = i\kappa\omega/2\pi\sigma(x)$ is the complex plasmon momentum and κ is the average permittivity of the media surrounding graphene [BFL$^+$14]. Therefore, s-SNOM images combined with the formula for q_p give a direct estimate of $\mathrm{Im}\,\sigma(x)$. The extraction of $\mathrm{Re}\,\sigma(x)$ requires an electromagnetic simulation of the coupled tip-graphene system, which was done using the numerical algorithm developed previously [FRA$^+$12, FRG$^+$13] (see Section 3.8). To facilitate connection with that previous work, we parametrized the conductivity via

$$\sigma(x) = \frac{e^2 v_F}{\pi\hbar\omega}\,\frac{ik_F(x)}{1+i\gamma(x)}\,, \tag{3.5}$$

which was modelled after the long-wavelength Drude (intraband) conductivity of graphene [BFL$^+$14] with Fermi momentum k_F and dimensionless damping factor γ. The goal of the data analysis was to determine the profiles of $k_F(x)$ and $\gamma(x)$ that yield the best fit to the s-SNOM data. In this parametrization, the presence of the bound states should increase the local damping, so the signature we were looking for was the enhanced value of $\gamma(x)$.

Our experimental data are presented in Fig. 3.4. The AFM topography image, Fig. 3.4(c), shows that the CNT does not produce any visible topographic features. However, in the near-field signal, up to two pairs of intereference fringes appear on each side of the CNT [the bright lines in Fig. 3.4(b)]. Similar twin fringes have been observed in prior s-SNOM imaging [FRG$^+$13,

SCH14, JSN$^+$15, NWW$^+$15] of linear defects in graphene. Importantly, the intensity and spacing of the fringes we observe here evolve with the CNT voltage V_g, which attests to their electronic (specifically, plasmonic) origin.

In addition to the controlled perturbation induced by the CNT, graphene contains uncontrolled ones due to random defects. To reduce the random noise caused by those, we averaged the near-field signal over a large number of linear traces taken perpendicular to the CNT. Thus obtained line profiles of both the amplitude $\bar{s}$ and the phase ϕ are plotted in Fig. 3.4(d) and (e). We focus on the $V_g = -2\,\mathrm{V}$ trace, which shows the strongest modulation. The accurate determination of functions $k_F(x)$ and $\gamma(x)$ is impacted by the s-SNOM resolution limit $\sim 20\,\mathrm{nm}$. In our fitting we assumed that $k_F(x)$ is given by the perfect screening model, $k_F^2(x) = [k_F^2(0)d^2 + k_F^2(\infty)x^2]/(d^2+x^2)$, which should be a good approximation for high doping [JF15]. The adjustable parameters are $k_F(0)$ and $k_F(\infty)$. For $\gamma(x)$ we considered trial functions in the form of a peak (dip) at $x = 0$, with adjustable width and height (depth), as sketched in Fig. 3.4(a). The trial $k_F(x)$ and $\gamma(x)$ were fed as an input to the electromagnetic solver described previously [FRA$^+$12, FRG$^+$13]. As detailed in Section 3.8, a good agreement with the observed form of the twin fringes requires a strong peak in $\gamma(x)$ near the CNT. The shape of the fringes was found to depend primarily on the integral of $\gamma(x) - \gamma(\infty)$, so in the end we modeled $\gamma(x)$ by a box-like discontinuity with a central region of a fixed width 13.5 nm and two adjustable parameters $\gamma(0)$, $\gamma(\infty)$. The best fits [the gray curves in Fig. 3.4(e)] to the $V_g = -2\,\mathrm{V}$ s-SNOM data were obtained using $\gamma(0) = 1.65$.

To establish a rough correspondence between the profiles of Fig. 3.4(e) and the square-well model we take d to be the thickness of the hBN spacer $d = 10\,\mathrm{nm}$ and U to give the same integrated weight $\int v(x)dx \equiv ud = \hbar v_F U = \hbar v_F \int [k_F(x) - k_F(\infty)]dx$. This prescription implies $E_F = 4$, $U = 13$, and $\bar{\sigma} = 3.5\,e^2/h$ for $\omega = 890\,\mathrm{cm}^{-1} = 1.7 v_F/d$, see Section 3.8. The square-well model in Fig. 3.3 yields a comparable optical conductivity $\bar{\sigma} = 4.7\,e^2/h$ although for a smaller $U = 5$. Given a number of simplifying assumptions we have made in the modelling, this level of agreement seems adequate.

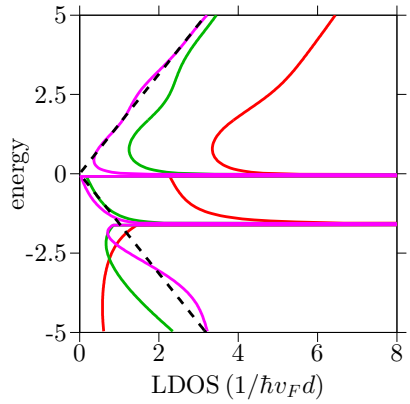

Figure 3.5: (Color online) The LDOS as a function of the dimensionless energy E for the $U = 5$ square-well model at the three fixed distances from the CNT: $x/d = 0$ (red), 0.6 (green), and 1.0 (violet). The dashed line shows the LDOS of unperturbed graphene.

3.4 Summary and future directions

In this Letter, we proposed a model for the anomalous plasmon reflection by ultranarrow electron boundaries in graphene. We validated this concept in experiments with electrostatically tunable line-like perturbations. One broad implication of our work is that nanoimaging of collective modes can reveal nontrivial electron properties, in this case, 1D bound states. Recent experiments have demonstrated that this technique is not limited to plasmons or graphene or 2D systems [DFM$^+$14, SHB$^+$15, GDG$^+$16, FSG$^+$]. We hope that our work stimulates even wider use of this novel spectroscopic tool.

A particularly intriguing future direction is to complement s-SNOM with scanned probe techniques other than AFM topography. For example, scanning tunneling microscopy, which has a superior spatial resolution, can be used to measure the local electron density of states (LDOS). For the particular model system studied here, the features exhibited by the LDOS should be quite striking, see Fig. 3.5 and Section 3.6. The origin of these features can be understood by examining the dispersions in Fig. 3.2(a). Within the selected energy interval there is the total of three bound states. The topmost one has a monotonic dispersion; the other two have energy minima at which the LDOS has van Hove singularities (diverges), see Fig. 3.5. The strength

of these singularities decreases exponentially with x because these bound states are localized near the well. At large x, the LDOS displays the V-shaped energy dependence characteristic of uniform graphene [BFL$^+$14]. We anticipate that the combination of optical and tunneling nanoimaging and nanospectroscopy could provide a refined information about the local electronic structure. One example of a possible application of this knowledge is the design of optimized plasmon switches (Fig. 3.1) for Dirac-material-based nanoplasmonics.

3.5 Local optical conductivity of a nonuniform graphene

In this section we describe the details of our 1D square well model, including the analytical expressions for the wavefunctions and the calculation of the optical conductivity in the vicinity the well. We start from the 2D Dirac Hamiltonian for the quasiparticles,

$$H = \hbar v_F (\sigma_z k_x + \sigma_y k_y) + v(x)\,, \tag{3.6}$$

where σ_i are the Pauli matrices acting on the sublattice pseudospin. The potential $v(x)$ is taken to be a square well,

$$v(x) = \begin{cases} -u, & |x| < d/2\,, \\ \\ 0, & |x| > d/2\,. \end{cases} \tag{3.7}$$

Without loss of generality, we take u to be positive. As the system is invariant in the y-direction, k_y is conserved, and so our problem is effectively one-dimensional. We use the following terminology: region I is the part of the system to the left of the well ($x < -d/2$); region II is the strip containing the well ($|x| < d/2$); region III is to the right of the well ($x > d/2$). For an eigenstate to have the same energy across all three regions, the magnitude k of its momentum

must obey the following relations

$$k_{\mathrm{I}} = k_{\mathrm{III}} = k_{\mathrm{II}} - \frac{u}{\hbar v_F}, \quad k_l^2 = k_{xl}^2 + k_y^2, \tag{3.8}$$

with $l = \mathrm{I}$, II, or III. When $k_l > 0$, the wavefunction belongs in the conduction band and has the form

$$\Psi_c(\mathbf{r}) \propto \psi_c(\theta_l) e^{ik_{xl}x + ik_y y}, \quad \psi_c = \begin{bmatrix} i\cos(\theta_l/2) \\ -\sin(\theta_l/2) \end{bmatrix}; \tag{3.9}$$

when $k_l < 0$, the wavefunction belongs to the valence band and is given by

$$\Psi_v(\mathbf{r}) \propto \psi_v(\theta_l) e^{ik_{xl}x + ik_y y}, \quad \psi_v = \begin{bmatrix} i\sin(\theta_l/2) \\ \cos(\theta_l/2) \end{bmatrix}. \tag{3.10}$$

The angle $\theta_l = \arctan(k_y/k_{xl})$ defines the direction of the momentum. In the following we use the notation k for k_{I} and $\bar{k}$ for k_{II} and similarly for all other quantities.

We find the complete expression for the wavefunction using the following device. We imagine that this wavefunction is generated by a plane wave incident from region I, which is then partially transmitted and reflected at each edge of the well. This yields

$$\Psi^r(\mathbf{r}) = \frac{e^{ik_y y}}{\sqrt{N^c}} \times \begin{cases} \psi_j(\theta) e^{ik_x x} + r_{jh}\, \psi_j(\pi - \theta)\, e^{-ik_x x}, & \mathrm{I} \\ \bar{t}_{jh}\, \psi_h(\bar{\theta}) e^{i\bar{k}_x x} + \bar{r}_{jh}\, \psi_h\left(\pi - \bar{\theta}\right) e^{-i\bar{k}_x x}, & \mathrm{II} \\ t_{jh}\, \psi_j(\theta) e^{ik_x x}, & \mathrm{III} \end{cases} \tag{3.11}$$

where $N^c = L_x L_y$ is normalization factor equal to the area of the system and j and h can be either c or v. The coefficients r, $\bar{t}$, $\bar{r}$ and t are determined by requiring the wavefunction to be continuous at the edges of the well, $x = \pm d/2$. If the wavefunction remains in the same band for all three

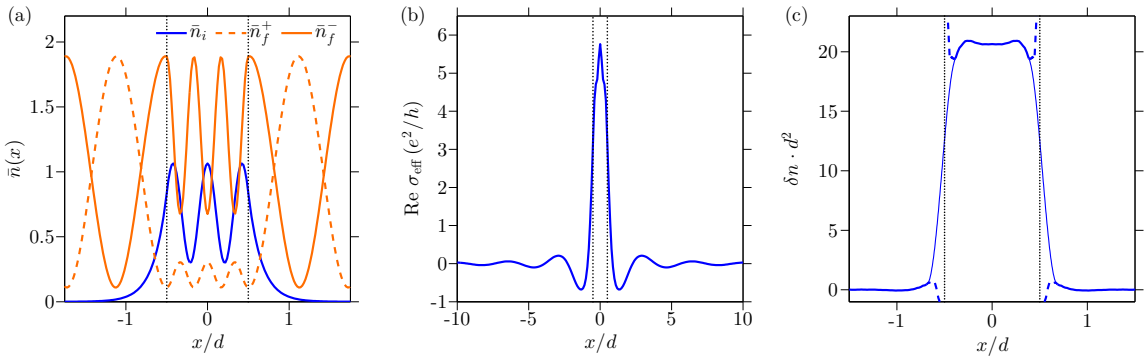

Figure 3.6: (a) The normalized density distribution $\bar{n}$ of the wave functions in a bound-to-continuum optical transition for $U \equiv ud/\hbar v_F = 5$, $K_y \equiv k_y d = 5$ and $\omega d/v_F = 1.7$. The bound state density (blue) is localized at the well, while the continuum state densities (orange) are extended. The bound state is on the $m = 3$ branch and has even parity, thus transition into the even parity final state (dashed curve) is forbidden. The vertical dotted lines indicate the boundaries of the square well. (b) The real part of the effective conductivity calculated with the same parameters as in (a) and $E_F = 4$. The peak at the center is caused by bound-to-continuum state transitions, while the long range oscillations with wavevector ω/v_F come from the transitions between states in the continuum. (c) Density perturbation induced by a square well of depth $U = 5$ at $E_F = 4$. The step-like discontinuity in the potential leads to a $1/x$ divergence at the edges of the well, shown by the dashed curves. For any realistic smooth potential the divergence is regularized, as shown schematically by the thin solid curves.

regions, the coefficients are

$$r_{cc} = r_{vv} = i\frac{2}{D}e^{i(\bar{\varphi}-\varphi)}\sin\bar{\varphi}\,(\sin\theta - \sin\bar{\theta})\,, \tag{3.12}$$

$$\bar{t}_{cc} = \bar{t}_{vv} = \frac{2}{D}e^{i\frac{\bar{\varphi}-\varphi}{2}}\cos\theta\cos\frac{\theta+\bar{\theta}}{2}\,, \tag{3.13}$$

$$\bar{r}_{cc} = \bar{r}_{vv} = \frac{2}{D}e^{i\frac{3\bar{\varphi}-\varphi}{2}}\cos\theta\sin\frac{\theta-\bar{\theta}}{2}\,, \tag{3.14}$$

$$t_{cc} = t_{vv} = \frac{2}{D}e^{i(\bar{\varphi}-\varphi)}\cos\theta\cos\bar{\theta}\,, \tag{3.15}$$

where $\varphi \equiv k_x d$ and $\bar{\varphi} \equiv \bar{k}_x d$, and

$$D(k,k_y) = 1 + \cos(\theta+\bar{\theta}) - e^{i2\bar{\varphi}}[1-\cos(\theta-\bar{\theta})]\,. \tag{3.16}$$

If the wavefunction switches band upon entering the well, the coefficients are

$$r_{vc} = r_{cv} = -r_{cc}(\theta \to \theta - \pi)\,, \tag{3.17}$$

$$\bar{t}_{vc} = -\bar{t}_{cv} = \bar{t}_{cc}(\theta \to \theta - \pi)\,, \tag{3.18}$$

$$\bar{r}_{vc} = -\bar{r}_{cv} = \bar{r}_{cc}(\theta \to \theta - \pi)\,, \tag{3.19}$$

$$t_{vc} = t_{cv} = t_{cc}(\theta \to \theta - \pi)\,. \tag{3.20}$$

If k_x is real, there is another wavefunction Ψ^l with the same magnitude k of the momentum (same energy), which corresponds to the wave incident from region III. A quick method to obtain Ψ^l is by reflecting Ψ^r with respect to the y-axis:

$$\Psi^l(x) = \sigma_y \Psi^r(-x)\,. \tag{3.21}$$

From Ψ^r and Ψ^l we construct the orthogonal eigenstates

$$\Psi^{\pm} = \frac{\Psi^r \pm \Psi^l}{\sqrt{2}}\,, \tag{3.22}$$

which are labeled by their parity P:

$$\Psi^P(x) = P\sigma_y \Psi^P(-x)\,, \quad P = \pm 1\,. \tag{3.23}$$

From the above expression we deduce that states localized within the potential well must also exist. Indeed, whenever $|\bar{k}| > |k|$ there exist states with $|k| < |k_y|$, so that k_x is imaginary and the wavefunction is evanescent outside the well. This happens when the denominators vanish, $D = 0$, so that the eigenstate exists without an incident plane wave from outside the well. The dispersion

of these bound states is found by solving

$$i(k\bar{k} - k_y^2)\tan\bar{\varphi} = k_x\bar{k}_x\,. \tag{3.24}$$

The wavefunction still has the form of Eq. (3.11), but with a normalization factor

$$N^b = 2L_y d\left[|\bar{t}|^2\left(1 + P\sin\bar{\theta}\,\frac{\sin\bar{\varphi}}{\bar{\varphi}}\right) + |t|^2\sin\theta\,\frac{e^{-qd}}{2qd}\right]. \tag{3.25}$$

Note that k_x is now imaginary, $k_x \equiv iq$ with $q > 0$. Each branch of solution except the one terminating at $k = k_y = 0$ is the continuation of Fabry–Pérot (FP) modes outside the continuum. The FP modes correspond directly to the supercritical or quasi-bound states. They satisfy the resonance condition $\bar{\varphi} = \pi m$ with $m = 1, 2, \ldots$, so branches of smaller m emerge at higher k_y. This condition can be expanded to find the expression for the critical point at which the mth bound state emerges from the valence band,

$$\frac{u}{\hbar v_F} = |k_y| + \sqrt{k_y^2 + (m\pi/d)^2}\,. \tag{3.26}$$

All the bound state branches asymptotically approach the line $k = |k_y| - u/\hbar v_F$ as $|k_y| \to \infty$. The wavefunction of each branch is alternatively even ($P = +1$) or odd ($P = -1$) with the lowest branch being even. An example of the normalized density of a bound state $\bar{n} \equiv |\Psi|^2 L_y d$ is shown in Fig. 3.6(a) along with the normalized density of continuum states $\bar{n} \equiv |\Psi|^2 L_x L_y$ for comparison.

Having found the expression for the eigenstate wavefunctions, we use the Kubo formula to calculate the nonlocal conductivity,

$$\begin{aligned}\sigma(\mathbf{r}, \mathbf{r}') = -\frac{1}{i\omega}\sum_{i,f}\frac{\nu_f - \nu_i}{\varepsilon_f - \varepsilon_i - (\omega + i0^+)}\times \\ \left\langle\Psi_i(\mathbf{r})\right|\hat{j}_x\left|\Psi_f(\mathbf{r})\right\rangle\left\langle\Psi_f(\mathbf{r}')\right|\hat{j}_x\left|\Psi_i(\mathbf{r}')\right\rangle,\end{aligned} \tag{3.27}$$

where i and f represent initial and final states, ν is the Fermi-Dirac occupation factor of the state with $\nu_i = 1$ and $\nu_f = 0$, $\varepsilon = \hbar v_F k$ is the energy of the state, and $\hat{j}_x = e v_F \sigma_z$ is the current operator. Assuming that the total field $\mathbf{E}$ is uniform and parallel to $\hat{x}$, $\mathbf{E} = E_x \hat{x}$, the current-field relation can be simplified to

$$j_x(x) = \int_{-\infty}^{\infty} d\mathbf{r}' \, E_x \sigma(\mathbf{r}, \mathbf{r}') \equiv E_x \sigma_{\text{eff}}(x) \,. \tag{3.28}$$

Note that the integral over y' enforces the conservation of k_y,

$$\int_{-\infty}^{\infty} dy' \, e^{ik_{yi}} e^{-ik_{yf}} = 2\pi\delta(k_{yi} - k_{yf}) \,, \tag{3.29}$$

so that all y and y' dependences cancel out in Eq. (3.28). We are interested in the real part of the effective conductivity. Transitions that contribute satisfy the relations

$$k_{yf} = k_{yi} \,, \quad k_f = k_i + \frac{\omega}{v_F} \,. \tag{3.30}$$

Additionally, the i/f states must have the opposite parity as the matrix element

$$M(x) \equiv \left\langle \Psi_i(x) \right| \sigma_z \left| \Psi_f(x) \right\rangle \tag{3.31}$$

is odd in x when the i/f states have the same parity. To proceed, we impose periodic boundary conditions and extend the system size to infinity, so that

$$\sum_{i,f} \to g \times \begin{cases} \dfrac{L_y}{2\pi} \displaystyle\int dk_y \,, & \text{(bound)} \\ \dfrac{L_x L_y}{(2\pi)^2} \displaystyle\int dk_x dk_y \,, & \text{(continuum)} \end{cases} \tag{3.32}$$

where k_y is taken to be positive and $g = 8$ is the total degeneracy, including spin, valley and

contribution from negative k_y. Applying the Sokhotski–Plemelj formula

$$\mathrm{Im}\,\frac{1}{x-i0^+} = \pi\delta(x) \tag{3.33}$$

to Eqs. (3.27) and (3.28), we find for the bound-to-bound state transitions,

$$\begin{aligned}\mathrm{Re}\,\sigma^{bb}_{\mathrm{eff}}(x) = g\pi\frac{e^2}{h}\frac{v_F}{\omega}L_y^2 \\ \times\left|\frac{dk_i}{dk_y}-\frac{dk_f}{dk_y}\right|^{-1}_{k_y^*} M(x)\int_{-\infty}^{\infty} dx'\,M^*(x')\,,\end{aligned} \tag{3.34}$$

where k_y^* satisfies Eq. (3.30). For bound-to-continuum transitions, we get

$$\begin{aligned}\mathrm{Re}\,\sigma^{bc}_{\mathrm{eff}}(x) = \frac{g}{2}\frac{e^2}{h}\frac{v_F}{\omega}L_xL_y^2 \\ \times\int_{k_y^{\min}}^{k_y^{\max}}\frac{dk_{yi}}{|\cos\theta_f|}M(x)\int_{-\infty}^{\infty} dx'\,M^*(x')\,.\end{aligned} \tag{3.35}$$

The limits of k_y are determined from the dispersion, the frequency ω and the doping level $k_F = \mu/(\hbar v_F)$. Continuum-to-bound state transitions result in the same expression except that the labels i and f are interchanged. The resultant conductivity has a peak around the well and decays quickly away from the well [Fig. 3.6(b)]. Finally, continuum-to-continuum transitions yield

$$\begin{aligned}\mathrm{Re}\,\sigma^{cc}_{\mathrm{eff}}(x) = \frac{g}{4\pi}\frac{e^2}{h}\frac{v_F}{\omega}L_x^2L_y^2 \\ \times\int_0^{k_F} dk_{xi}\int_{k_y^{\min}}^{k_y^{\max}}\frac{dk_{yi}}{|\cos\theta_f|}M(x)\int_{-\infty}^{\infty} dx'\,M^*(x')\,,\end{aligned} \tag{3.36}$$

where

$$k_y^{\min} = \sqrt{\max\left[0, \left(k_F - \frac{\omega}{v_F}\right)^2 - k_{xi}^2\right]}, \qquad (3.37)$$
$$k_y^{\max} = \sqrt{k_F^2 - k_{xi}^2}.$$

This results in an oscillating conductivity with a period of $2\pi v_F/\omega$ which can be negative, that is, the local current in real space can go in the opposite direction as the field. This is however no cause for alarm. Consider the case of uniform undoped graphene where the conductivity as a function of momentum at fixed ω is

$$\sigma_0(q) = -i\frac{e^2}{4\hbar}\frac{\omega}{\sqrt{v_F^2 q^2 - \omega^2}}. \qquad (3.38)$$

The corresponding conductivity in the real-space is

$$\sigma_0(x) = \int_{-\infty}^{\infty} \frac{dq_x}{2\pi}\,\sigma_0(q_x)e^{iq_x x} = \frac{e^2}{8\hbar}\frac{\omega}{v_F}H_0^{(1)}\left(\frac{\omega}{v_F}x\right), \qquad (3.39)$$

so that $\mathrm{Re}\,\sigma_0(x) \propto J_0\left(\frac{\omega}{v_F}x\right)$ can be negative. Thus conductivity oscillations with period $\sim 2\pi v_F/\omega$ is a general property in the presence of nonuniformity. For $\omega = 890\,\mathrm{cm}^{-1}$ in our s-SNOM experiment this period is 37 nm and such oscillations cannot be easily resolved. Therefore, we draw the reader's attention to another feature of the computed local conductivity, which is the prominent peak near the origin, see Fig. 3.6(b). We conclude that our simple model does predict a strong enhancement of dissipation near the nanotube, in agreement with the experiment. The extra dissipation is caused by the bound-to-bound state optical transitions. Of course, these calculations are not meant to be quantitatively compared with the experiment because our model of the square-well potential is not fully realistic. The quantity more suitable for the purposes of a

qualitative comparison is the average value of the effective conductivity

$$\bar{\sigma} \equiv \frac{1}{W} \int_{-\infty}^{\infty} dx \, \mathrm{Re}\, \sigma_{\mathrm{eff}}(x) \,. \tag{3.40}$$

similar to Eq. (4) of the main text. As long as W is larger than d but smaller than the spatial resolution, the near-field profile is sensitive only to the product $\bar{\sigma}W$ not the precise value of W, see Sec. 3.8. The results of our calculations of $\bar{\sigma}$ are shown in Fig. 3 of the main text. For simplicity, in these calculations we excluded the part of $\mathrm{Re}\,\sigma_{\mathrm{eff}}$ resulting from continuum-to-continuum transitions because it is relatively small at the potential well.

3.6 Local density and density of states

For better understanding of the effect of the potential well on electronic properties it is instructive to consider two other local observables: the carrier density and the local density of states. We begin with the density perturbation $\delta n(x)$. To find this quantity we first find the square of the absolute value of the wavefunctions $|\Psi(x,y)|^2$ of the occupied eigenstates under the potential well, integrated over k_y and k. The integration is done over the energies k bounded from above by the Fermi energy and from below by a cutoff energy k_m, a large negative number. The same procedure is then repeated for the unperturbed eigenstates (without the potential well). The difference of the two results is δn. There is however a complication to this procedure rooted in the "chiral anomaly" in the quantum field-theory of free Dirac fermions. For x inside the well, the lower cutoff k_m for the unperturbed eigenstates must be changed to $\bar{k}_m = k_m - u/\hbar v_F$. Without this redefinition, $\delta n(x)$ would diverge as k_m is decreased. Once this proper background subtraction is done, the integration converges to a finite value $\delta n(x)$ everywhere except at the edges of the well, $x \to \pm d/2$. The remaining divergence can be traced to the discontinuity in the potential $v(x)$. To arrive at this conclusion we reasoned that the divergence is produced by large

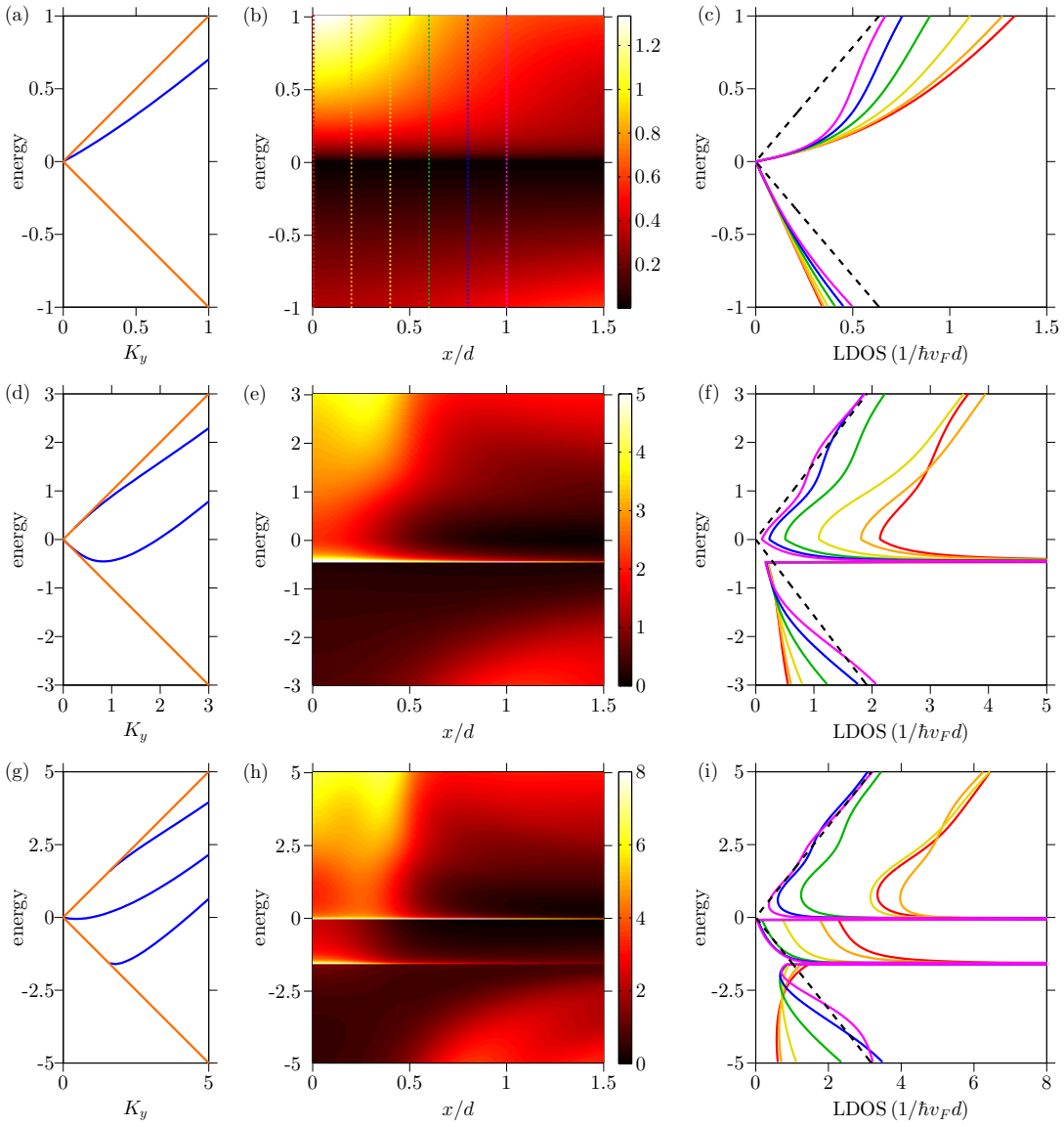

Figure 3.7: The local density of states ν in the vicinity of a square well potential. (a) The dispersion of the bound states for well depth $U = 1$. (b) A false color plot of the LDOS as a function of the dimensionless energy $E = kd$ and distance x/d. (c) Cross sections of (b) taken at several distances from $x = 0$ to $x = d$. The dashed black lines represent the LDOS of unperturbed graphene, $\nu_0 = 2|k|/\pi\hbar v_F$. Similar quantities are shown in (d)-(f) for $U = 3$ and in (g)-(i) for $U = 5$.

negative energies, and so it could be investigated using the perturbation theory. Therefore, we considered the linear-response theory expression for the density perturbation:

$$\delta n(x) = \int \frac{dq}{2\pi} e^{iqx} \Pi(q) v(q)\,, \tag{3.41}$$

where

$$v(q) = -\frac{2U_0}{q} \sin\left(\frac{qd}{2}\right) \tag{3.42}$$

is the Fourier transform of the potential $v(x)$ and $\Pi(q)$ is the static polarization function of graphene. At large q this function behaves as

$$\Pi(q) = -\frac{|q|}{4\hbar v_F} \tag{3.43}$$

regardless of the doping level. [WSSG06] Evaluation of the integral for δn using this asymptotic form yields

$$\frac{u}{4\pi\hbar v_F}\left(\frac{1}{x+\frac{d}{2}} - \frac{1}{x-\frac{d}{2}}\right), \tag{3.44}$$

which matches the numerical results calculated as described above for undoped graphene [Fig. 3.6(c)]. In reality, i) the linear dispersion of Dirac fermions does not extend to infinite momenta and ii) the potential must be smooth. Either way the divergence is regularized and the perturbed density is smooth as well, see Fig. 3.6(c). Not surprisingly, this box-like density profile is different from the more realistic Lorentzian function [Eq. (3.66)] we used to fit the experimental data in the main text and in Sec. 3.8 below. However, as we argued in Sec. 3.5, a qualitative comparison between the present model and the experiment is still meaningful.

Let us now turn to the local density of states (LDOS) ν. Previously, the LDOS of graphene around clusters of pointlike charged impurities has been measured by scanning tunneling spectroscopy. [WWS$^+$13] These experiments discovered peaks in LDOS, which were attributed to the emergence of the supercritical quasi-bound states. [SKL07a] We find that for a 1D perturbation

the bound states, instead of the quasi-bound ones, produce the dominant features in the LDOS.

The contribution of the bound states to the LDOS is given by

$$\nu(k,x) = \frac{g}{2\pi}\frac{L_y}{\hbar v_F}\sum_i \left|\frac{d\tilde{k}_{yi}}{dk}\right| \left|\Psi(\tilde{k}_{yi},x)\right|^2, \tag{3.45}$$

where $\tilde{k}_{yi}$ are positive solutions of Eq. (3.24) at a given k. The contribution of the delocalized states in the continuum is

$$\nu(k,x) = \frac{g}{4\pi^2}\frac{L_x L_y}{\hbar v_F}|k| \sum_{P=\pm}\int_0^{\pi/2} d\theta \left|\Psi^P(k,\theta,x)\right|^2. \tag{3.46}$$

We show in Fig. 3.7(a) the dispersion of bound states and in Fig. 3.7(b) the false color plot of the LDOS for the case of $U = 1$. The bound states contribute to the significant increase in ν at positive energies and for distances x inside the well. This is seen more clearly in Fig. 3.7(c), where vertical cross sections of the false color plot is taken at several distances inside and outside the well. The contribution of the bound states drops quickly outside the well and approaches the unperturbed LDOS $\nu_0 = 2|k|/(\pi\hbar v_F)$ shown in dashed lines. Similar plots for cases $U = 3$ and $U = 5$ are shown in Fig. 3.7(d)-(f) and in (g)-(i), respectively. In these two cases the LDOS inside the well are similarly increased due to the bound states. However, the most prominent features of the LDOS are the van Hove singularities that are caused by the extrema in the bound state dispersions. For $U = 3$ the singularity occurs at dimensionless energy $E = kd = 0.45$, while for $U = 5$ they occur at $E = -0.05$ and $E = -1.59$. A quasi-bound state is present for the case of $U = 5$, whose contribution is manifest as the increase of the LDOS inside the well just before the van Hove singularity at $E = -1.59$, as shown in Fig. 3.7(i). The former is a relatively weak feature in comparison to the latter.

We think that these properties of the LDOS should be quite generic for hypercritical potentials in graphene. Therefore, despite the oversimplification of the square-well potential model, our analysis may provide a useful reference for future scanning tunneling experiments

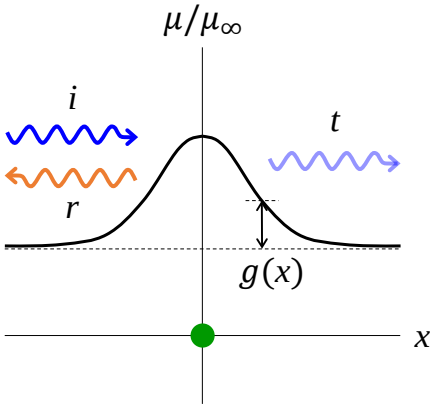

Figure 3.8: Schematic of an incident, reflected, and transmitted plasmon wave near the electronic inhomogeneity $g(x)$ caused by the linelike charge perturbation (green).

with such potentials.

3.7 Plasmon reflection from a linelike charge perturbation

In this section we summarize the theory of plasmon reflections from a linelike charge perturbation. [FRG+13] At this stage we are not yet discussing how the incident plasmon wave is created or how the reflected wave can be measured. Those questions are addressed in Sec. 3.8 devoted to realistic simulations of s-SNOM experiment. Here our purpose is to specify the model assumptions and to present the analytical results.

Our main assumption is that we can describe the response of graphene by a local conductivity $\sigma(x)$. This assumption is readily justified if the density $n(x)$ and the chemical potential $\mu(x)$ of graphene are smoothly varying, see Fig. 3.8. Thus, if the plasmon energy $\hbar\omega$ is much smaller than μ everywhere, $\sigma(x)$ is given by [Eq. (5) of the main text],

$$\sigma(x) = \frac{i}{\pi\omega}\frac{\mathcal{D}(x)}{1+i\gamma(x)}, \quad \mathcal{D}(x) = \frac{e^2}{\hbar^2}|\mu(x)|. \tag{3.47}$$

Here $\mathcal{D}$ is the Drude weight and the dimensionless function γ is the phenomenological damping rate. We assume that the system remains uniform in the y-direction at all x. The legitimacy of the

local conductivity approximation is less obvious if the carrier density varies sharply, e.g., in a box-like fashion depicted in Fig. 3.6(c). However, it should be indeed valid in the context of the plasmon propagation if the plasmon wavelength is longer than the characteristic length scale of nonlocality (the Fermi wavelength or the characteristic width of the inhomogeneity, whichever is larger). The effective local conductivity can then be defined by averaging the nonlocal one over a suitable interval W, see Eq. (3.40). In this case, Eq. (3.47) should be considered a formal parametrization of function $\sigma(x)$. In particular, $\gamma(x)$ should be understood as damping averaged over the lengthscale W.

Let us suppose that at $x \to \pm\infty$, $n(x)$ and $\mu(x)$ approach constant values n_∞ and μ_∞, respectively and let us define the dimensionless function

$$g(x) = \frac{\sigma(x)}{\sigma_\infty} - 1\,, \quad \sigma_\infty \equiv \sigma(\infty)\,. \tag{3.48}$$

If γ were constant, this function would be equal to $g = |\mu/\mu_\infty| - 1$, see Fig. 3.8.

We want to study how an incident plasmon plane wave with momentum (q_x, q_y) is scattered by the inhomogeneity. We will show that the corresponding reflection coefficient is given by the formula

$$r \simeq i\frac{q_x^2 - q_y^2}{q_x}\tilde{g}(-2q_x)\,, \tag{3.49}$$

where

$$\tilde{g}(k) = \int\limits_{-\infty}^{\infty} dx\, g(x) e^{-ikx}\,. \tag{3.50}$$

In particular, for normal incidence, $q_x = q_\infty$, $q_y = 0$, the reflection coefficient is

$$r \simeq iq_\infty \tilde{g}(-2q_\infty)\,, \tag{3.51}$$

similar to the usual first Born approximation. Note that because of the translational invariance in y, the momentum q_y is conserved.

Let us outline the derivation. Assuming $q \gg \omega/c$, which is satisfied in our experiment, we can neglect retardation and treat the problem in the quasistatic approximation. As our main dependent variable we choose the electric potential $\Phi = \Phi(\mathbf{r})$. In general, Φ is the sum of the external potential Φ_{ext} and the potential induced by charge density ρ in graphene,

$$\Phi(\mathbf{r}) = \Phi_{\text{ext}}(\mathbf{r}) + (V * \rho)(\mathbf{r}), \tag{3.52}$$

where $V(\mathbf{r}) = 1/\kappa r$ is the Coulomb kernel and the asterisk denotes convolution,

$$(A * B)(\mathbf{r}) \equiv \int d^2 r' A(\mathbf{r} - \mathbf{r}') B(\mathbf{r}'). \tag{3.53}$$

Combining together these equations plus the continuity equation for current and charge density, we obtain

$$\Phi(\mathbf{r}) = \Phi_{\text{ext}}(\mathbf{r}) - V(\mathbf{r}) * \nabla \left(\frac{\sigma(\mathbf{r})}{i\omega} \nabla \Phi(\mathbf{r}) \right). \tag{3.54}$$

For an ideal uniform sample the solution of this equation has the form of a Fourier integral:

$$\Phi(\mathbf{r}) = \int \frac{d^2 q}{(2\pi)^2} e^{i\mathbf{q}\mathbf{r}} \frac{\tilde{\Phi}_{\text{ext}}(\mathbf{q})}{\varepsilon(q)}, \quad \varepsilon(q) = 1 - \frac{q}{q_p}. \tag{3.55}$$

The zero of the dielectric function $\varepsilon(q)$ defines the plasmon momentum

$$q_p = \frac{i\kappa\omega}{2\pi\sigma} \tag{3.56}$$

introduced in the main text. The momentum q_p is complex for any finite damping, $\gamma > 0$, with $\operatorname{Im} q_p > 0$ having the physical meaning of the inverse propagation length. Indeed, in the absence of the external potential, one can find (unbounded) solutions $\Phi = e^{iq_x x + iq_y y}$ with real q_y and complex $q_x = \sqrt{q_p^2 - q_y^2}$, $\operatorname{Im} q_x > 0$, which can be thought of as decaying plane waves that are incident from the far left at some oblique angle. In the problem we study graphene is inhomogeneous, q_p

is x-dependent,

$$\frac{1}{q_p(x)} = \frac{1+g(x)}{q_\infty}, \quad q_\infty \equiv q_p(\infty), \tag{3.57}$$

and so the solution would contain the incident and the scattered (reflected plus transmitted) waves, see Fig. 3.8.

Setting $\Phi_{\rm ext}(\mathbf{r}) \to 0$ and $\Phi(\mathbf{r}) \to \Phi(x)e^{iq_y y}$ in Eq. (3.54), we obtain the equation for $\Phi(x)$:

$$\Phi(x) = V_1 * \left(\frac{1+g(x)}{q_\infty} q_y^2 \Phi(x) - \partial_x \frac{1+g(x)}{q_\infty} \partial_x \Phi(x) \right). \tag{3.58}$$

Here $V_1(x) = K_0(|q_y x|)/\pi$ is the 1D Coulomb kernel and $K_0(z)$ is the modified Bessel function of the second kind. Using the Green's function

$$G(x, q_y) = \int_{-\infty}^{\infty} \frac{dk}{2\pi} e^{ikx} \varepsilon^{-1}\left(\sqrt{k^2 + q_y^2}\right), \tag{3.59}$$

we find the equation for the scattered wave $\psi \equiv \Phi(x) - e^{iq_x x}$:

$$\psi(x) = q_\infty^{-1} (G * V_1) * \left(g(x) q_y^2 \Phi(x) - \partial_x g(x) \partial_x \Phi(x) \right). \tag{3.60}$$

We expect $\psi(x) \simeq re^{-iq_x x}$ at large negative x, which implies

$$r = -\frac{i}{q_x} \int_{-\infty}^{\infty} dx\, e^{iq_x x} \left\{ q_y^2 g(x) \Phi(x) - \partial_x \left[g(x) \partial_x \Phi(x) \right] \right\}. \tag{3.61}$$

To the first order in the small parameter $g(x)$ we can replace $\Phi(x)$ with $e^{iq_x x}$ in the integral, which leads to Eqs. (3.49). A particularly simple result is obtained if the plasmon wavelength

$$\lambda_\infty = 2\pi/q_\infty \tag{3.62}$$

is much larger than the characteristic width d of the inhomogeneity, in which case $\tilde{g}(-2q_x) \simeq \tilde{g}(0)$.

Using Eqs. (3.47) and (3.51), we find the reflection coefficient

$$r \simeq iq_\infty \int_{-\infty}^{\infty} dx \left[\frac{\sigma(x)}{\sigma_\infty} - 1 \right] \tag{3.63}$$

for the normal incidence. This simple equation gives a basic idea how r may depend on d and the local change in σ. A more detailed discussion on the reflection coefficient for any form of $g(x)$ can be found in Section 5.8.

3.8 Fitting the near-field profiles

As described in the main text, the near-field amplitude $\bar{s}(x)$ and phase $\phi(x)$ measured in our imaging experiments reveals the presence of interference fringes, i.e., spatial modulations near the nanotube. For example, $\sim 20\%$ variations of $\bar{s}(x)$ are seen in Fig. 4(e) of the main text. Assuming these relative modulations should be of the order of the plasmon reflection coefficient r, we can use Eq. (3.63) to estimate the perturbation of the conductivity caused by the nanotube. Using the representative value of $\lambda_\infty \sim 200\,\mathrm{nm}$ at frequency $\omega = 890\,\mathrm{cm}^{-1}$ at which the effective permittivity is equal to

$$\kappa(\omega) = \frac{\varepsilon_{\mathrm{vacuum}}(\omega) + \varepsilon_{\mathrm{SiO_2}}(\omega)}{2} = 2.2\,, \tag{3.64}$$

we find $\sigma_\infty \approx 5ie^2/h$ from Eq. (3.56). Hence, we can reproduce $|r| \sim 0.2$ if we assume, for example, that the reactive part of the conductivity $\mathrm{Im}\,\sigma$ is constant, while the dissipative part is enhanced to about $\mathrm{Re}\,\sigma \sim 3e^2/h$ over an interval of width $d = 10\,\mathrm{nm}$ near the origin. These numbers are generally consistent with the estimates in the main text.

To go beyond such rough estimates, additional modeling is required. For it to be more realistic, several important issues need to be accounted for. First, the plasmon waves launched and detected by the s-SNOM tip are not simple plane waves because the tip is positioned very close to the nanotube. Second, the intensity of such waves depends in a nontrivial way on the

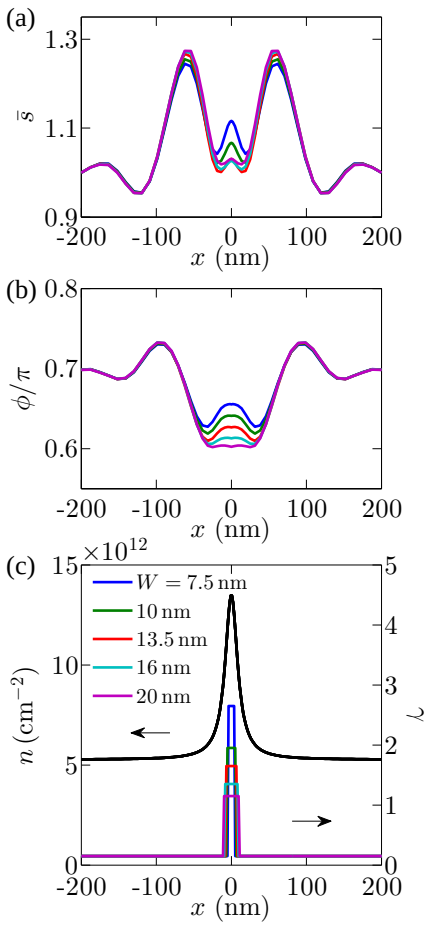

Figure 3.9: Simulated near-field amplitude ($\bar{s}$) and phase (ϕ) profiles for varying width W of the box-like discontinuity in the damping rate γ. The height of the box is chosen so that the resultant near-field profiles are as close as possible to the one used in the actual fitting (red). The detailed shape of the box affects only small-distance features in the signal.

electric field concentration that occurs inside the tip-sample nanogap. Third, in the experiment the tip-sample distance is varying periodically with the tapping frequency Ω. The complex near-field amplitude $s_3 e^{i\phi_3}$ corresponds to the signal demodulated at the third harmonic 3Ω. The normalized signal $\bar{s}(x)$ is the ratio $s_3(x)/s_3(L)$, where L is a coordinate point giving a fair approximation of the $x \to \infty$ limit. [$L = 200$ nm in Fig. 4(d)-(e) of the main text.] Because of these complications, the quantitative modeling of $\bar{s}(x)$ and $\phi(x)$ is possible only through numerical simulations.

Previously, an electromagnetic solver was developed, [FRA$^+$12, FRG$^+$13] which takes these issues into account. The algorithm implemented in the solver [FRG$^+$13] finds a numerical

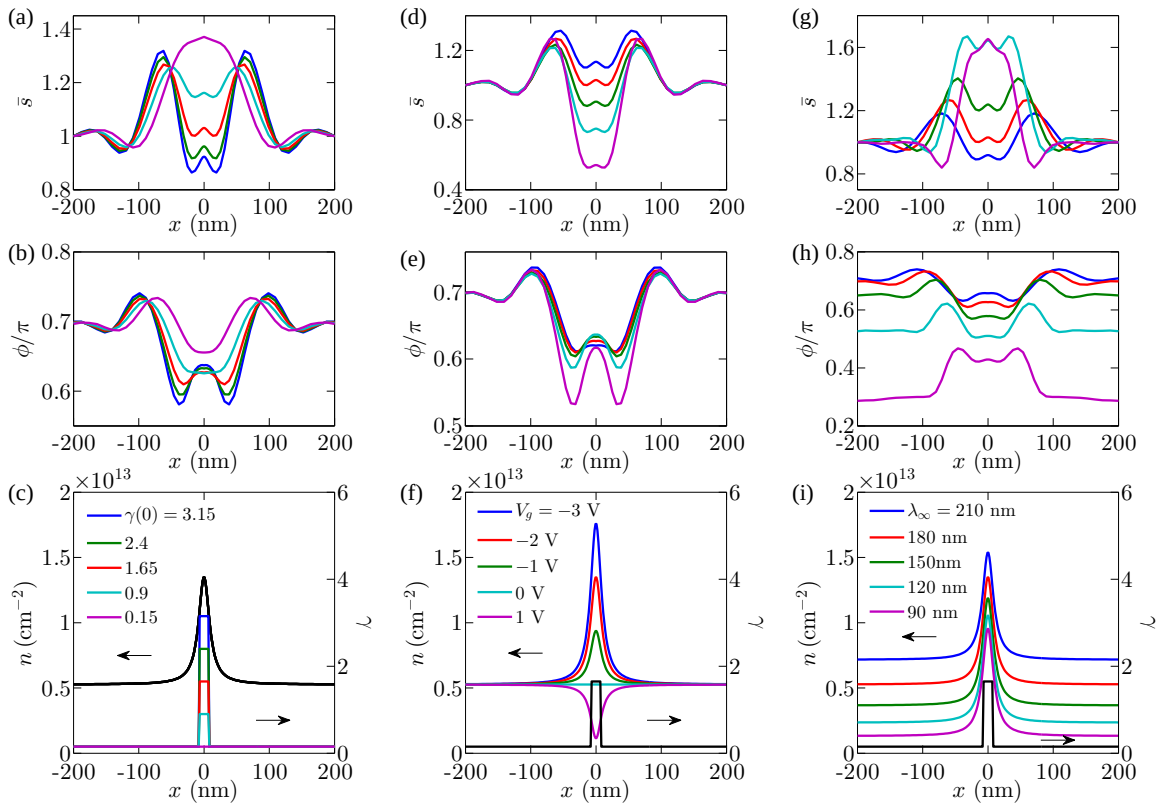

Figure 3.10: Simulated near-field amplitude ($\bar{s}$) and phase (ϕ) profiles along with the density (n) and damping profiles (γ) for (a)-(c) varying height $\gamma(0)$ of the box-like discontinuity in the damping rate, (d)-(f) gate voltage V_g, or (g)-(i) background plasmon wavelength λ_∞. The red and black curves correspond to the profiles used to produce the fits in Fig. 4(e) of the main text.

solution of Eq. (3.54) discretized on a double grid of q_y and x. The external field is taken to be the sum of two terms. The first one, a constant, represents the incident infrared beam. The second one approximates the field created by the tip, modeled as an elongated metallic spheroid. The charge density distribution on the spheroid surface is found self-consistently from the condition that this surface is an equipotential. The total dipole moment of the tip, which represents the instantaneous amplitude s of the scattered electromagnetic field is computed. Finally, $s_3 e^{i\phi_3}$ is calculated by taking the appropriate Fourier transform and normalized to the reference point $x = L$ in order to yield $\bar{s}$ and ϕ. The calculation is repeated for each tip position along the x-axis.

Using this solver we were able to produce simulated near-field profiles that matched well the measured ones using a set of adjustable parameters. We will now describe this fitting

procedure and the results, Figs. 3.9 and 3.10. As explained in Sec. 3.7, these fitting results should be considered an estimate of the nonlocal conductivity of graphene averaged over the lengthscale d.

We took the trial damping function to be

$$\gamma(x) = \gamma_\infty + [\gamma(0) - \gamma_\infty]\,\Theta\left(\frac{W}{2} - |x|\right), \tag{3.65}$$

where $\Theta(z)$ is the step-function, see the colored boxes in Fig. 3.9(c). For the carrier density profile we assumed the Lorentzian form [Fig. 3.9(c), black curve]

$$n(x) = n_\infty + \frac{C}{\pi}\frac{V_g}{e}\frac{d}{x^2+d^2}, \tag{3.66}$$

where $d = 10\,\text{nm}$ is the graphene-nanotube distance and C is the capacitance (per unit length) between them,

$$C = \frac{1}{2}\frac{\kappa_0}{\ln(2d/l)}. \tag{3.67}$$

The effective static permittivity κ_0 of the dielectric environment around the nanotube is

$$\kappa_0 = \frac{\varepsilon_{\text{hBN}}(0) + \varepsilon_{\text{SiO}_2}(0)}{2} = 3.7. \tag{3.68}$$

Equation (3.66) for $n(x)$ is appropriate for our relatively highly doped ($n > 10^{12}\,\text{cm}^{-2}$) graphene which screens the electric field of the nanotube like a good metal. [JF15] We have not measured the radius $l \sim 1\,\text{nm}$ of the nanotube directly, so there is an uncertainty in C. This uncertainty is however small due to the logarithmic form of C. On the other hand, the voltage difference V_g between the nanotube and graphene is measured. Hence, our model contains four adjustable parameters: $\gamma(0)$, γ_∞, W, and n_∞. Instead of the last of these we can use the asymptotic plasmon wavelength λ_∞ because they are directly related via Eqs. (3.47), (3.56), (3.62), and one more

equation,

$$\mu(n) = \hbar v_F (\pi |n|)^{1/2}. \tag{3.69}$$

A brief comment on the trial form of γ and n may be in order. The discontinuous box-like profile of the dimensionless damping rate $\gamma(x)$ may seem artificial; however, since the plasmon wavelength is much larger than the width of the box $W \sim d$, the near-field amplitude is largely insensitive to the precise functional form of $\gamma(x)$. In principle, we could also choose a box-like profile for $n(x)$. However, Eq. (3.66) is just as convenient and is better physically motivated.

In Fig. 3.9(a, b) we show the simulated profiles of the near-field amplitude $\bar{s}$ and phase ϕ for several W's for fixed $V_g = -2\,\mathrm{V}$, $\lambda_\infty = 180\,\mathrm{nm}$, and $\gamma(\infty) = 0.15$. The profiles for $W = 13.5\,\mathrm{nm}$ and $\gamma(0) = 1.65$ match those measured in the experiment rather well. This fitting suggests that the electrified nanotube causes the density change by almost a factor of three and the damping enhancement by more than an order of magnitude. The former conclusion should be robust as it is a consequence of the simple electrostatics. On the other hand, the latter should be considered the experimental discovery. An explanation for this surprisingly high local damping was presented in the main text and the technical details were given in Sec. 3.5.

A rough correspondence between the numbers obtained above and the parameters of the square-well model can be established as follows. The depth u of the well is taken as the integrated potential divided by the width d of the well, $u = \frac{1}{d}\int v(x)dx$. The potential $v(x)$ can be found through $v(x) = \mu(x) - \mu_\infty$ and Eq. (3.69). This results in the dimensionless well depth $U = ud/\hbar v_F = 13$. The estimation of the integrated conductivity $\bar{\sigma}$ is more complicated. Due to the presence of a nonzero background γ_∞, density variations will also contribute to the real part of the optical conductivity. To isolate the contribution of the optical transitions, the integrated conductivity is calculated as $\bar{\sigma} = \frac{1}{d}\mathrm{Re}\int[\sigma(x) - \sigma'(x)]dx$, where $\sigma'(x)$ is the conductivity of a comparison system, which has the same density profile $n(x)$ but the constant damping rate $\gamma = \gamma_\infty$. This prescription yields $\bar{\sigma} = 3.5\,\mathrm{e}^2/\mathrm{h}$.

The dependence of $\bar{s}$ and ϕ profiles on the other adjustable parameters, such as $\gamma(0)$, λ_∞,

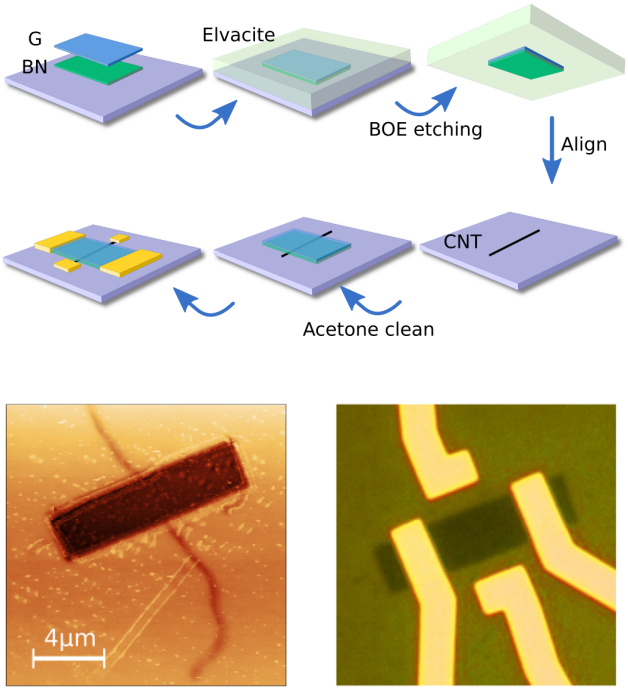

Figure 3.11: (Top) Fabrication process of the graphene/hBN/CNT heterostructure on a SiO_2/Si substrate. (Bottom left) AFM image of the device before addition of the contacts. (Bottom right) Optical image of the completed device.

and also on the gate voltage V_g is illustrated in Fig. 3.10(a, b), (g, h), and (d, e), respectively. The profiles vary dramatically with the changes in these parameters, and so the determination of the best-fitting values of the adjustable parameters has very little uncertainty. This analysis is yet another illustration of how the s-SNOM nanoimaging can be a powerful and sensitive technique for probing the local surface conductivity of graphene and perhaps many other 2D materials as well.

3.9 Device Fabrication

Our device consists of (from top to bottom) a graphene monolayer, a 10 nm-thick hBN flake, a metallic single-wall CNT, and a SiO_2/Si substrate. The CNT was grown by chemical vapor deposition, located using a scanning electron microscope, and transferred onto a SiO_2/Si substrate. Monolayer graphene was mechanically exfoliated. Using a PPC/PDMS stamp, the

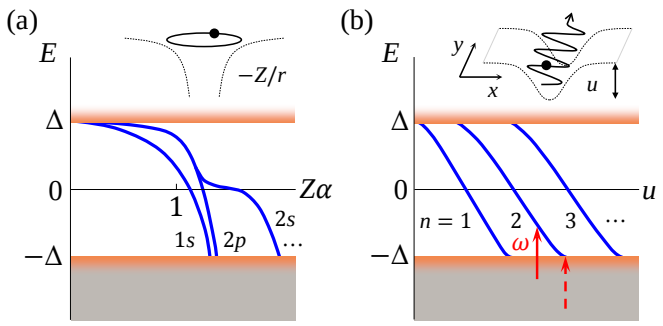

Figure 3.12: (Color online) Bound-states energies (blue) of Dirac fermions in strong external potentials. The orange lines represent the boundaries of continuum states. The insets depict classical trajectories (solid) and the potential profiles (dashed). (a) Low-lying states of large-Z atoms. [Adopted from Ref. [ZP72].] (b) Fixed-k_y states in a 1D potential well in graphene. The solid red arrow shows an optical transition of frequency ω from the valence band to a bound state. This transition disappears at some critical well depth u (dashed arrow).

graphene/hBN stack was transferred onto a separate SiO_2/Si substrate. The stack was then picked up with an acrylic resin Elvacite [HPT+15]. It was subject to buffered oxide etch (BOE), aligned, and transferred to cover the CNT. The Elvacite was cleaned away with acetone. The final step of the fabrication was adding metallic contacts to the CNT and graphene. These steps are summarized in Fig. 3.11 (top). The AFM and optical images of the device are shown in Fig. 3.11 (bottom left and right).

3.10 Supercritical transitions

Our problem has an intriguing parallel to the collapse of superheavy atoms in nuclear physics, which is as follows. For a very large nuclear charge $Z > Z_c \sim 1/\alpha$, the extrapolated values of the first few atomic levels sink below the top of the positron energy band, [ZP72] $-\Delta = -m_0c^2$, see Fig. 3.12(a). (Here $\alpha = e^2/\hbar c$ is the fine-structure constant and m_0 is the electron mass.) Such supercritical states can no longer be bound to the nucleus but should be quasi-bound, being hybridized with the extended states in the positron band. In graphene where the role of c is played by the Fermi velocity $v_F \sim c/300$, the critical charge is rather small, $Z_c \sim 1$ [SKL07b, PNCN07]. This has made it possible to observe the long-sought su-

percriticality experimentally by measuring the tunneling density of states near charged impurities [WWS$^+$13]. Analogous transitions [Ken02] are possible for the bound states studied in this Letter, Fig. 3.12(b). Compared to prior studies of a single [SKL07b, PNCN07, FNS07] or a few pointlike charges [WWS$^+$13, PNCN07, DMKME14, GGS15], the 1D geometry examined in this work has several advantages. The gapless 2D Dirac spectrum is replaced by a gapped one with $\Delta = |\hbar k_y v_F|$ [Fig. 3.12(b)], making the analogy to the atomic collapse problem [ZP72] closer. Alternatively, it prompts an analogy to a hypothetical cosmic string [NCV99], previously used in graphene literature [PCNLM10, CGS13] in a different context. More importantly, our "1D atom" permits continuous *in situ* tunability in terms of at least two parameters: the gate voltage and the optical excitation frequency. Experimental verification of these supercritical transitions can be attempted via two complementary approaches. One is to examine the changes in the LDOS. In fact, the detailed calculations of the LDOS presented in Sec. 3.6 were done exactly with such experiments in mind. The other approach is to look for abrupt drops in the local optical conductivity $\bar{\sigma}$ caused by the liquidation of the optical transitions, see Fig. 3(b) of the main text. Unfortunately, in either the conductivity or the LDOS, the supercritical signatures are very subtle compared to those of, say, van Hove singularities. Also, pinpointing these transitions requires an exhaustive search of the parameter space, which, for technical reasons, has not been possible in the devices we fabricated so far. Nevertheless, this can be an interesting and challenging problem for future work.

3.11 Electronic response of graphene to linelike charge perturbations

In this section, the problem of electrostatic screening of a charged line by undoped or weakly doped graphene is treated beyond the linear-response theory. The induced electron density is found to be approximately doping independent, $n(x) \sim x^{-2}\log^2 x$, at intermediate distances x

from the charged line. At larger x, twin p-n junctions may form if the external perturbation is repulsive for graphene charge carriers. The effect of such inhomogeneities on conductance and quantum capacitance of graphene is calculated. The results are relevant for transport properties of graphene grain boundaries and for local electrostatic control of graphene with ultrathin gates.

3.11.1 Introduction

One of the properties that make graphene an attractive platform for electronic devices is tunability of its charge carrier density through electrostatic gating. The gates can be brought into immediate proximity of graphene, which enables one to control doping of this material very efficiently. Local gating on ultrasmall lengthscales is attainable by utilizing nanowire [YK09] or nanotube [NJB$^+$15] gates [Fig. 3.13(a)]. Physical insight into fundamental characteristics of such devices can be gained from a simplified problem of how graphene responds to a linelike external charge [Fig. 3.13(c)]. This problem is also relevant for understanding properties of grain-boundary defects [Fig. 3.13(b)] in graphene grown by chemical-vapor deposition. [BL13, CDN$^+$14, YC14]

The problem of screening of a linelike charge by electrons in graphene is an interesting challenge because the usual linear-response theory [RSZF13] fails when the line is highly charged and/or when graphene is lightly doped. Previous studies of one-dimensional charge perturbations in graphene eschewed solving this difficult problem resorting instead to *ad hoc* approximations for the induced carrier density profile. [YK09, FRG$^+$13] Accurate determination of this profile requires numerical calculations, e.g., finding the self-consistent solution of the Dirac equation for electron wavefunctions and the Poisson equation for the electrostatic potential $\Phi(x)$ as a function of the in-plane coordinate x transverse to the charged line. However, if $\Phi(x)$ varies smoothly on the scale of the local Fermi length $k_F^{-1}(x)$, a simpler approach based on the Thomas-Fermi approximation (TFA) can be applied. [LL91] We show that approximate solutions of the TFA equations for the electron density can be derived in certain limits. Using these analytical solutions and numerical simulations, we compute two other important observables amenable to

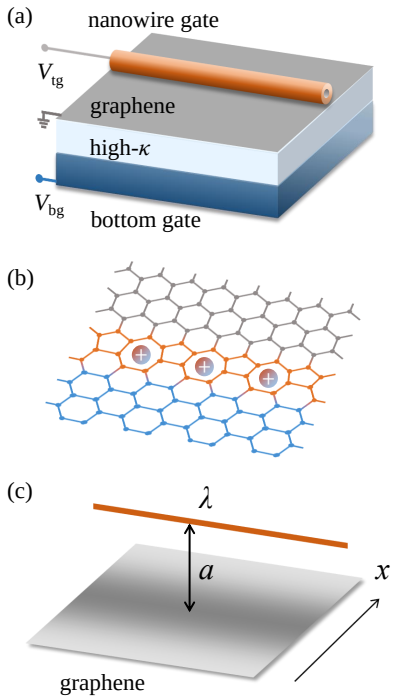

Figure 3.13: (Color online) Models of linear charged perturbations in graphene devices. (a) Nanowire top gate. The bottom gate separated from graphene by an insulator of high dielectric constant κ may be useful for additional control. (b) Grain boundary (pentagon-heptagon chain) with charged adsorbates (circles). (c) Charged string off the graphene plane.

experimental probes: the conductance and the gate capacitance.

Let us introduce our key notations and assumptions. We denote the linear charge density of the external perturbation by $e\lambda$. Without loss of generality, we take λ to be positive (unless specified otherwise). We assume that the unperturbed electron density n_∞ of graphene is uniform. To distinguish between n-type and p-type doping, we define the Fermi momentum corresponding to n_∞ as a signed quantity,

$$k_\infty = \mathrm{sgn}(n_\infty)|\pi n_\infty|^{1/2}. \tag{3.70}$$

We assume that the external charge $en_{\mathrm{ext}}(x)$ has a Lorentzian density distribution,

$$n_{\mathrm{ext}}(x) = \frac{\lambda}{\pi}\frac{a}{x^2+a^2}\,. \tag{3.71}$$

The actual profile of the external charge may of course be somewhat different. For example, the charge distribution of a grain boundary probably does not have power-law tails. However, the role of parameter a in Eq. (3.71) is mainly to regularize the response of graphene at very short distances, [FNS07] i.e., it serves as a short-distance cutoff. As long as we are not interested in microscopic physics at $|x| \lesssim a$, Eq. (3.71) can be adopted as a convenient model. In all examples mentioned above (nanowire and nanotube gates and also grain boundaries in graphene) realistically achievable a can be as small as a few nanometers. The particular functional form of Eq. (3.71) corresponds to the idealized model shown in Fig. 3.13(c) where the external charge is located off the graphene plane and is truly one-dimensional. This can be seen from the fact that the electrostatic potential created in the graphene plane by the out-of-plane charged line is equal to that created by the *in-plane* Lorentzian charge distribution (3.71):

$$\Phi_{\text{ext}}(x) = \frac{e}{\kappa}\int_{-\infty}^{\infty} dx' n_{\text{ext}}(x') \log\frac{1}{(x-x')^2} \tag{3.72}$$

$$= \frac{e\lambda}{\kappa}\log\frac{1}{x^2+a^2}\,. \tag{3.73}$$

In general, the effective width parameter a should be taken as the larger of the actual physical width of the charge distribution and its distance to the graphene plane. In this article we assume that electron-electron interaction in graphene is weak, i.e., we consider the dimensionless coupling constant $\alpha = e^2/\kappa\hbar v$ a small parameter in the problem. [KUP+12] Here $v = 10^8\,\text{cm}\,\text{s}^{-1}$ is the graphene Fermi velocity. By choosing a substrate with a large dielectric constant κ [Fig. 3.13(a)], it is possible to make $\alpha \simeq 2.2/\kappa$ significantly smaller than unity. However, it is difficult to make α truly small, so as a rule we do not treat $\log\alpha$ as a small parameter. Finally, we assume that the graphene is not too highly doped,

$$|k_\infty| \ll \frac{1}{\alpha a}\,. \tag{3.74}$$

Depending on the relation between λ, k_∞, and x, the response of graphene can be either weak or strong and either linear or nonlinear (Fig. 3.14). The degree of nonlinearity is controlled

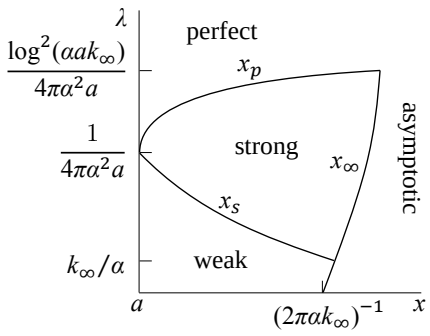

Figure 3.14: Schematic diagram of screening regimes. In the 'weak' region screening of the external potential is poor; the response is linear [Eq. (3.84)] at $\lambda < k_\infty/\alpha$ and nonlinear [Eq. (3.93)] at larger λ. In the 'strong' regime the external potential is greatly reduced and the induced density profile is given by Eq. (3.93). In the 'asymptotic' regime the density profile follows Eq. (3.77). In the 'perfect' regime graphene maintains local charge neutrality, apart from small corrections [Eq. (3.99)]. This diagram is drawn assuming graphene is not too heavily doped, Eq. (3.74); otherwise, the 'weak' and 'strong' screening regions would disappear.

by the dimensionless parameter

$$\tilde{\lambda} = \frac{\alpha\lambda}{k_\infty}\,. \tag{3.75}$$

Linear screening is realized if $\tilde{\lambda} \ll 1$ (the bottom part of the 'weak' region in Fig. 3.14) where the induced electron density

$$n_{\rm ind}(x) \equiv n(x) - n_\infty \tag{3.76}$$

scales linearly with λ. On the other hand, if $\tilde{\lambda} > 1$, the response is nonlinear. A conspicuous manifestation of the nonlinearity is found in the region labeled 'strong' in Fig. 3.14, where the induced density can be approximated by a 'universal' (doping-independent) form $n_{\rm ind}(x) \sim x^{-2}\log^2 x$. In the large-$x$ 'asymptotic' regime of Fig. 3.14, the induced density exhibit a power-law decay

$$n_{\rm ind}(x) \simeq \frac{b}{x^2}\,, \tag{3.77}$$

where the dependence of b on λ is linear [Eq. (3.80) of Sec. 3.11.2] if $\tilde{\lambda}$ is small and logarithmic if $\tilde{\lambda}$ is large [Eq. (3.96)]. The 'perfect' screening region is included in Fig. 3.14 for completeness. Here graphene maintains charge neutrality locally, i.e., the induced density is close to the external

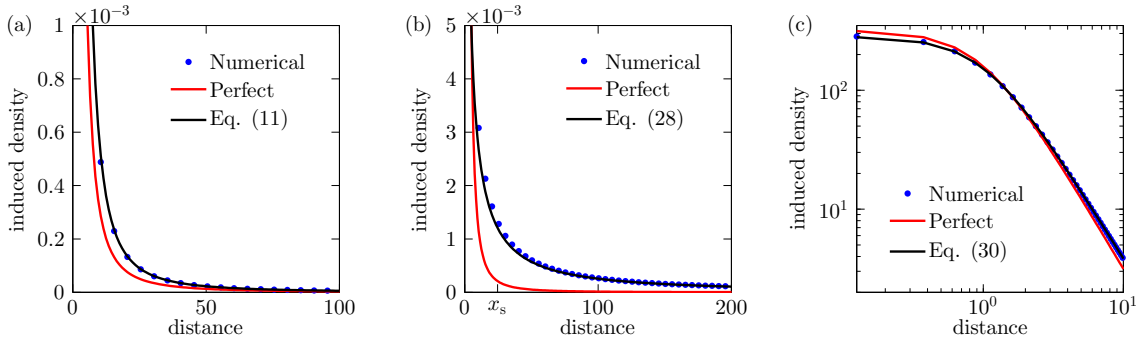

Figure 3.15: (Color online) Comparison of induced electron density profile from the numerical solutions of the TFA equations (blue dots) and analytical formulas (black lines) in three screening regimes. The red curves (labeled 'Perfect') in each panel represent $n_{\rm ext}(x)$. Units for density n, distance x, and linear number density λ are respectively $(4\pi^3\alpha^2a^2)^{-1}$, a, and $(4\pi^3\alpha^2a)^{-1}$. (a) Linear screening regime realized for $\lambda = 0.1$ and $n_\infty = 4$. The induced density is much smaller than n_∞. (b) Nonlinear regime realized for $\lambda = 0.4$ and $n_\infty = 0$. Density profile characterized by screening length $x_s \approx 25$. (c) Near-perfect screening for $\lambda = 1000$ and $n_\infty = 0$. Note the double logarithmic scale.

one, $n_{\rm ind}(x) \simeq n_{\rm ext}(x)$, so that the model assumption (3.71) must be critically revisited.

The crossovers among the predicted screening regimes can be systematically studied in experiments using devices that have both top and bottom gates, Fig. 3.13(a). High linear charge densities with $\alpha\lambda \sim (1\,{\rm nm})^{-1}$ are quite feasible to achieve with top gate voltages $V \sim 1\,{\rm V}$. On the other hand, applicable graphene carrier density is bounded from below by the formation of electron-hole puddles [DSAHR11, BFL$^+$14] and from above by Eq. (3.74). For high quality graphene, this allows the range of n from $\sim 10^9$ to $\sim 10^{12}\,{\rm cm}^{-2}$. The corresponding $\tilde\lambda$ can exceed one hundred, which is deep in the nonlinear screening regime.

In the remainder of this article we derive detailed formulas for the carrier density profiles, verify them by numerical simulations, then make predictions for capacitance and transport measurements.

3.11.2 Linear response

Linear screening of linelike charges by doped graphene has been studied in previous literature. [RSZF13] We include a brief summary of the relevant results for later comparison with our nonlinear response theory. Linear screening is realized when the external charge is not too high or when graphene doping level is not too low. The quantitative criterion $\tilde{\lambda} \ll 1$ is derived below. Within the linear-response theory, the induced electron density is given by

$$n_{\mathrm{ind}}(x) = \int_{-\infty}^{\infty} \frac{dq}{2\pi} \left[1 - \frac{1}{\varepsilon(q)}\right] \lambda e^{-a|q|+iqx}, \tag{3.78}$$

where the term $\lambda e^{-a|q|}$ is the Fourier transform of the effective external charge $n_{\mathrm{ext}}(x)$ [Eq. (3.71)] at momentum q. As we are primarily concerned with distances $x \gg k_\infty^{-1}$ where the TFA is valid, the dielectric function of graphene can be approximated by [KUP+12, AZS02, WSSG06]

$$\varepsilon(q) = 1 + \frac{q_{\mathrm{TF}}}{|q|}, \quad q_{\mathrm{TF}} = 4\alpha|k_\infty|. \tag{3.79}$$

A particularly simple analytical expression for $n_{\mathrm{ind}}(x)$ can be derived in the 'asymptotic' regime, cf. Fig. 3.14. Carrying out the integration in Eq. (3.78) by the steepest-descent method, one finds the leading-order approximation for the induced density to be [RSZF13]

$$n_{\mathrm{ind}}(x) \simeq \frac{b}{x^2}, \quad b = \frac{\lambda}{\pi}\left(a + q_{\mathrm{TF}}^{-1}\right). \tag{3.80}$$

(In contrast, $n_{\mathrm{ind}} \propto x^{-3}$ for a pointlike charge perturbation. [WSSG06]) Note that the coefficient b in Eq. (3.80) is much larger than $\lambda a/\pi$ under the assumed condition (3.74), so that $n_{\mathrm{ind}}(x) \gg n_{\mathrm{ext}}(x)$. Such an 'overshoot' is typical for screening of localized perturbations in metals. Metallic systems possess the overall charge neutrality. However, at distances shorter than the local screening length from the perturbation screening is necessarily weak. Therefore, there is a

missing charge at short distances, which has to be compensated at large x. However, the electric field is *not* overscreened: it is of the same sign as the external one but reduced in magnitude.

The analytical Eq. (3.80) agrees well with our numerical simulations shown in Fig. 3.15(a). For these simulations we used previously developed codes [ZF08] with suitable modifications. In brief, the electron density $n(x)$ to be found was defined on a grid of x with periodic boundary conditions. The solution was obtained by minimizing the total energy of the system (kinetic plus electrostatic) within the TFA using standard technical computing software. [MAT]

Refinement of Eq. (3.79) can be obtained through the random-phase approximation (RPA). Within the RPA, the dielectric function of graphene coincides with Eq. (3.79) at $0 < |q| < 2\,|k_\infty|$ but at $|q| > 2\,|k_\infty|$ it is given by a different formula [AZS02, WSSG06]

$$\begin{aligned}\varepsilon(q) = 1 + \frac{q_{\rm TF}}{|q|} - \frac{q_{\rm TF}}{2|q|}\sqrt{1-\left|\frac{2k_\infty}{q}\right|^2} \\ + \alpha\cos^{-1}\left|\frac{2k_\infty}{q}\right|, \quad |q| > 2|k_\infty|\,.\end{aligned} \tag{3.81}$$

Notably, $\varepsilon \simeq 1 + \pi\alpha/2$ becomes doping independent at $|q| \gg 2\,|k_\infty|$ where the response is dominated by virtual interband transitions. (For corrections to the last result beyond RPA, see Refs. [KUP+12, SF12].)

Substituting Eq. (3.81) in Eq. (3.78) and using contour integration techniques to evaluate the integral, we find the RPA correction to $n_{\rm ind}$:

$$\Delta n_{\rm ind}^{\rm RPA}(x) \simeq b_1\,\frac{\cos\left(2\,|k_\infty x| + \frac{\pi}{4}\right)}{|k_\infty x|^{5/2}}\,, \quad |k_\infty x| \gg 1\,, \tag{3.82}$$

$$b_1 = -\frac{\lambda}{2\sqrt{\pi}}\,\frac{\alpha|k_\infty|}{(1+2\alpha)^2}\,, \tag{3.83}$$

which is a particular case of the Friedel oscillations. [AM76] At intermediate distances, Eq. (3.78) yields [RSZF13]

$$n_{\rm ind}(x) \simeq \frac{\lambda q_{\rm TF}}{\pi}\log\frac{0.561}{|q_{\rm TF}x|}\,, \quad 1 \ll |k_\infty x| \ll \alpha^{-1}\,. \tag{3.84}$$

Finally, let us estimate the region of validity of the linear-response theory. This theory applies if the induced carrier density is smaller than the original one, $n_{\rm ind} \ll n_\infty$ or, equivalently, if the local Fermi momentum,

$$k_F(x) = \mathrm{sgn}\big(n(x)\big)\,|\pi n(x)|^{1/2} \tag{3.85}$$

is perturbed slightly compared to k_∞ [Eq. (3.70)]. Naively, one may require the condition $[k_F(x) - k_\infty]/k_\infty \ll 1$ to hold at all distances of interest, $x \gg a$. In fact, the validity region is wider because the nonlinearity affects only the response at momenta $q < 2\,|k_\infty|$. (As mentioned above, the response at $q > 2\,|k_\infty|$ is essentially doping-independent.) Therefore, the smallest number we should use in the argument of $n_{\rm ind}(x)$ for our estimate is $x \sim |k_\infty|^{-1}$. From Eqs. (3.76), (3.84), and (3.85) we get

$$\max\left[\frac{k_F(x)}{k_\infty} - 1\right] \simeq \frac{2\alpha\lambda}{k_\infty}\log\left(\frac{1}{\alpha}\right). \tag{3.86}$$

Neglecting the logarithmic factor, which is never large in practice, we arrive at $|\tilde{\lambda}| \ll 1$ as the criterion of linear screening.

3.11.3 Nonlinear response

In this Section we treat a more difficult case $\tilde{\lambda} \gg 1$ where screening is nonlinear. Our approach to this problem is to solve the TFA equations analytically and numerically. The first of these equations is

$$\mu(x) - e\Phi(x) = 0\,, \tag{3.87}$$

where

$$\mu(x) = \hbar v k_F(x) \tag{3.88}$$

is the local chemical potential of graphene (assuming the linear Dirac dispersion of quasiparticles). The second equation links the total charge density and the electrostatic potential,

$$\Phi(x) = \frac{2e}{\kappa} \int dx' \log|x - x'| [n_{\rm ind}(x') - n_{\rm ext}(x')] \,. \tag{3.89}$$

This relation can be inverted by exploiting techniques from the theory of analytic functions:

$$n_{\rm ind}(x) - n_{\rm ext}(x) = \frac{\kappa}{\pi^2 e} \mathcal{P} \int_0^\infty \frac{x' dx'}{x'^2 - x^2} \frac{d\Phi}{dx'} \,, \tag{3.90}$$

where $\mathcal{P}$ stands for the Cauchy principal value. If desired, Eqs. (3.87)–(3.90) can be combined into a single nonlinear integral equation for $n(x)$. The TFA is valid if [LL91]

$$\frac{d}{dx} k_F^{-1}(x) \ll 1 \,. \tag{3.91}$$

The problem we want to solve can be separated into two parts, depending on whether $n_{\rm ind}(x)$ is greater or smaller than n_∞. The latter situation occurs at large x, where we expect screening behavior akin to linear response. Indeed, it is easily seen that $n_{\rm ind}(x)$ follows Eq. (3.77) provided the integral in Eq. (3.90) is dominated by $x' \ll x$. Invoking Eq. (3.87), we then obtain the asymptotic behavior $\Phi \sim x^{-2}$ for $x \gg x_\infty$, where

$$x_\infty = |k_\infty|^{-1} |\pi b|^{1/2} \tag{3.92}$$

and b is to be determined below.

The analytical form of $n_{\rm ind}(x)$ at $x \ll x_\infty$ is not immediately obvious. Fortunately, were are able to find (by trial and error) the following approximate solution:

$$n_{\rm ind}(x) \simeq \frac{1}{4\pi^3 \alpha^2} \frac{1}{(x - x_{\rm s})^2} \log^2 \left(\frac{x}{x_{\rm s}} \right) , \tag{3.93}$$

which is characterized by a nonlinear screening length x_s. This length is found from the argument that at small x the external field is nearly unscreened, so that the total and the external potentials differ only by some constant: $\Phi(x) \simeq \Phi_{\rm ext}(x) + {\rm const} = -2\lambda(e^2/\kappa)\log x + {\rm const}$ [Eq. (3.73)]. Comparing with Eq. (3.93) at $x \ll x_s$, we get

$$x_{\rm s} = \frac{1}{4\pi\alpha^2\lambda}\,. \tag{3.94}$$

From Eqs. (3.91) and (3.94) we see that Eq. (3.93) is valid at $x \gg 2\pi\alpha x_s$. At smaller x the density and the Fermi momentum presumably tend to a finite maximum, i.e,

$$\max k_F(x) \simeq k_F(2\pi\alpha x_s) \simeq 2\alpha\lambda\log\left(\frac{1}{\alpha}\right)\,. \tag{3.95}$$

The last logarithmic factor is valid if $\alpha \ll 1$; otherwise, it should be replaced by a number of the order of unity. Equation (3.95) is consistent with Eq. (3.86) for the linear regime. Hence, in both regimes the nonlinearity parameter $\tilde{\lambda}$ [Eq. (3.75)] has the physical meaning of the maximum relative change in the Fermi momentum $(k_F - k_\infty)/k_\infty$ caused by the perturbation.

To fix the so far undetermined coefficient b, we require a smooth matching of Eq. (3.77) and Eq. (3.93), which yields

$$b \simeq \frac{1}{4\pi^3\alpha^2}\log^2|\tilde{\lambda}|\,, \quad x_\infty \simeq \frac{1}{2\pi\alpha k_\infty}\log|\tilde{\lambda}|\,. \tag{3.96}$$

More accurate expression for $n_{\rm ind}(x)$ can be obtained by iterations using Eq. (3.93) as the input to the TFA equations. Namely, substituting it into Eq. (3.87), we can get $\Phi(x)$, which we can then insert into Eq. (3.90) to obtain an improved approximation for $n_{\rm ind}(x)$. The first iteration

yields

$$n_{\rm ind}^{(1)}(x) = \frac{1}{4\pi^3\alpha^2}\left[\frac{\pi^2}{2}\frac{1}{(x+x_{\rm s})^2} + \log^2\left(\frac{x}{x_s}\right)\frac{x^2+x_s^2}{(x^2-x_s^2)^2} - 2\frac{\log(x/x_{\rm s})}{x^2-x_{\rm s}^2}\right], \tag{3.97}$$

which demonstrates a good agreement with our numerical simulations, see Fig. 3.15(b).

For completeness, we consider the case of

$$\lambda > \frac{1}{4\pi\alpha^2 a}, \tag{3.98}$$

where nonlinear screening affects distances shorter than our cutoff length a. At such large λ, a highly doped region appears at $x < a$, where screening is nearly perfect, i.e., where the difference between $n_{\rm ind}(x)$ and $n_{\rm ext}(x)$ is relatively small and the system is able to maintain the local charge neutrality. The solution for $n_{\rm ind}(x)$ can again be obtained by iterations. The input to Eq. (3.90) is now $n_{\rm ind}(x) = n_{\rm ext}(x)$. The first correction is

$$n_{\rm ind}(x) - n_{\rm ext}(x) \simeq \frac{\sqrt{\lambda a}}{\pi^2\alpha}\left[\frac{x\sinh^{-1}\frac{x}{a}}{(x^2+a^2)^{\frac{3}{2}}} - \frac{1}{x^2+a^2}\right]. \tag{3.99}$$

Its range of validity $x < x_p$ can be estimated from the requirement that at $x = x_p$ this correction is no longer small, which implies

$$\log(x_p/a) \sim \pi\alpha\sqrt{\lambda a}. \tag{3.100}$$

Furthermore, one can check that at $x = x_p$ Eq. (3.99) matches by the order of magnitude with $n_{\rm ind}(x)$ given by Eq. (3.93) with x_s set to a. This leads to us conclude that at $x > x_p$ Eq. (3.93) and then at $x > x_\infty$ Eq. (3.77) must still hold with $x_s \sim a$. Because of the exponentially large magnitude of x_p, we could verify numerically only the $x < x_p$ case. Comparison of simulations with Eq. (3.99) in Fig. 3.15(c) indeed shows a good agreement.

If $x_p > \sqrt{\lambda a}/k_\infty$, which corresponds to

$$\lambda > \frac{1}{\pi^2 \alpha^2 a} \log^2 \left(\frac{1}{\alpha a k_\infty} \right), \tag{3.101}$$

the screening should be nearly 'perfect' for all $x > a$, cf. Fig. 3.14.

When λ and n_∞ have opposite signs, twin *p-n* junctions appear at $|x| \sim x_\infty$. The density profile near a graphene *p-n* junction has been studied in Ref. [ZF08]. These results should more or less carry over to the present problem, so they will not be repeated. In the following Sections we focus on computing the effect of nonlinear screening on transport and capacitance characteristics of the system.

3.11.4 Quantum Capacitance

The charge density λ is a natural parameter for graphene grain boundaries [Fig. 3.13(b)]. In contrast, linelike perturbation created by means of narrow gates [Fig. 3.13(a)] are controlled by the voltage V with respect to graphene while λ has to be found by integrating the differential gate capacitance

$$C(V) = e \frac{d\lambda}{dV}. \tag{3.102}$$

It is important that V is not simply equal to the electrostatic potential difference $\Delta\Phi$ between the gate and the graphene sample. It has another contribution from the graphene chemical potential:

$$V = \Delta\Phi + (\mu/e). \tag{3.103}$$

As a result, the differential gate capacitance C has two components, the classical (or geometrical) one C_g and the so-called quantum one C_q, which add in series:

$$C^{-1} = C_g^{-1} + C_q^{-1}. \tag{3.104}$$

Our goal in this Section is to derive the quantum capacitance of a device with an ultranarrow top gate, Fig. 3.13(a). However, it is useful to review the conventional planar-gate geometry first. Here the classical capacitance per unit area $C_g = \kappa/(4\pi d)$ is inversely proportional to the separation d between graphene and the planar gate. In turn, the quantum capacitance is proportional to the thermodynamical density of states, $C_q = e^2(dn/d\mu)$. This quantity can also be written in terms of the inverse Thomas-Fermi screening length $q_{\rm TF}$:

$$C_q = \frac{\kappa}{2\pi} q_{\rm TF} \,. \tag{3.105}$$

The net result of having a finite screening length is equivalent to replacing the physical gate-graphene separation by an effective one:

$$d_{\rm eff} = d + \frac{1}{2} q_{\rm TF}^{-1} \,. \tag{3.106}$$

Although in this article we use the free-fermion approximation [Eq. (3.79)] for the linear-response screening length $q_{\rm TF}^{-1}$, in reality it is modified by many-body interactions and disorder (see, e.g., Refs. [Fog04, AGS$^+$06]). Recently quantum capacitance measurements have been used to probe such effects of graphene. [YJB$^+$13, BFL$^+$14]

If the gate is now a long metallic string or radius $l \ll a$, the capacitances per unit length are relevant. In order to derive C^{-1} we start with a general expression for the electrostatic potential difference between the string and the graphene sample,

$$\begin{aligned} \Delta\Phi(x) = {} & \frac{e\lambda}{\kappa} \log\left(\frac{x^2+a^2}{l^2}\right) \\ & + \frac{e}{\kappa}\int_{-\infty}^{\infty} dx' n_{\rm ind}(x') \log\frac{x'^2+a^2}{(x-x')^2} \,, \end{aligned} \tag{3.107}$$

which follows from Eqs. (3.72) and (3.89). If we set $n_{\rm ind} = n_{\rm ext}$, we obtain

$$\Delta\Phi = (e\lambda/\kappa)\log(4a^2/l^2) = \text{const}\,. \tag{3.108}$$

Hence, the geometric capacitance of the system is

$$C_g^{-1} = \frac{\Delta\Phi}{e\lambda} = \frac{2}{\kappa}\log\left(\frac{2a}{l}\right), \tag{3.109}$$

which can be alternatively derived by the method of images. Next, combining Eqs. (3.87), (3.89), (3.102), (3.103), (3.107), and subtracting C_g^{-1}, we find

$$C_q^{-1} = \frac{1}{e\lambda}\int_{-\infty}^{\infty} dx\, n_{\rm ext}(x)\frac{\partial}{\partial\lambda}\Phi(x)\,. \tag{3.110}$$

We can now use this expression for analytical and numerical calculations. Our analytical formulas for positive $\tilde{\lambda} \equiv \lambda/(\alpha k_\infty)$ are as follows:

$$\frac{\kappa}{2}C_q^{-1} \simeq \begin{cases} -e^{2aq_{\rm TF}}\mathrm{Ei}(-2aq_{\rm TF})\,, & \tilde{\lambda} \ll 1, \quad \text{(3.111a)} \\ -\log(4\pi\alpha^2 a\lambda)\,, & \tilde{\lambda} \gg 1, \quad \text{(3.111b)} \end{cases}$$

where $\mathrm{Ei}(z)$ is the exponential integral. These equations describe, respectively, the linear and the 'strong' regimes of Fig. 3.14. They can be commonly written as

$$C_q^{-1} \simeq \frac{2}{\kappa}\log\left(\frac{x_{\rm sc}}{2a}\right), \tag{3.112}$$

where $x_{\rm sc}$ is equal to (in the same order) $q_{\rm TF}^{-1}$ and x_s. We conclude that the total capacitance can be modeled after the geometric one,

$$C^{-1} = \frac{2}{\kappa}\log\left(\frac{2a_{\rm eff}}{l}\right) \tag{3.113}$$

with the effective gate-graphene separation

$$a_{\rm eff} = a + \frac{x_{\rm sc}}{2}\,, \tag{3.114}$$

where a on the right-hand side was added by hand to recover the result $a_{\rm eff} = a$ expected for a perfect metal, $x_{\rm sc} = 0$. The similarity of Eqs. (3.106) and (3.114) illustrates once again that quantum capacitance is a measure of the screening length of a system. Whereas in the planar-gate geometry this length is formally divergent for undoped graphene, for the linelike gate the divergence is regularized by nonlinearity. Numerically, we find that C_q^{-1} as a function of λ approaches a universal envelope curve (3.111b) shown by the dashed line in Fig. 3.16.

For completeness, we consider the 'perfect' screening regime where

$$e\Phi(x) \simeq n_{\rm ext}(x)\,\frac{d\mu}{dn} \tag{3.115}$$

for all relevant x, so that Eq. (3.110) becomes

$$C_q^{-1} \simeq \frac{1}{e^2}\int_{-\infty}^{\infty}\frac{dx}{\lambda^2}\,n_{\rm ext}^2(x)\frac{d\mu}{dn}\,. \tag{3.116}$$

The inverse thermodynamic density of states $d\mu/dn$ in Eqs. (3.115) and (3.116) is to be evaluated at $n = n_\infty + n_{\rm ext}(x)$. For $n_\infty = 0$ where the integrand scales as $|\lambda|^{1/2}$, we find the analytical result

$$\frac{\kappa}{2}C_q^{-1} \simeq \frac{1}{2\pi\alpha\sqrt{\lambda a}}\,, \quad \lambda > \frac{1}{\alpha^2 a}\,, \tag{3.117}$$

which agrees with our numerical simulations (the dashed-dotted curve in Fig. 3.16).

If $\tilde{\lambda}$ is negative, i.e., if λ and k_∞ have opposite signs, the twin p-n junctions form at some λ which can be estimated from Eq. (3.86) setting $k_F(0)$ to zero. This event — the onset of the ambipolar regime — is marked by a maximum in κC_q^{-1}, which is absent in the unipolar trace for

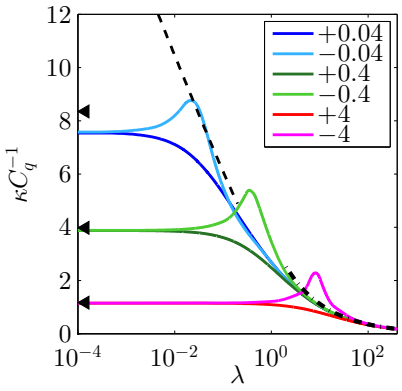

Figure 3.16: (Color online) Inverse quantum capacitance as a function of the gate charge. The curves are from numerical calculations for k_∞ specified in the legend. The units of λ and k_∞ are $(4\pi^3\alpha^2 a)^{-1}$ and $(\pi^3\alpha a)^{-1}$. The triangles, the dashed line, and the dash-dotted line correspond, respectively, to the asymptotic limits of the linear, nonlinear, and near-perfect screening, Eqs. (3.111a), (3.111b), and (3.117). The difference between the curves and the triangles at smaller k_∞ are due to finite-size effects in the simulation. The peaks of the $k_\infty < 0$ curves signal the formation of the twin *p*-*n* junctions.

the same $|k_\infty|$, see Fig 3.16. From dimensional arguments, as k_∞ approaches zero, the height of the maximum in κC_q^{-1} measured with respect to its plateau at $\lambda = 0$ should approach a universal number. Figure 3.16 suggests that number is about 1.5. However, such dimensional arguments assume the TFA is valid, which, similar to the case of a single *p*-*n* junction, [ZF08] is the case at small α. To treat a more typical case $\alpha \sim 1$ one needs to go beyond the TFA, which may be a problem for future research.

3.11.5 Conductance

Charged linelike defects are known to significantly influence electron transport in graphene. Grain boundaries strongly reduce the sheet conductivity of large-area graphene, [TBL$^+$12] while bipolar junctions created by nanowire gates cause conductance oscillations. [YK09] In this section we find expressions for the graphene conductance G relevant for both situations.

We consider a scattering problem for a massless Dirac particle with initial momentum

$\mathbf{k} = |k_\infty|(\cos\theta, \sin\theta)$ subject to the potential perturbation

$$-e\Phi(x) = \mu(n_\infty) - \mu(n) = \hbar v[k_\infty - k_F(x)]\,. \tag{3.118}$$

The intermediate equation follows from the TFA, Eq. (3.87). For a weak potential, we can apply the standard perturbation theory to the massless Dirac equation to obtain the reflection coefficient

$$r(\theta) = i\tan\theta \int_{-\infty}^{\infty} e^{2ik_x x} k_F(x) dx\,, \tag{3.119}$$

which is similar to the first Born approximation formula for the Schrödinger equation. [LL91] The region of validity of this formula can be extended beyond the perturbative regime if in the argument of the exponential we replace k_x by $k_F(0)\cos\theta$, the local momentum at the $x = 0$ point where the scattering potential is the 'most' nonanalytic. [LL91] However, this is permissible only if $k_F(x)$ is real at all x, i.e., if all points on the quasiparticle path are classically allowed (no quantum tunneling occurs).

The conductance G is found by summing the transmittances $T(\theta)$ of all the $k_y = k_\infty \sin\theta$ channels:

$$G = \frac{4e^2}{h} \sum_{k_y} T(\theta)\,, \quad T(\theta) = 1 - |r(\theta)|^2\,. \tag{3.120}$$

In the absence of scattering, $\lambda = 0$, the conductance is

$$G_0 = \frac{4e^2}{h} \frac{|k_\infty| W}{\pi}\,, \tag{3.121}$$

where W is the width of the graphene sheet. Assuming $W \gg k_\infty^{-1}$, we can compute G at $\lambda \neq 0$ by replacing the summation with the integration over θ. However, $|r(\theta)|$ diverges as $|\theta|$ tends to $\pi/2$, so that the first Born approximation cannot be used. In fact, the absolute value of the exact reflection coefficient must be approaching unity instead of diverging. We account for this

by cutting off the integration limits at $\bar{\theta}$ where $|r(\bar{\theta})| \sim 1$:

$$G = G_0 \int_0^{\bar{\theta}} T(\theta) \cos\theta d\theta \,. \tag{3.122}$$

For the linear screening regime, $\tilde{\lambda} \ll 1$ [Eq. (3.75)], we obtain

$$\frac{G_0 - G}{G_0} \simeq \begin{cases} \dfrac{\tilde{\lambda}^2}{\alpha^2} \log^2 |\tilde{\lambda}| \,, & |\tilde{\lambda}| \ll \alpha^2 \,, \\ \dfrac{\pi}{\left|\tilde{\lambda}^{-1} + \delta^{-1}\right|} \,, & \alpha^2 \ll |\tilde{\lambda}| \ll \delta \,. \end{cases} \tag{3.123}$$

Accordingly, the conductance G as a function of $\tilde{\lambda}$ exhibits an asymmetric maximum at $\tilde{\lambda} = 0$ drawn schematically in Fig. 3.17. The asymmetry becomes pronounced when $\tilde{\lambda}$ approaches $\delta = 1/(2\log\alpha^{-1})$, where the screening begins to cross over into the nonlinear regime. If $\tilde{\lambda}$ is positive, $k_F(x)$ becomes essentially independent of k_∞. Using Eqs. (3.93) and (3.95) and changing the integration variable $x \to \lambda x$ in Eq. (3.119) one can show that $r(\theta)$ does not depend on λ any more in this regime. Actually, this is clear from dimensional argument: since $r(\theta)$ is dimensionless, it may not depend on λ, which has the units of inverse length. This implies that G ceases to decrease with $\tilde{\lambda}$, leveling at a plateau, $(G_0 - G)/G_0 \simeq \pi\delta$. If $\tilde{\lambda}$ is negative, the situation is quite different. The scattering potential is repulsive. At large $|\tilde{\lambda}|$ it causes the twin *p-n* junctions to appear at $|x| \sim x_\infty$ [Eq. (3.96)], which act as tunneling barriers. Here even the modified Born approximation fails completely. The transmittance $T_{pn}(\theta)$ of each *p-n* junction is given instead by [CF06]

$$T_{pn}(\theta) = \exp\left(-\frac{\pi\hbar v k_\infty^2}{F} \sin^2\theta\right), \tag{3.124}$$

where

$$F = 2.5\,\hbar v \alpha^{1/3} |n'|^{2/3} \tag{3.125}$$

is the electric field at the junction, [ZF08] with n' being the density gradient. Combining these

equation, we obtain

$$T_{pn}(\theta) = \exp\left(-\frac{b_2}{\alpha}\sin^2\theta\right), \quad b_2 \sim \log^{2/3}|\tilde{\lambda}|. \tag{3.126}$$

Multiple reflections of the quasiparticles in the region between the *p*-*n* junctions lead to the conductance oscillations and resonances. The net transmittance is given approximately by the Fabry-Pérot-like formula [YK09]

$$T(\theta) = \left|\frac{T_{pn}(\theta)}{1-[1-T_{pn}(\theta)]e^{i\phi}}\right|^2, \tag{3.127}$$

where ϕ is the phase acquired by a quasiparticle after one roundtrip between the junctions:

$$\phi \simeq 2\int_{-x_\infty}^{x_\infty} dx\, k_F(x) \simeq \frac{1}{\pi\alpha}\log^2|\tilde{\lambda}|. \tag{3.128}$$

The last estimate is obtained using Eqs. (3.93) and (3.96). Conductance minima arise at $\phi = (2m+1)\pi$, where m is an integer. They have the magnitude $G_{\min}/G_0 \sim \sqrt{\alpha/b_2}$ because the transmittance of each junction is appreciable only at small angles [CF06] $\theta < \sqrt{\alpha/b_2}$, cf. Eq. (3.126). Conductance maxima are found at $\phi_m = 2m\pi$, which correspond to $|\tilde{\lambda}_m| \sim \exp(\pi\sqrt{2\alpha m})$. The widths of these maxima and the distance from one to the next increase exponentially as a function of m. The heights of these maxima approach G_0, which is a manifestation of the resonant tunneling phenomenon. In practice, observation of the resonant tunneling requires samples with the mean free path longer than the roundtrip distance $4x_\infty$; otherwise, the conductance is influenced by diffusive transport between and across the *p*-*n* junctions. [FNGS08] In previous experiments with nanowire-gated graphene devices, [YK09] the conductance maxima were found to be significantly lower than G_0 and decreasing with the top gate voltage, i.e., $|\lambda|$, presumably due to disorder scattering.

If the conductance can also be measured in the direction parallel to the gate, we expect it to

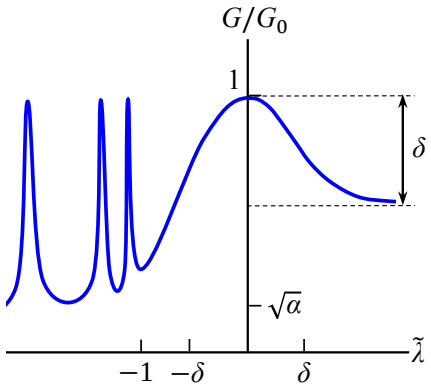

Figure 3.17: Reduced conductance G/G_0 as a function of $\tilde{\lambda}$ (schematically). A monotonic decrease is expected in the unipolar case, $\tilde{\lambda} > 0$, while Fabry-Pèrot oscillations with a period determined by Eq. (3.128) should dominate for negative and large $\tilde{\lambda}$ where the *p-n* junctions form. Parameter δ is defined in the text.

show a dependence consisting of a smooth increase with $|\lambda|$ with superimposed small oscillations due to quantization of the quasi-bound resonant modes. This oscillating part would have the same period as the Fabry-Pérot oscillations in the transverse conductance discussed above. At small gate voltages, both longitudinal and transverse conductances are expected to show additional fine features related to the analog of the Goos-Hänchen effect in graphene, [BSAT09] which is a lateral displacement of a quasiparticle trajectory along the *p-n* interface during reflection. This effect can be included using a more accurate equation for ϕ that incorporates both the path length contribution expressed by Eq. (3.128) and the phase shift of the reflections at the *p-n* interfaces.

3.11.6 Discussion

In this work we considered linelike charged perturbations on graphene and derived analytical expressions for the induced density profiles in both linear and nonlinear screening regimes. These results were applied to the analysis of two types of electronic properties. The first one is the quantum correction to the classical capacitance between the narrow gate and graphene as a function of the top gate voltage. Measuring this quantity will be a direct way for observing the crossovers among different screening regimes and testing our predictions. For

example, we showed that the divergence of the inverse quantum capacitance of undoped graphene predicted from the naive linear-response theory will be curbed by nonlinear screening effects. If the gate creates a strong repulsive potential for charge carriers in graphene, it can induce twin *p*-*n* junctions. Our calculations indicate that the onset of this ambipolar regime is signaled by a peak in the inverse quantum capacitance.

The second quantity we studied is the transverse electrical conductance of the system. Our predictions for the ambipolar regime, where the conductance oscillates as a function of gate voltage, include formulas for the maxima, minima, and the oscillation period. Our theory holds in the ballistic transport regime, which was difficult to probe in earlier experiments on such systems. [YK09] We hope that modern higher-quality devices that utilize graphene encapsulated in boron nitride [DYM$^+$10] and single-wall nanotube gates of smallest possible diameter, would enable a systematic investigation of nonlinear screening and resonant tunneling phenomena we discussed.

Conductance of graphene with charged grain boundaries was previously studied analytically and numerically in Refs. [FXT$^+$11, RSZF13, IZ13]. In Ref. [RSZF13] transport properties were computed modeling charged grain boundaries as short line segments of length $W \ll k_\infty^{-1}$. However, the screening was treated assuming that the boundaries are infinitely long, which seems to require the opposite inequality. These incompatible assumptions make a direct comparison between our analytical results for the conductance difficult. As for screening, only the linear regime was considered in Ref. [RSZF13], and for this our results agree.

Our findings have further implication for experiments using novel scanned-probe techniques. Linear charged defects in the form of grain boundaries have been shown to reflect surface plasmon polaritons [FRG$^+$13, SCH14, GBOR14] and induce photocurrent, [WAGL$^+$15] both of which can be imaged with nanoscale resolution using scanning near-field optical microscopy. Scanning tunneling microscopy is another avenue of approach to measure local density of states. [KWE$^+$13, NIVO$^+$13] We will apply our theory to interpretation of such measurements

in a future work. [NJB+15]

Finally, let us comment on the proposals [KN07] that graphene is a condensed-matter laboratory for exotic effects predicted in other fields of physics. For example, electronic response of graphene to a pointlike charge has an interesting analogy to the atomic collapse of superheavy elements. [PS45, ZP72] It has been shown [Kat06, Nov07, SKL07b, SKL07a, FNS07, PNCN07, KUP+12] that in graphene subcritical $Z < Z_c$ and supercritical $Z > Z_c$ charges produce qualitatively different behavior of the screened electrostatic potential at large distances from the perturbation, the critical charge Z_c being of the order of $1/\alpha$. Characteristic oscillations of the local density of states that appear in the supercritical case have been recently detected experimentally. [WWS+13] Phenomena similar to atomic collapse have been also studied theoretically in the context of narrow-band gap semiconductors and Weyl semimetals. [KSZ13] In turn, our problem of screening of a linelike charge perturbation in graphene have interesting analogies in cosmology (screening of a hypothetical cosmic string by vacuum polarization [NCV99]) and polyelectrolyte physics (Onsager-Manning condensation of counterions [Man69, Oos68]). Our results imply that in graphene nonlinear screening plays a greater role for linelike charges compared to the pointlike ones: the former are *always* supercritical, e.g., there is no threshold λ for the appearance of Friedel oscillations. Finally, our analytical formulas assume $\alpha \ll 1$, which can be realized using a high-κ dielectric substrate [Fig. 3.13(a)], such as $SrTiO_3$ [Refs. [CSM11, DSL12, SLS14, SKJ+14]]. For such gate dielectrics it may be important to consider electric-field dependence of κ in nonlinear screening regimes. [FRS]

3.12 Acknowledgment

Chapter 3, in full, is a reprint of the materials as they appear in Physical Review Letters: Jiang, B.-Y.; Ni, G.; Pan, C.; Fei, Z.; Cheng, B.; Lau, C.; Bockrath, M.; Basov, D.; Fogler, M. "Tunable Plasmonic Reflection by Bound 1D Electron States in a 2D Dirac Metal", Phys. Rev.

Lett. 117, 086801 (2016), and in Physical Review B: Jiang, B.-Y.; Fogler, M. M. “Electronic response of graphene to linelike charge perturbations”. Phys. Rev. B 91, 235422 (2015). The dissertation author was the primary investigator and author of these papers.

Chapter 4

Plasmon reflections by topological electronic boundaries in bilayer graphene

4.1 Introduction

Topological band theory has become a valuable tool for interpreting ground-state properties and low frequency transport in electronic materials with nontrivial momentum-space geometry. [BLD16] In this paper we demonstrate that topological states also play a significant role in the response of such materials at finite frequencies. Our objects of study are domain walls in bilayer graphene that separate regions of local AB and BA stacking order. Because of their prevalence in exfoliated samples[BHH$^+$12, ATH$^+$13, LFZ$^+$13, BDN$^+$13] and their intriguing electronic[Kos13, SJGG$^+$14, JSN$^+$15] and optical properties,[JSZ$^+$16] these domain walls have been under intense investigation.[YJQH16, ZMM13, VLN$^+$13, Kos13] From the point of view of the crystal structure, the stacking wall is a line of partial dislocations. The magnitude of its Burgers vector $|\vec{b}| = a/\sqrt{3}$ is equal to the bond length rather than the lattice constant $a = 0.246\,\text{nm}$.[BDN$^+$13] The domain wall can have an arbitrary angle α with respect to $\vec{b}$, the two limiting varieties being the tensile wall [$\alpha = 0$, Figure 4.1a] and the shear wall [$\alpha = \pi/2$,

Figure 4.1d]. The width $l = 6$–$10\,\mathrm{nm}$ of a domain wall is determined by a competition between the stacking-dependent interlayer interaction and intralayer elastic strain.[ATH+13, LLPK16] The electronic structure of the domain walls and their topological properties are best elucidated using a long-wavelength effective theory[MK13, Kos13] valid for states near the Brillouin zone corners. In this approach the two inequivalent corners are assigned different valley quantum numbers that are practically decoupled since $l \gg a$. Far from the wall, the limit of perfect AB or BA stacking is approached. There, neglecting small "trigonal warping" effects, the dispersion of each valley consists of parabolic conduction and valence bands touching at a point. An electric field applied normal to the bilayer can be used to separate the bands by a tunable energy gap, Figure 4.1a and d. However, at the domain wall gapless electron states must remain since the valley-Chern number of the filled valence band (of a given valley) differs by two in the AB and BA stacked regions. Accordingly, the wall must host a minimum of two co-propagating one-dimensional (1D) conducting channels per spin per valley (PSPV).[ZMM13, LWM+16, MBM08] In particular, for the tensile wall which runs along a zigzag direction (henceforth, the x-direction), the two valleys project to widely separated conserved momenta parallel to the wall. These 1D channels are protected if one neglects intervalley scattering, and so they can support unidirectional valley currents along the wall.

While the above arguments have been used to interpret dc transport experiments,[JSN+15] topological considerations alone do not delineate the response at infrared frequencies ω where optical transitions between many different electron states may contribute. Recent scattering-type scanning near-field optical microscopy (s-SNOM) experiments have demonstrated that in this frequency range the AB-BA walls act as *reflectors* for surface plasmons.[JSZ+16] Here we explicitly show that these reflections probe a structure-sensitive local conductivity of the wall. Specifically, our experiments show that the standing waves formed by the superposition of plasmons launched by the microscope with plasmons scattered by the domain wall generically depend on the type of wall (e.g., tensile or shear), chemical potential, and the potential bias

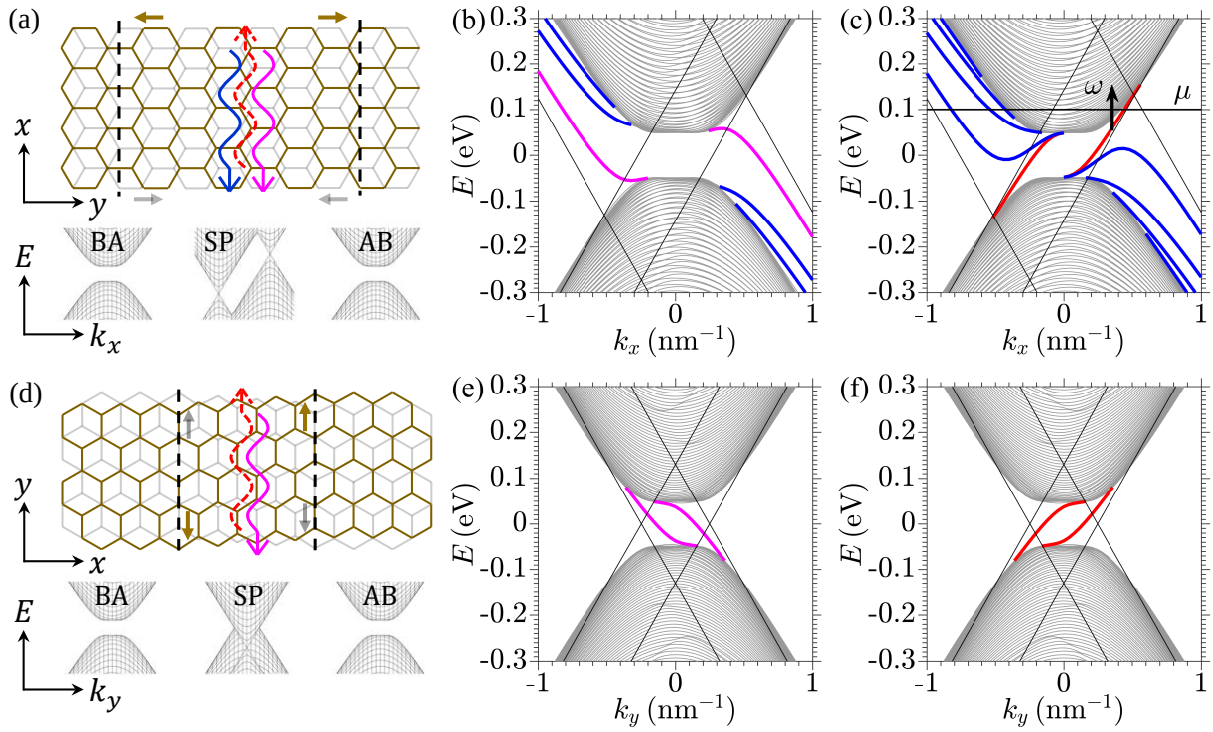

Figure 4.1: (a) Schematic representation of a tensile domain wall and the electronic structure of BA, SP and AB stacking. Brown (gray) arrows indicate the direction of strain for the top (bottom) layer. Red and magenta wavy arrows represent chiral topological modes bound to the domain wall, while the blue wavy arrow represent conventional confined modes. (b), (c) Electronic structure of the wall under a positive (negative) interlayer bias $V_i = V_{\text{top}} - V_{\text{bot}}$ for the K valley. The bound states exist outside the continua but within the boundaries of the SP bands indicated by the thin black lines. The direction of propagation of the topological states is determined by the sign of V_i, while the direction for the conventional ones are fixed by the structure of the SP band and unchanged by V_i. (d)-(f), Similar plots for the shear wall. The shear wall has a smaller width, hosts no conventional bound states, and thus have a distinctly different optical response from the tensile wall.

between the layers. Our modeling demonstrates that while the charge density is nearly unchanged across a domain wall, the local ac conductivity tensor $\sigma_{ij}(\mathbf{r})$ and thus electrodynamic impedance can be changed significantly because of the presence of both topological and conventional 1D conducting channels. (Unlike the topological bound states, the conventional ones do not cross the band gap, see below.) Incorporating the position-dependent $\sigma_{ij}(\mathbf{r})$ into a long-wavelength theory for the plasmon dispersion, we develop a quantitative description of the observed standing waveform and make predictions for future experiments.

4.2 Electronic structure

The presence of the extra 1D channels at the domain wall can be understood as arising from a smooth variation of the stacking order across the wall. If we envision this variation is very gradual ("adiabatic"), then each point is characterized by a local band structure of bilayer graphene with one layer shifted uniformly relative to the other layer by a certain $\vec{\delta} = (0, \delta)$.[Kos13] In this limit the local band structure at the middle of the wall is approximated by SP (saddle point) stacking, while far away the band structure reverts to that of AB or BA, as shown in Figure 4.1a and d for the tensile and shear walls, respectively. These Figures can be understood as projections from the 2D momentum space to the 1D momentum axis parallel to the wall: $k_{\|} = k_x$ for the tensile wall and $k_{\|} = k_y$ for the shear wall. The 2D SP dispersion has two Dirac points shifted in both momentum and energy by the amount determined by the interlayer hopping amplitude $\gamma_1 = 0.40\,\mathrm{eV}$,[ZLB+08] the bias voltage V_i, and the valley index. For a tensile wall, these Dirac points project to different k_x and remain distinct. For a shear wall, they occur at the same k_y and overlap with other states. In both cases the dispersion at the middle of the wall is gapless. This is a crucial property which implies there is a range of energies E in the SP stacking that fall into the gap of the AB stacking. The states at such energies must be confined to the wall. These states can be thought of as electronic "waveguide modes." For an infinitely wide waveguide there are infinitely many such states, and the band structure of the entire system is essentially an overlap of the SP and the AB bands. For the realistic situation of finite-width domain walls, the number of 1D bound states is finite because the quantum confinement produces a finite number of dispersing 1D branches. Figure 4.1b and c illustrate this for a positive and negative interlayer bias V_i, respectively. Among all the confined branches, only one pair PSPV crosses the gap. These are the topological chiral modes inferred from the valley-Chern number mismatch mentioned earlier.[ZMM13, MBM08] The propagation direction of these states is determined by the sign of V_i and the valley index. The remaining confined branches are clearly inherited from the SP

band structure, in agreement with our qualitative picture. These waveguide branches have mostly fixed direction regardless of the sign and magnitude of the bias. Actually, their existence is also a topological effect to some extent since it is facilitated by the gaplessness of the SP dispersion, which is ensured by the persistence of the Dirac points. The latter is a topological property protected by the spatial inversion and the time-reversal symmetries.[Kos13]

The electronic structure of the shear wall, Figure 4.1e and f, can be understood in a similar manner. The shear wall is narrower than the tensile wall, so that the quantum confinement effect is stronger. As a result, the dispersions of the waveguide modes are pushed so close to the boundaries of the conduction and valence bands that they are hardly visible in Figure 4.1e and f. For all practical purposes, the waveguide modes are absent and only the gap-crossing doublets of the chiral states survive.

4.3 Experimental observations

These differences in the band structures for the two types of walls result in different local optical responses, which we observed by imaging plasmonic reflections using s-SNOM. The principles of this experimental technique have been presented in prior works[FRA$^+$12, FRG$^+$13, MKG$^+$14, JZCN$^+$16] and review articles.[KH04, ABJR12, BFL$^+$14] In brief, the s-SNOM utilizes a sharp metallized tip of an atomic force microscope as a nano-antenna that couples incident infrared light to the surface plasmons in the bilayer graphene sheet (Figure 4.2e, inset). These plasmons propagate radially away from the tip and are subsequently reflected by inhomogeneities, in this case, the AB-BA walls. The intensity of the total electric field underneath the tip has a correction determined by the interference of the launched and reflected plasmon waves. The amplitude of this interference term oscillates as a function of the distance between the tip and the reflector, with the period equal to one-half of the plasmon wavelength. The detection of these interference fringes is made by measuring the light backscattered by the tip into the

far field and isolating the genuine near-field signal s. For further details of the experimental procedures, see the end of this Letter.

The interference patterns are visualized in the s-SNOM images of the tensile and the shear wall, shown in Figure 4.2a and b. These images are acquired from two samples of bilayer graphene, each deposited on a SiO_2 substrate grown on top of a Si back gate. For the tensile wall, the pattern contains a single faint interference fringe observed at a gate voltage V_g corresponding to a nearly charge neutral graphene, Figure 4.2a(i). As V_g decreases, the pattern evolves into twin interference fringes, Figure 4.2a(ii-v). For the shear wall, the pattern also starts from a single fringe, which then splits into three, Figure 4.2b. The evolution of the interference patterns can be seen more clearly in Figure 4.2c and d, which show the transverse s-SNOM line profiles averaged over a $1\,\mu$m-long section of the wall.

The observed plasmonic scattering and interference patterns are related to the spatial dependence of the optical conductivity $\sigma_\perp$ of the bilayer graphene sheet. This parameter determines the momentum q_p of the plasmons according to the formula $q_p = \frac{i\kappa\omega}{2\pi\sigma_\perp}$ where κ is the effective permittivity of the environment.[BFL+14] Depending on the type of wall and their respective local electronic structure, the perturbation of local q_p by the wall will vary, leading to distinct plasmonic signatures for each type of wall. Additionally, as the gate voltage V_g is tuned, the carrier density and chemical potential of the bilayer graphene sheet are varied. By tuning V_g and studying the corresponding s-SNOM signal, one effectively probes different parts of the electronic spectrum.

4.4 Local optical conductivity

For our quantitative analysis, we calculate the local optical conductivity of the domain wall following these steps. We start with a model[Kos13] of 4×4 Dirac-type Hamiltonian $H = H(k_\parallel, k_\perp)$ of bilayer graphene with a uniform arbitrary stacking δ. We modify it by allowing

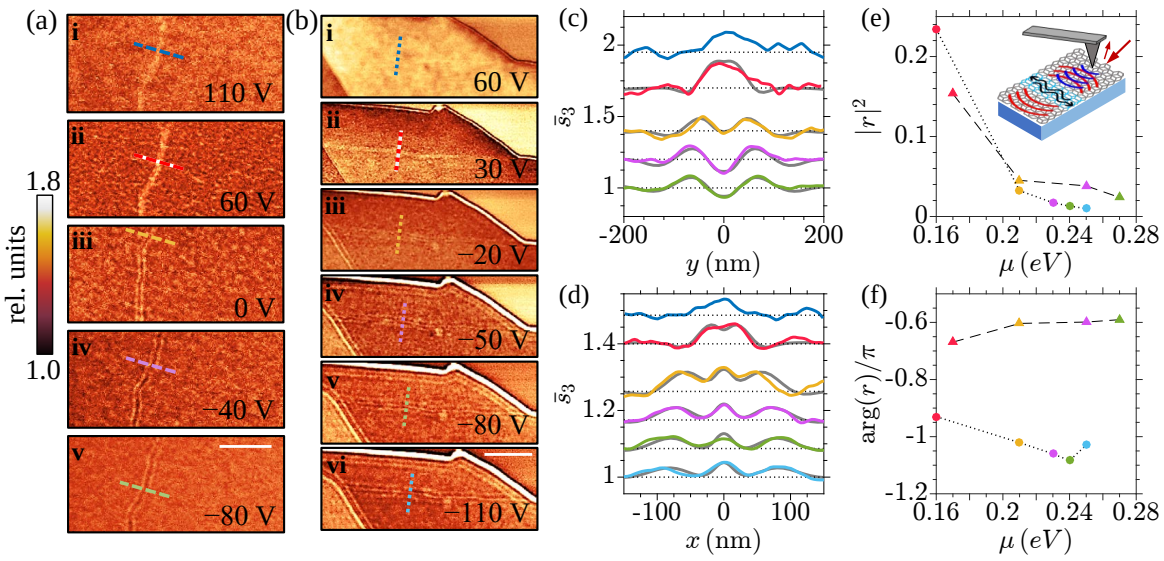

Figure 4.2: (a), (b) Plasmonic interference patterns around a tensile (shear) wall across a range of gate voltages V_g. Scale bar 1 μm. (c), (d) Line profiles across the tensile (shear) wall taken at locations indicated by the colored dashed (dotted) lines in (a) [(b)]. Here the s-SNOM amplitude $\bar{s}_3$ is normalized to its value at $y = -200$ nm ($x = -150$ nm) and offset for clarity, while fits to the experimental profiles are shown in gray. (e), (f) Magnitude (phase) of the plasmon reflection coefficient r of the domain walls calculated from the fits in (c) and (d). The triangles represent the tensile wall while the circles represent the shear wall. (Inset) Schematic of the s-SNOM experiment. The external light source (red arrow) is converted by the AFM tip into plasmons (blue), which are partially reflected by the domain wall (cyan) due to the presence of bound states (black wavy arrows). The plasmonic interference pattern (blue and red) is detected in the form of back-scattered light (small red arrow).

smooth spatial variations of the stacking parameter $\delta = \delta(x_\perp)$. Note that $x_\perp = y$ for the tensile wall and $x_\perp = x$ for the shear wall where x is the zigzag direction of the graphene lattice, see Figure 4.1. The momentum $k_\parallel$ remains a good quantum number. The momentum $k_\perp$ perpendicular to the wall is replaced by the operator $-i\partial/\partial x_\perp$. The resulting eigenproblem is solved numerically on a 1D grid of $x_\perp$ to obtain electronic band dispersions such as those shown in Figure 4.1. Next, these energy dispersions and the wave functions are used to evaluate the Kubo formula for the nonlocal conductivity $\Sigma(x_\perp, x'_\perp)$. We further define an effective local conductivity $\sigma_\perp(x_\perp) \equiv \int \Sigma(x_\perp, x'_\perp) dx'_\perp$, which is appropriate when the total electric field on the sheet varies slowly on the scale of the wall width.[JNP+16] The presence of the bound states is manifest in a strongly enhanced $\sigma_\perp$ at the tensile domain wall. The profile of $\sigma_\perp(x_\perp)$ obtained from this

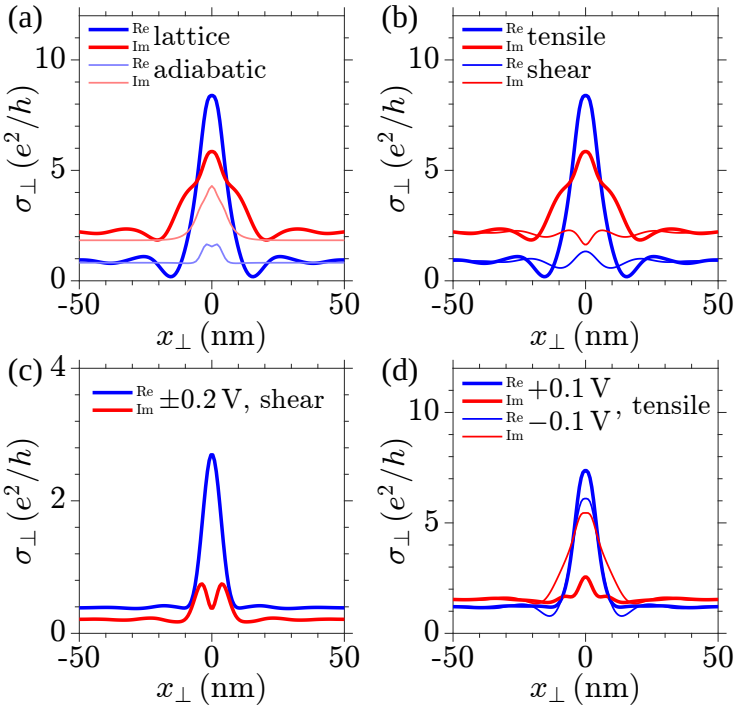

Figure 4.3: (a) Local optical conductivity $\sigma_\perp$ for the tensile wall at interlayer bias $V_i = 0$, calculated using the lattice approach and the adiabatic approach. Conventional bound states residing at the wall lead to the prominent peak in the real part of the conductivity, an effect missing in the adiabatic approach. (b) Comparison between the tensile and shear walls. At zero bias the shear wall hosts no confined modes and the conductivity is relatively flat. (c) Conductivity of the shear wall at $V_i = \pm 0.2\,\mathrm{V}$. The bias opens the band gap, induces the topological chiral modes, and gives rise to the peak in $\mathrm{Re}\,\sigma_\perp$. Reversing the bias has no effect as it is equivalent to interchanging the valleys. (d) Conductivity of the tensile wall at $V_i = \pm 0.1\,\mathrm{V}$. Reversal of the sign of the bias alters the dispersion of the conventional bound states, changing the conductivities. All quantities were calculated at $\mu = 0.1\,\mathrm{V}$, $\eta = 0.1$ and $\omega = 890\,\mathrm{cm}^{-1}$.

nonlocal model can be compared with the results for an adiabatic model in which the Hamiltonian for the uniform stacking is used instead at every point $x_\perp$. As shown in Figure 4.3a, the real part of $\sigma_\perp$ is much larger than that obtained from the adiabatic approximation. The difference occurs precisely because of the lack of the bound states in the latter. On the other hand, for shear walls, which host no bound states when $V_i = 0$, the conductivity is relatively flat at the wall, see Figure 4.3b.

The large contrast of the local optical conductivity for the two types of walls leads directly to the distinct plasmonic profiles they produce. To simulate the measured s-SNOM profile, we

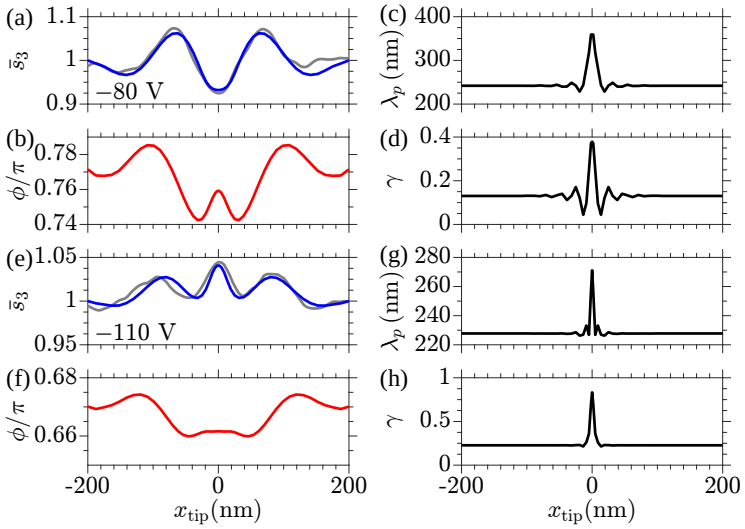

Figure 4.4: (a), (b) Simulated s-SNOM amplitude ($\bar{s}_3$) and phase (ϕ) profiles for the tensile wall at $V_g = -80\,\mathrm{V}$. Experimental data is shown in gray in (a). (c), (d) The plasmon wavelength (λ_p) and damping (γ) profile used for the fit. Peaks in these profiles arise from optical transitions involving the bound states and determine the plasmonic response around the domain walls. (e)-(h) Similar plots for the shear wall at $V_g = -110\,\mathrm{V}$.

converted $\sigma_\perp$ into the plasmon momentum q_p and used it as an input to the electromagnetic solver developed in previous work.[FRA+12, FRG+13, JNP+16, NWW+15, NWG+16, GNP+15] The chemical potential μ, interlayer bias V_i and the phenomenological damping rate η are tuned until the output matches the experimental profile. An example of a fit for the tensile wall is shown in Figure 4.4a-d and for the shear wall in Figure 4.4e-h, where the magnitude and phase of the simulated s-SNOM signal are shown. Here the plasmon momentum q_p is parametrized by the plasmon wavelength λ_p and plasmonic damping rate γ, $q_p = \frac{2\pi}{\lambda_p}(1+i\gamma)$. The interaction of the walls with plasmons can be characterized by the plasmon reflection coefficient r. According to the first-order perturbation theory,[FRG+13] this reflection coefficient is proportional to the amount of excess conductivity at the wall with respect to the background value σ_∞ far from the wall: $r = -iq_\infty \int dx_\perp (\sigma_\perp(x_\perp) - \sigma_\infty)/\sigma_\infty$. The tensile wall has a larger $|r|$; therefore, it reflects plasmons more strongly than the shear wall, see Figure 4.2e. The phase θ of the reflection coefficient $r = |r|e^{i\theta}$, (Figure 4.2f) dictates the position of the extrema of the interference pattern $\sim \cos(2q_p|x_\perp| + \theta + \theta_t)$, where $\theta_t \approx -\pi/2$ is the tip-dependent phase shift.[FRG+13] At the

tensile wall ($x_\perp = 0$, $\theta \approx -\pi/2$) the pattern has a minimum; at the shear wall ($\theta \approx -\pi$) it has a small local maximum, as observed in the experiments. Note that we are treating the interlayer bias V_i as a fitting parameter because in the experiment the value of V_i is determined primarily by (uncontrolled) dopants in the SiO_2 substrate.[ZLB+08] The calculated s-SNOM profiles are strongly V_i-dependent because the sign and value of V_i can alter the bound state dispersions (Figure 4.1b and c) and the conductivity $\sigma_\perp$ dramatically (Figure 4.3c and d). This suggests that determination of V_i from the fits to these profiles is a robust procedure. More generally, the remarkable fidelity of our fits across a range of gate voltages indicates that s-SNOM can be a reliable method for probing the local band structure of electronic materials.

4.5 1D plasmons

Note that prior to the discovery of AB-BA walls, topological confinement of electrons at the "electric-field walls" in bilayer graphene was studied.[MBM08, ZMM13, YJQH16, QJNM11] These are the regions where V_i and therefore, the valley Chern number changes sign. Hence, there is a minimum of $N = 2$ such chiral states PSPV, same as for the AB-BA walls. Based on our study, we expect that N can be system-dependent and larger than two. Recently 1D plasmons confined to "electric-field walls" were also studied theoretically.[HS17] Below we discuss similar collective modes for the AB-BA walls (Figure 4.2e, inset).

The propagating 1D plasmons may exist if their Landau damping by the surrounding bulk is weak enough. This condition can be met if i) the chemical potential μ resides in the band gap so that the confined single-particle modes are the only low-energy degrees of freedom and ii) the frequency ω is small enough so that the optical transitions to bulk bands do not occur. Devices where these requirements can be achieved would likely need a thin top gate in addition to the back gate to control μ and also substrates/gate dielectrics other than SiO_2 to reduce disorder and unintentional doping. The 1D plasmon mode may then be amenable for study by the

s-SNOM nanoimaging, similar to plasmons in carbon nanotubes.[SHB+15] The dispersion of the domain-wall plasmon is determined by the divergences of the loss function $-\mathrm{Im}\,\varepsilon_{1\mathrm{D}}^{-1}\left(k_{\parallel},\omega\right)$. Here $\varepsilon_{1\mathrm{D}}$ is the effective dielectric function[KBF97]

$$\varepsilon_{1\mathrm{D}}\left(k_{\parallel},\omega\right)=\kappa-\frac{8e^2}{h}\ln\left(\frac{A}{k_{\parallel}l}\right)\sum_{j=1}^{N}\frac{k_{\parallel}^2|v_j|}{\omega^2-k_{\parallel}^2v_j^2}\,, \tag{4.1}$$

κ is the dielectric constant of the environment, N is the total number PSPV of the 1D electron states, v_j is the velocity of jth state, and $A\sim 1$ is a numerical coefficient. The dispersion curves calculated using Eq. (4.1) (cf. Supplementary Material for details) are approximately linear in the experimentally accessible range of momenta $k_{\parallel}$, see Figure 4.5a. The slope of each curve, i.e., the plasmon group velocity scales as $\sqrt{N}$, similar to the case of carbon nanotubes.[SHB+15] For shear walls $N=2$ (Figure 4.1e and f), and so the plasmon wavelength $\lambda_p=2\pi/k_{\parallel}$ as a function of μ at given fixed ω and V_i is approximately constant, see the red and green curves in Figure 4.5b. For tensile walls, we can have $N=2$, 4, or 6, depending on the chemical potential and interlayer bias (Figure 4.5b, right panel). Sharp changes in λ_p should therefore occur at some μ where N changes in steps of two, see the blue curve in Figure 4.5b. These properties of plasmons at AB-BA boundaries may be interesting for exploring fundamental physics of interacting 1D systems (so-called Luttinger liquids[KBF97]) or for implementing ultrasmall plasmonic circuits. Note that theoretical calculations for "electric-field walls" predict macroscopically large electron mean-free paths of chiral electronic[QJNM11] (or plasmonic[HS17]) states if the intervalley scattering is dominated by long-range disorder. This prediction should equally apply to the AB-BA boundaries but it is yet to be verified experimentally.

4.6 Conclusion

In summary, the local electrodynamic impedance of bilayer graphene is strongly sensitive to the atomic-scale stacking order, leading to plasmonic reflection that depends on the strength

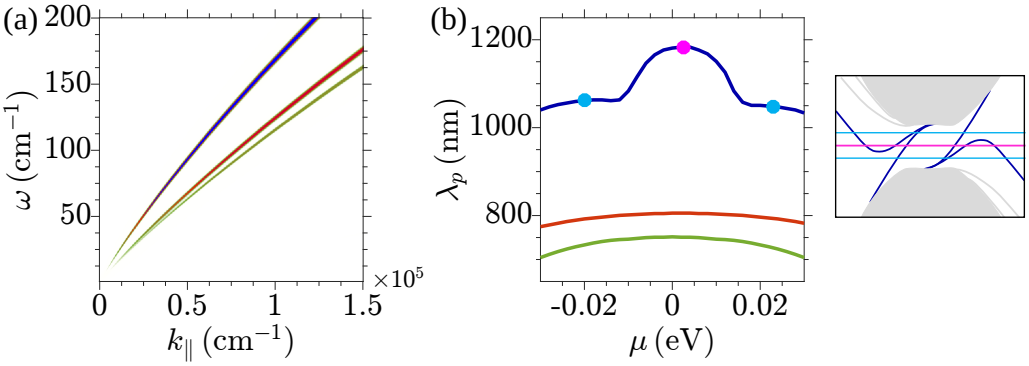

Figure 4.5: (a) Dispersion of 1D plasmons at $\mu = 0$ for **i**. the shear wall at $V_i = 0.1\,\mathrm{V}$ (green, $N = 2$), **ii**. the tensile wall at $V_i = 0.1\,\mathrm{V}$ (orange, $N = 2$) and **iii**. the tensile wall at $V_i = -0.1\,\mathrm{V}$ (blue, $N = 6$). The difference in the plasmon velocity $v_p = \partial\omega/\partial k_\parallel$ is due to the different number of plasmonic channels N, $v_p \propto \sqrt{N}$. (b) (Left panel) Plasmon wavelength for the three cases in (a) at $\omega = 100\,\mathrm{cm}^{-1}$. For case **iii** the plasmon wavelength changes sharply at particular μ's due to a change in N. (Right panel) Corresponding band structure for case **iii**, where N varies between 4 (cyan) and 6 (magenta) at different chemical potentials μ.

and orientation of the strain at the domain wall. Topological arguments give an important initial insight into the physics of this reflection. However, further microscopic analysis of the electronic structure proves to be necessary to account quantitatively for the plasmonic interference fringes observed in near-field nanoimaging. We anticipate many useful applications of these theoretical concepts and experimental approaches to other types of electronic boundaries found in a wide family of van der Waals heterostructures.

4.7 Experimental methods

Experimental procedures are as follows. Graphene flakes were exfoliated onto a 285 nm-thick SiO_2 layer on top of a highly doped Si substrate. Regions of bilayer graphene were identified by their contrast under optical microscopy. Metal contacts were defined on graphene using shadow masks. The infrared nanoimaging experiments were performed at ambient conditions using an s-SNOM based on an atomic force microscope operating in the tapping mode. Infrared light ($\lambda = 11.2\,\mu$m) was focused on the tip of the microscope. A pseudo-heterodyne interferometric

detection was used to extract the scattering amplitude s and phase ϕ of the near-field signal. To remove the background, the signal was demodulated at the third harmonic of the tapping frequency 270 kHz.

4.8 Model for bilayer graphene

Our low energy four-band Hamiltonian for homogeneous BLG is adopted from Ref. [Kos13],

$$\bar{H} = \begin{bmatrix} H_0^+ & U^\dagger \\ U & H_0^- \end{bmatrix}, \tag{4.2}$$

with basis $(F_{At}^{K^\xi}, F_{Bt}^{K^\xi}, F_{Ab}^{K^\xi}, F_{Bb}^{K^\xi})$, where F denotes envelope function, $\xi = \pm 1$ for the K and K' valley, and t stands for top layer and b for bottom layer. The two-band Dirac Hamiltonian for a single layer is

$$H_0^\pm = \hbar v \begin{bmatrix} \pm\frac{1}{\hbar v}\frac{V}{2} & \xi k_x + i k_y \\ \xi k_x - i k_y & \pm\frac{1}{\hbar v}\frac{V}{2} \end{bmatrix}, \tag{4.3}$$

where $V = eV_i$ is the interlayer potential. The interlayer interaction is

$$U = \frac{\gamma_1}{3}\left(1 + 2\begin{bmatrix} \cos\frac{2\pi}{3}\delta & \cos\frac{2\pi}{3}(\delta+1) \\ \cos\frac{2\pi}{3}(\delta-1) & \cos\frac{2\pi}{3}\delta \end{bmatrix}\right), \tag{4.4}$$

where the interlayer coupling[ZLB$^+$08] $\gamma_1 = 0.4\,\mathrm{eV}$ and the stacking order $\delta \in [1,2]$ with $\delta = 1, 1.5, 2$ corresponding to *AB*, *SP* and *BA* stacking. To describe the domain walls, the homogeneous Hamiltonian $\bar{H}$ has to be modified to account for the local change in stacking. This is done by replacing the momentum perpendicular to the wall $k_\perp$ by the operator $-i\partial/\partial x_\perp$ and making the stacking parameter spatially dependent, $\delta(x_\perp)$, resulting in the real space Hamiltonian $H(x_\perp)$. For the tensile wall $x_\perp = y$ while for the shear wall $x_\perp = x$. The distribution $\delta(x_\perp)$ is

found in Ref. [ATH+13] to be

$$\delta(x_\perp) = \frac{2}{\pi}\arctan\left(e^{\pi x_\perp/l}\right) + 1\,, \tag{4.5}$$

where the width $l = 10.1\,\text{nm}$ for the tensile wall and $l = 6.2\,\text{nm}$ for the shear wall.

4.9 Optical conductivity of the domain wall

There are two ways to approximate $\sigma(x_\perp)$. The first is to diagonalize the homogeneous Hamiltonian $\bar{H}$ for a given stacking order δ, use the Kubo formula to find the homogeneous optical conductivity $\bar{\sigma}(\delta)$, then map it to $\bar{\sigma}(x_\perp)$ using the stacking distribution $\delta(x_\perp)$ of the domain wall. This is what we call the "adiabatic" approach and cannot account for the presence of the edge states. The second is to diagonalize the real space Hamiltonian H in coordinate basis and use the Kubo formula to find the nonlocal conductivity $\Sigma(x_\perp, x'_\perp)$, which is then localized by $\sigma(x_\perp) = \int \Sigma(x_\perp, x'_\perp) dx'_\perp$. This "lattice" approach is what we use for our calculations.

Let us start from the calculation of $\bar{\sigma}(\delta)$. The conductivity consists of two parts, an interband conductivity $\bar{\sigma}^I$ from optical transitions between the four bands, and a Drude-like intraband conductivity $\bar{\sigma}^D$. Except for specific stacking orders such as *AB* and *BA* stacking, the conductivities are anisotropic. We consider only the diagonal elements of the conductivity, $\bar{\sigma}_{xx}$ and $\bar{\sigma}_{yy}$, and neglect $\bar{\sigma}_{xy}$ and $\bar{\sigma}_{yx}$ which are small.[SYB+15] The interband conductivity is calculated using the Kubo formula,

$$\bar{\sigma}^I_{\alpha\alpha} = \frac{g_s g_v i\hbar}{4\pi^2}\int dk_x dk_y \sum_{n\neq m} -\frac{f_m - f_n}{E_m - E_n}\frac{e^2 v^2 M^*_\alpha M_\alpha}{\hbar\omega(1+i\eta) - (E_m - E_n)}\,. \tag{4.6}$$

Here $\alpha = x$ or y. The spin and valley degeneracy are $g_s = g_v = 2$. The summation goes over all pairs of states $|n\rangle$ and $|m\rangle$, where the energy of the state $|n\rangle$ is E_n and its occupation number f_n is given by the Fermi-Dirac distribution, $f_n = 1/(1 + e^{(E_n-\mu)/k_B T})$. The matrix element is defined as

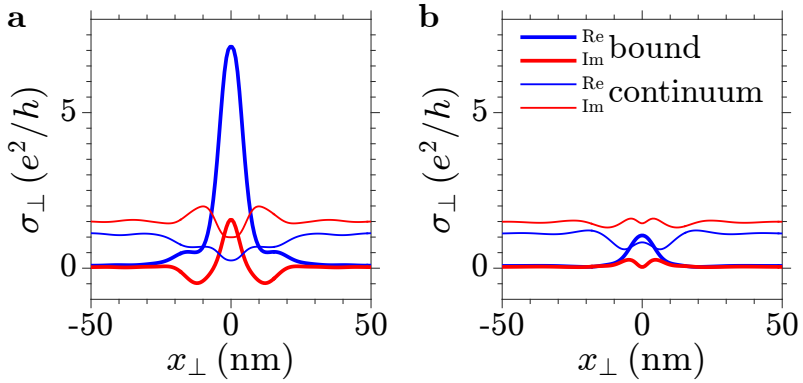

Figure 4.6: **a**. Local conductivity $\sigma_\perp$ for the tensile wall, where the contribution from optical transitions involving the bound states (thick curves) are separated from the contribution of transitions involving only the continuum (thin curves). Parameters: $\mu = 0.1\,\text{eV}$, $T = 300\,\text{K}$, $\omega = 890\,\text{cm}^{-1}$, $V_i = 0.1\,\text{V}$ and $\eta = 0.1$. **b**. Similar quantities for the shear wall. In both cases the bound states produce a prominent peak in the real part of $\sigma_\perp$.

$M_\alpha = \langle m|s_\alpha \otimes \tau_0|n\rangle$ where s_α are the Pauli matrices acting on the sublattice and τ_0 is the identity matrix acting on the layer degree of freedom. The phenomenological damping rate is η.

The intraband conductivity $\bar{\sigma}^D$ arises from the $n = m$ part of the summation, where the fraction $\frac{f_m - f_n}{E_m - E_n}$ is replaced by the derivative $\frac{df_n}{dE_n}$ while $E_m - E_n = 0$, so that

$$\bar{\sigma}^D_{\alpha\alpha} = \frac{g_s g_v i\hbar}{4\pi^2} \int dk_x dk_y \sum_n -\frac{df_n}{dE_n} \frac{e^2 v^2 M^*_\alpha M_\alpha}{\hbar\omega(1+i\eta)}. \tag{4.7}$$

The total conductivity $\bar{\sigma} = \bar{\sigma}^I + \bar{\sigma}^D$ can be readily found given the Hamiltonian $\bar{H}(\delta)$, the chemical potential μ, the temperature T, the frequency ω, the interlayer bias V_i and the damping rate η.

The calculation of the nonlocal conductivity Σ is very similar. The system is discretized in the $x_\perp$ direction into a grid of size N, so that the Hamiltonian H has $4N$ bands. The integration over $k_\perp$ is removed, and the matrix element is calculated at every grid point, $M_\alpha(x_\perp) = \langle m(x_\perp)|s_\alpha \otimes \tau_0|n(x_\perp)\rangle$, leading to the following nonlocal conductivities

$$\Sigma^I_{\alpha\alpha}(x_\perp, x'_\perp) = \frac{g_s g_v i\hbar}{4\pi^2} \int dk_\parallel \sum_{n\neq m} -\frac{f_m - f_n}{E_m - E_n} \frac{e^2 v^2 M^*_\alpha(x_\perp) M_\alpha(x'_\perp)}{\hbar\omega(1+i\eta) - (E_m - E_n)}, \tag{4.8}$$

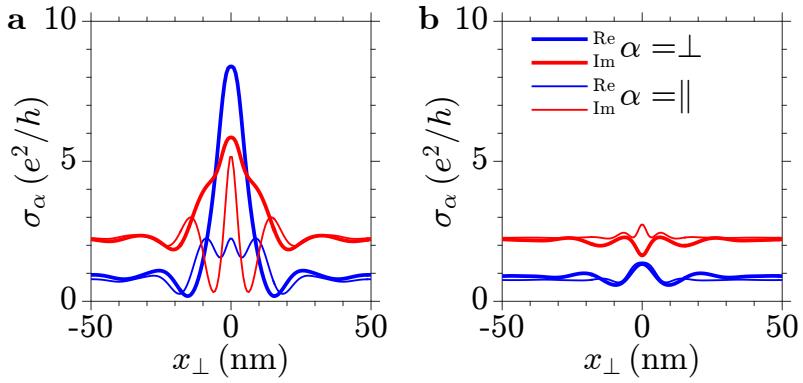

Figure 4.7: **a**. The local conductivity σ_α is highly anisotropic at the tensile wall. Parameteres: $\mu = 0.1\,\text{eV}$, $T = 300\,\text{K}$, $\omega = 890\,\text{cm}^{-1}$, $V_i = 0$ and $\eta = 0.1$. **b**. Similar quantities for the shear wall.

$$\Sigma^D_{\alpha\alpha}(x_\perp, x'_\perp) = \frac{g_s g_v i\hbar}{4\pi^2} \int dk_\parallel \sum_n -\frac{df_n}{dE_n} \frac{e^2 v^2 M^*_\alpha(x_\perp) M_\alpha(x'_\perp)}{\hbar\omega(1+i\eta)}\,. \tag{4.9}$$

For our calculations the nonlocal Σ is then localized by integration over $x'_\perp$ and denoted $\sigma_\alpha \equiv \sigma_{\alpha\alpha}$, where $\alpha = \perp$ or $\parallel$.

As the bound state wavefunctions are localized at the domain wall, optical transitions involving these bound states give rise to conductivity peaks at the wall, as shown in Figure 4.6. The domain wall also introduces anisotropy to the local conductivity, as shown in Figure 4.7a for the tensile wall and 4.7b for the shear wall. Away from the wall, conductivities in the $\perp$ and the $\parallel$ direction have the same value as expected, but at the wall they can be drastically different.

4.10 Fitting the s-SNOM profiles

To fit the experimental s-SNOM profiles, we calculate conductivities at $T = 300\,\text{K}$ and $\omega = 890\,\text{cm}^{-1}$, while treating the chemical potential μ, the interlayer bias V_i and the damping rate η as fitting parameters. As shown in Figure 4.8, changes to these three parameters have drastic effects on the resulting s-SNOM signal around the domain wall. An increase in μ increases the plasmon wavelength and decreases the strength of the signal, a change to V_i changes the signal

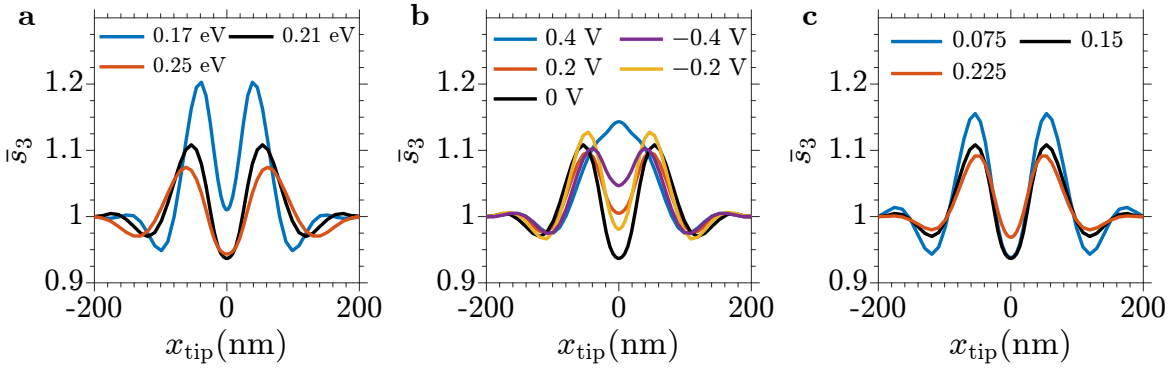

Figure 4.8: Comparison of near-field amplitude profiles under different fitting parameters. The black curve in every panel is calculated at $\mu = 0.21\,\mathrm{eV}$, $V_i = 0\,\mathrm{V}$, and $\eta = 0.15$ for the tensile wall. In each panel one of the three parameters is varied. **a**. Varying the chemical potential changes the plasmon wavelength and the overall amplitude. **b**. Changing the interlayer bias V_i alters signal strength at the wall. **c**. Increasing the damping rate η decreases the overall amplitude of the oscillations.

strength at the wall, while an increase in η decreases the overall amplitude of the oscillations. This shows that one can reliably determine these three parameters in the fitting procedure.

In Figure 4.9 we show our fits to the experimental near-field amplitude profiles for the tensile wall along with the phase ϕ of the s-SNOM signal. Also shown are the plasmonic wavelength profile λ_p and the plasmonic damping profile γ used for the fit. Parameters used for the series of fits for $V_g = (60, 0, -40, -80)\,\mathrm{V}$ are: $\mu = (0.17, 0.21, 0.25, 0.27)\,\mathrm{eV}$, $V_i = (0.25, 0.2, -0.1, -0.2)\,\mathrm{V}$, and $\eta = (0.2, 0.15, 0.1, 0.12)$. Fits for the shear wall are shown in Figure 4.10. The parameters used for $V_g = (30, -20, -50, -80, -110)\,\mathrm{V}$ are: $\mu = (0.16, 0.21, 0.23, 0.24, 0.25)\,\mathrm{eV}$, $V_i = (0.3, 0.35, 0.35, 0.35, 0.4)\,\mathrm{V}$, and $\eta = (0.2, 0.2, 0.2, 0.2)$.

4.11 Dielectric function in the band gap

In this section we derive the effective 1D dielectric function $\varepsilon_{1\mathrm{D}}(k_{\parallel}, \omega)$ of the domain wall when the chemical potential lies within the band gap. The pole of $1/\varepsilon_{1\mathrm{D}}$ determines the dispersion of the 1D plasmon propagating along the wall.

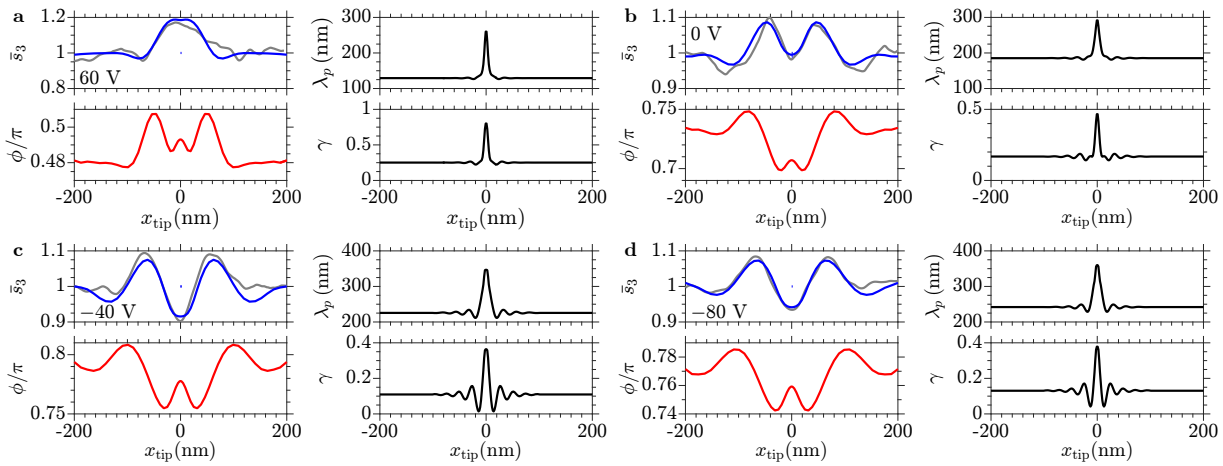

Figure 4.9: Fits for the near-field profiles for the tensile wall. **a**. $V_g = 60\,\mathrm{V}$. **b**. $V_g = 0\,\mathrm{V}$. **c**. $V_g = -40\,\mathrm{V}$. **d**. $V_g = -80\,\mathrm{V}$. In each panel the normalized experimental near-field amplitude profile $\bar{s}_3$ is shown in gray, the simulated amplitude $\bar{s}_3$ and phase ϕ profiles are shown in blue and red. Also shown are the plasmon wavelength profile λ_p and damping profile γ used for the fit.

In the absence of external fields, the total electric potential Φ of the sheet in the quasistatic limit is determined by the charge density ρ and current density j on the sheet,

$$\Phi = V_2 * \rho = V_2 * \frac{i}{\omega} \nabla \cdot \mathbf{j}\,, \tag{4.10}$$

where the Coulomb kernel $V_2 = 1/\kappa r$, $\mathbf{r} = (x,y)$ and $*$ denotes convolution, $A * B = \int A(r - r')B(r')dr'$. For ease of notation we assume that the domain wall lies on the y-axis. When the chemical potential is in the gap, $|\mu| < V/2$ (and the temperature and frequency are low, $k_B T \ll V$ and $\hbar\omega \ll V$), only the bound states contribute to the optical response. The charge density is zero on the sheet and the current only flows along the domain wall, so we can make the simplification $j_x = 0$ and $\Phi = \phi(x)e^{iq_y y}$. Eq. (4.10) can then be rewritten as

$$\phi(x) = V_1 * \frac{i}{\omega}\partial_y j_y = V_1 * \frac{k_y^2}{i\omega}\int \Sigma_{yy}(x,x')\phi(x')dx'\,, \tag{4.11}$$

Here the 1D Coulomb kernel is $V_1(x) = \int V_2 dy = \frac{2}{\kappa}K_0(q_y|x|)$, where K_0 is the modified Bessel

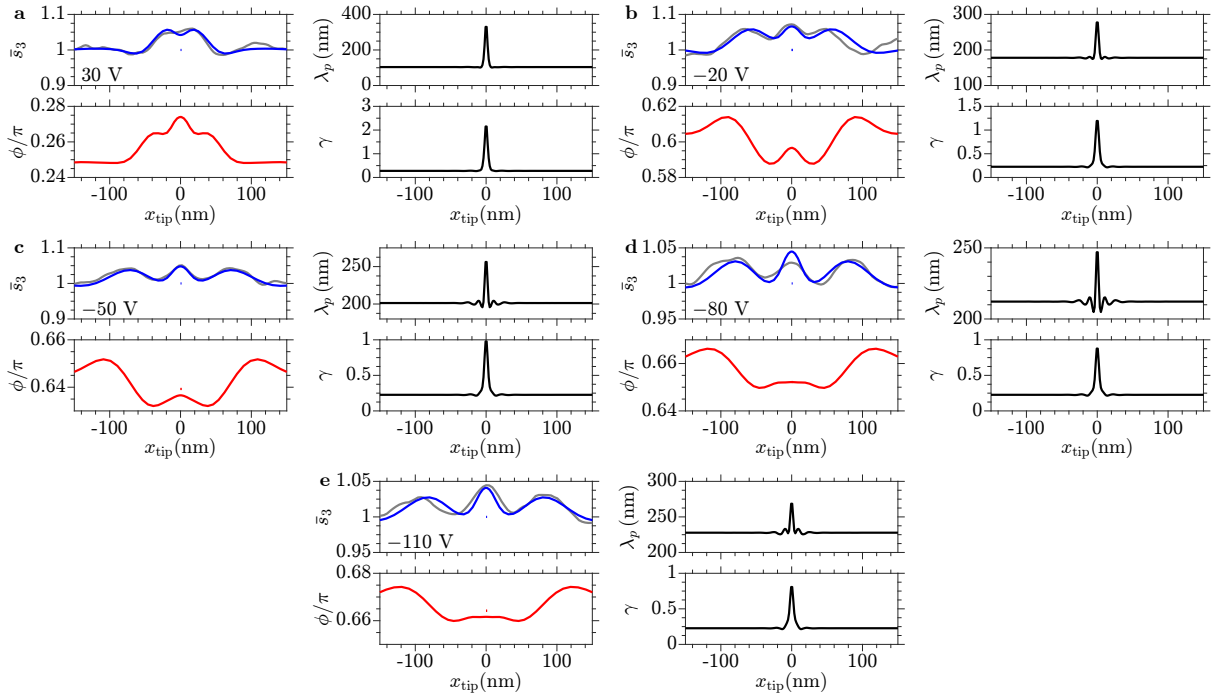

Figure 4.10: Fits for the near-field profiles for the shear wall. **a**. $V_g = 30\,\mathrm{V}$. **b**. $V_g = -20\,\mathrm{V}$. **c**. $V_g = -50\,\mathrm{V}$. **d**. $V_g = -80\,\mathrm{V}$. **e**. $V_g = -110\,\mathrm{V}$. In each panel the normalized experimental near-field amplitude profile $\bar{s}_3$ is shown in gray, the simulated amplitude $\bar{s}_3$ and phase ϕ profiles are shown in blue and red. Also shown are the plasmon wavelength profile λ_p and damping profile γ used for the fit.

function of the second kind. Note that we removed the k_y dependence in Σ_{yy} as the plasmon wavelength $\lambda_y \sim k_y^{-1}$ is much larger than all other length scales in the problem, so that we can make the approximation $\Sigma(x,x',k_y) \simeq \Sigma(x,x',0)$.

At frequencies $\hbar\omega \ll V$, there are no allowed optical transitions and the conductivity comes purely from the Drude response,

$$\Sigma_{yy}(x,x') = g_s \sum_{K,K'} \sum_{j=1}^{N} \frac{iD_{yy,j}}{\pi(\omega - \xi_j v_j q_y)} |\psi_j(x)|^2 |\psi_j(x')|^2 \,, \tag{4.12}$$

where ψ_j is the wavefunction of the j-th bound state at energy $E_j = \mu$ and $g_s = 2$ is the spin degeneracy. The Drude weight $\frac{1}{\pi} D_{yy,j} = \frac{e^2}{h} |v_j|$ is directly proportional to the particle group velocity $v_j = \partial E_j / \hbar \partial k_y$, and ξ_j is the sign of v_j. The summation over the K and the K' valleys can

be reduced by noting that every bound state has a counterpart in the other valley with a velocity that is equal in magnitude but opposite in direction.

Since the width of the wavefunctions $\sim l$ is much smaller than the plasmon wavelength λ_y, the particle density distributions $|\psi_j(x)|^2$ can be roughly approximated as a δ-function of characteristic width l. Eq. (4.11) then becomes

$$\phi(l) \simeq \left(\sum_{j=1}^{N} \frac{k_y^2}{\kappa(\omega^2 - v_j^2 k_y^2)} 2K_0(k_y l) g_s \frac{2e^2}{h} |v_j| \right) \phi(l) = \left(1 - \frac{\varepsilon_{1D}}{\kappa}\right) \phi(l) \,. \tag{4.13}$$

For small arguments $K_0(z) \simeq \log(A/z)$ where $A \simeq 2e^{-0.577} = 1.12$, and so the 1D dielectric function is

$$\varepsilon_{1D}(k_y, \omega) = \kappa - \frac{8e^2}{h} \ln\left(\frac{A}{k_y l}\right) k_y^2 \sum_{j=1}^{N} \frac{|v_j|}{\omega^2 - k_y^2 v_j^2} \,. \tag{4.14}$$

4.12 Acknowledgment

Chapter 4, in full, is a reprint of the material as it appears in Nano Letters: Jiang, B.-Y.; Ni, G.; Addison, Z.; Shi, J. K.; Liu, X; Zhao, S. Y. F.; Kim, P.; Mele, E. J.; Basov, D. N.; Fogler, M. M. "Plasmon reflections by topological electronic boundaries in bilayer graphene", Nano Lett. 17, 7080 (2017). The dissertation author was the primary investigator and author of this paper.

Chapter 5

Theory of plasmon reflection by a 1D junction

5.1 Introduction

Plasmonics aims to combine the advantages of nanometer scale electronics with the high operating frequency (terahertz and beyond) of photonics.[Ozb06] A promising platform for plasmonics is graphene, which features high confinement, wide range of operating frequencies ω, long lifetimes, and tunability.[GPN12, FRA$^+$12] Recent experiments demonstrated long plasmon propagation distance and high quality factor for graphene,[WLG$^+$14, NWG$^+$16] making 2D plasmonic circuitry feasible. An important basic element in a plasmonic circuit is a switch that has a small size and a large on-off ratio. Numerous numerical studies of the interaction of plasmons with 1D obstacles has been done in search of such a device.[RSK$^+$04, PTSK05, SRC05, PPN12, GPNMM13, PFM13, WGT$^+$17] It has been shown that a narrow 1D junction of low conductivity in an otherwise uniform 2D conductive sheet can potentially serve as a plasmonic switch.[GDPC13] This is a surprising result as deeply subwavelength obstacles usually cannot impede the propagation of a wave. In this work we explain the physical principles behind the

plasmonic interaction with this type of inhomogeneities and provide analytical solutions for the reflection coefficient, which can be simply understood in terms of equivalent circuits.

Plasmons propagating in a uniform conducting film or a two-dimensional electron gas (2DEG) have a momentum q that is inversely proportional to the (frequency-dependent) sheet conductivity σ, $q = \frac{i\kappa\omega}{2\pi\sigma}$, where κ is the dielectric function of the environment exterior to the sheet. A local variation in the conductivity causes a change in q and thus acts as a scatterer for plasmons. Controlled conductivity variations can be realized in graphene which has a conductivity σ determined by its chemical potential. Using patterned electric gates, the chemical potential can be tuned locally. We consider an idealized case where the sheet conductivity has a piecewise constant 1D profile with a value σ inside a strip of width $2a$ and another value σ_0 in the semi-infinite leads on both sides (Figure 5.1). In practice, the width of the junction is determined by the geometry of the gate and can be as narrow as a few nanometers for a nanowire or nanotube gate.[JF15, JNP$^+$16] Two types of phenomena govern the propagation of plasmons across such a junction, the cavity resonances inside the junction and the capacitive coupling between the leads. Depending on how the width $(2a)$ of the junction compares to the plasmon wavelength in the leads $(\lambda_0 = 2\pi/\mathrm{Re}\, q_0)$ and in the junction (λ), the strength of these phenomena varies. For wide junctions $\lambda_0 \ll a$, the reflectivity is determined chiefly by the cavity resonances while the capacitive coupling is negligible. The resonances give rise to alternating maxima and minima in the reflectance, which are, however, quickly suppressed by plasmonic damping. For narrow junctions $\lambda_0 \gg a$, the capacitive coupling dominates. This regime can be further divided into two, depending on how a compares to λ. When $a \gg \lambda$, the cavity resonances are present but have their amplitudes modified by the capacitive coupling. When $a \ll \lambda$, instead of the oscillating resonant fields, the junction has a constant electric field like a capacitor and becomes a parallel *RC* circuit in the dc limit. A near perfect plasmon reflection occurs at a specific conductivity σ where the impedance of the *RC* circuit diverges. As there is no wave propagation in the junction in this limit, the anomalously strong reflection is robust against plasmonic damping.

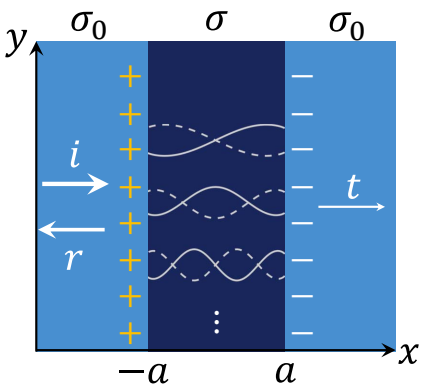

Figure 5.1: A normally incident plasmon is partially reflected and transmitted by a 1D junction of conductivity σ different from the background value σ_0. The strength of reflection is determined by two types of effects, the capacitive coupling of the two edges of the leads (represented by the $+$ and $-$ electric charges) and the cavity resonances in the strip. The field profiles of the first few resonant modes (white solid and dashed curves) are calculated for the case of a narrow junction with an infinite conductivity contrast between the gap and the leads.

5.2 Wide junctions

. A wide junction contains two interfaces that partially reflect and transmit plasmons but otherwise do not interact with each other, i.e., it is a plasmonic cavity. Using the analytical Wiener-Hopf method, the reflection coefficient r for a normally incident plasmon wave from the left at the left (L) and the right (R) interfaces was found in Ref. [RK15] to be

$$r_L = e^{i\theta}\frac{\sigma_0-\sigma}{\sigma_0+\sigma}, \quad r_R = e^{-i\theta}\frac{\sigma-\sigma_0}{\sigma+\sigma_0}, \quad t_L = t_R = \frac{2\sqrt{\sigma_0\sigma}}{\sigma_0+\sigma}, \tag{5.1}$$

where the reflection phase shift is given by

$$\theta = \frac{\pi}{4} - \frac{2}{\pi}\int_0^\infty du\,\frac{\tan^{-1}(\frac{\sigma}{\sigma_0}u)}{u^2+1}, \tag{5.2}$$

which approaches $\pi/4$ in the limit $\sigma \to 0$, as in the case where the lead-junction boundary corresponds to a physical termination of the lead.[NLMM14, KWS+17] We define the reflection coefficient r to be the prefactor in the asymptotic form of the scattered potential $\phi_s \simeq re^{-iq_0x}$ at

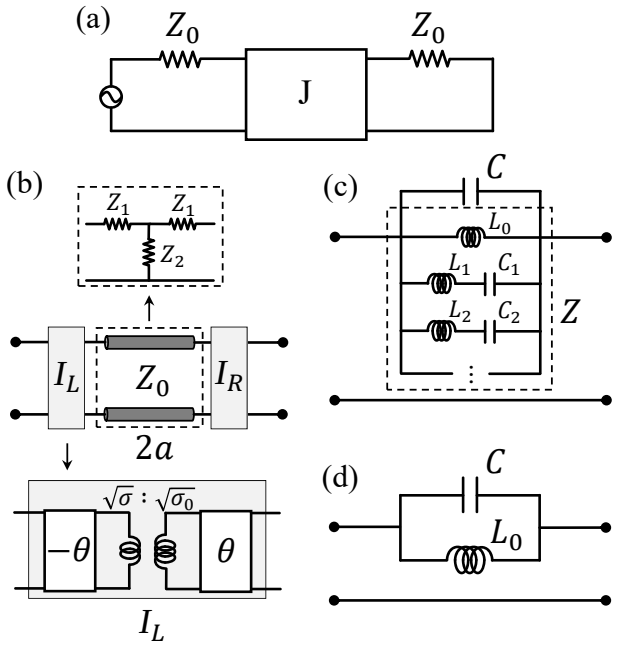

Figure 5.2: (a) Equivalent circuit for the system. The leads have impedance Z_0, while the junction J has different representations depending on its width a and conductivity σ. (b) For wide junctions J is represented by two interfaces and a waveguide of length $2a$. The waveguide can be replaced by a T-junction, while the interface I consists of two phase shifters and an ideal transformer. The right interface I_R has reversed number of coils and signs of θ compared to I_L. (c) For narrow junctions J is a capacitor in parallel with a LC network that describes the cavity resonances. (d) In the dc limit, $a \ll \lambda$, the LC network reduces to L_0 and J becomes a parallel RC circuit.

large negative x for an incident plasmon potential $\phi_0 = e^{iq_0 x}$. (Our definition differs in the overall sign from Ref.[RK15] where r is the reflection coefficient for the current.) The total reflection coefficient from the left interface of the junction is then found from the usual Fabry-Pérot (F-P) formula to be

$$r_{\rm FP}^{-1} = \frac{\sigma_0^2+\sigma^2}{\sigma_0^2-\sigma^2} + i\frac{2\sigma_0\sigma}{\sigma_0^2-\sigma^2}\cot(\phi-\theta)\,, \tag{5.3}$$

where $\phi = 2qa$ is the phase accumulated across the junction. Equation (5.3) yields a reflectance that has local minima at resonances, $\phi = n\pi+\theta$, and local maxima at anti-resonances, $\phi = (n+\frac{1}{2})\pi+\theta$, where n is a positive integer. These alternating maxima and minima can be seen in Figure 5.3a, where an example of the reflectance $R = |r|^2$ for a lossless junction of width

$a = \lambda_0$ is shown. Here we have parametrized the plasmon momentum as $q = \frac{2\pi}{\lambda}(1 + i\gamma)$ with the dimensionless damping factor defined as $\gamma = \mathrm{Im}\, q/\mathrm{Re}\, q$. Under these circumstances, the junction acts as a plasmonic switch with a high on-off ratio, tuned by the conductivity contrast σ_0/σ (or alternatively, by varying the width a). However, the resonances are quickly suppressed by plasmonic damping as shown in Figure 5.3b, removing the switching behavior and rendering wide junctions less desirable for nanoplasmonic applications. Note that we assume the same damping in the leads and the junction, $\gamma_0 = \gamma$, in this work, so that the conductivity ratio σ_0/σ is real. A higher damping in the junction than in the leads further diminishes the resonances but does not alter the results qualitatively (see Supporting Information, Sec. 9.)

A wide junction can be described with an equivalent circuit. The reflection and transmission coefficients described in Eq. (5.1) is what one would expect at the interface of two waveguides with impedances σ_0 and σ (up to the phase factor in r).[MDP48] However, the plasmonic wave impedance of a 2DEG, $Z_0 = \pi/\omega\kappa$, was found to be independent of the sheet conductivity[FAA+16] (see Supporting Information, Sec. 2.) To retain the analogy to conventional waveguides, we treat the interface as a composite object consisting of an ideal transformer and two phase shifters (Figure 5.2b). The transformer effectively rescales the junction impedance into $Z_j = Z_0\frac{\sigma_0}{\sigma}$ while the complementary phase shifters provide the requisite phase factor $e^{\pm i\theta}$ to the reflection coefficients. The junction itself can be replaced by a piece of waveguide of the effective length $2a - (\theta/q)$, which is equivalent to a T-junction with impedances $Z_1 = -iZ_j \tan(qa - \theta/2)$ and $Z_2 = iZ_j \csc(2qa - \theta)$.[MDP48] The total reflection coefficient from the left interface [coinciding with Eq. (5.3)] can then be found directly from the standard waveguide-theory formula,

$$r = (Z_L - Z_0)/(Z_L + Z_0)\,, \tag{5.4}$$

where the total load impedance, calculated by the rules of parallel and series connections, is $Z_L = Z_1 + (Z_2^{-1} + (Z_1 + Z_0)^{-1})^{-1}$.

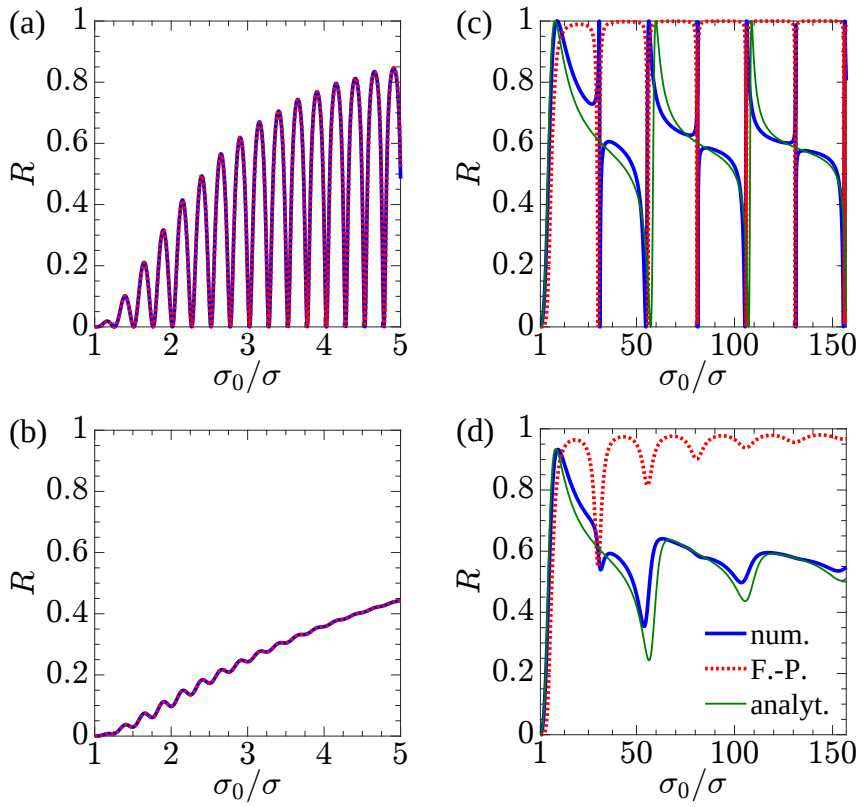

Figure 5.3: (a) Reflectance of a "wide" junction, $a = \lambda_0$ without damping. (b) Similar quantities for damping $\gamma = 0.05$. (c) Reflectance of a "narrow" junction, $a = 0.01\lambda_0$ without damping. Blue curves are numerical results, dotted red curves are from the F-P formula (5.3), and green curves are from Eq. (5.15). (d) Similar quantities for $\gamma = 0.05$.

5.3 Narrow junctions

. The F-P formula Eq. (5.3) remains numerically accurate for junctions as narrow as $a \simeq \lambda_0$. However, as the two leads get closer, they start to couple capacitively. The system resembles a leaky capacitor. The capacitive coupling becomes dominant when the junction is narrow. The cavity resonances remain but their amplitude gets strongly modified. This can be seen in Figure 5.3c, where the reflectance for a narrow junction $a = 0.01\lambda_0$ deviates greatly from the F-P formula but retains the same resonant locations. In the following we present key intermediate steps in calculating the analytical form of this modified reflectance, leaving the detailed derivation to the Supporting Information.

We start by discussing the general form of r for a narrow junction. The reflection coefficient r can be found by comparing the incident in-plane plasmon field $\mathbf{E_i} = E_0 e^{iq_0 x}\hat{x}$ with the asymptotic form of the scattered field $\mathbf{E_s} \simeq -rE_0 e^{-iq_0 x}\hat{x}$ at large distances from the junction. The scattered field E_s can be completely determined by the total field E inside the junction via the Green's function G,

$$E_s(|x| > a) = \frac{\Delta\sigma}{\sigma_0}\int_{-a}^{a} dx' \, G(x-x')E(x'), \tag{5.5}$$

where $\Delta\sigma \equiv \sigma - \sigma_0$, and the Green's function G is the 1D Fourier transform of the dielectric function $\varepsilon = 1 - |q|/q_0$ of the 2D sheet,

$$G(x) = \int_{-\infty}^{\infty} \frac{dq}{2\pi} e^{iqx}\varepsilon^{-1}(q). \tag{5.6}$$

By analyzing the asymptotic form of Eq. (5.6), one arrives at the relation

$$r = i\frac{\Delta\sigma}{\sigma_0}\frac{q_0}{E_0}\int_{-a}^{a} dx E(x)e^{iq_0 x}, \tag{5.7}$$

which is valid for junctions of *any* width. For example, it can be used to evaluate the reflection coefficient numerically (see Supporting Information, Sec. 6.) For narrow junctions, Eq. (5.7) can be simplified and written in terms of the voltage drop across the junction $V = \int_{-a}^{a} dx E(x)$,

$$r \simeq i\frac{\Delta\sigma}{\sigma_0} q_0 \frac{V}{E_0}, \quad (a \ll \lambda_0) \tag{5.8}$$

since $e^{iq_0 x} \approx 1$ in the junction.

Equation (5.8) has simple solutions in limiting cases. In the perturbative limit $\sigma \lesssim \sigma_0$, the current density $j(x)$ is approximately constant across the junction, which implies that the field

inside the junction is $E \simeq \frac{\sigma_0}{\sigma} E_0$. This yields the reflection coefficient

$$r = 2iq_0 a \frac{\Delta\sigma}{\sigma}, \quad (|\Delta\sigma| \ll \sigma_0) \tag{5.9}$$

in agreement with Ref. [FRG+13]. The reflection coefficient depends linearly on the strength of the perturbation $\Delta\sigma$, as expected. In fact, the perturbative expression is also valid for any $\sigma > \sigma_0$ as long as the resulting reflection coefficient is small, $|r| \ll 1$. However, it is clearly inadequate as σ is decreased toward zero, i.e., the case of a vacuum gap, where the perturbative r diverges [Eq. (5.9)]. For such a vacuum gap, the current j is stopped and completely reflected by the gap, i.e., the two edges of the junction act as a (non-leaky) capacitor. By considering the charge conservation equation on either edge of the capacitor, $j_i + j_r = \sigma_0 E_0 (1 - r) = \dot{Q} = -i\omega V C$, the reflection coefficient for the vacuum gap can be found:

$$r = \frac{i\kappa}{2\pi C + i\kappa} = \frac{i\pi}{\log \frac{2}{q_0 a} - c + i\pi}, \quad (\sigma = 0\,, a \ll \lambda_0)\,. \tag{5.10}$$

Here the junction capacitance per unit length

$$C = \frac{\kappa}{2\pi^2} \left(\log \frac{2}{q_0 a} - c \right), \quad c = 0.577\ldots \tag{5.11}$$

can be derived by considering the charge distribution on perfectly conducting leads (see Supporting Information, Sec. 5.) The sublinear dependence on a is due to the 2D geometry of the system, where the fringing fields, i.e., fields exterior to the sheet, dominate. Equation (5.10) is in good agreement with our numerical result (Figure 5.4) and with Ref. [GPNMM13]. It can also be simply explained by the equivalent circuit for the system, a resistor-capacitor-resistor series with impedances Z_0, $Z_c = i/\omega C$, and Z_0. The circuit yields the load impedance $Z_L = Z_c + Z_0$, so that $r = Z_c/(Z_c + 2Z_0)$, matching Eq. (5.10).

For intermediate conductivities $0 < \sigma < \sigma_0$, the capacitive component is diminished

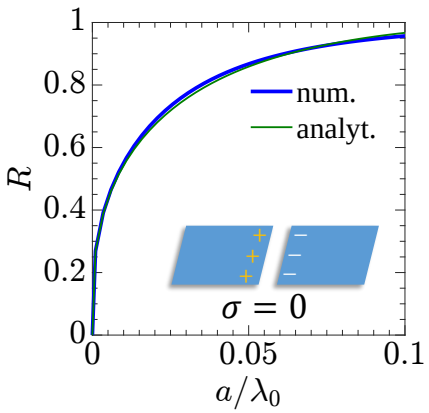

Figure 5.4: Reflectance of a narrow vacuum gap in a lossless sheet.

compared to the vacuum gap, but the junction can now host plasmonic resonances. This is seen in the modified charge conservation equation,

$$\sigma_0 E_0(1-r) = -i\omega V C' + \int_{-a}^{a} dx \frac{\sigma E(x)}{\pi\sqrt{a^2-x^2}}, \quad C' = -\frac{\Delta\sigma}{\sigma_0} C, \tag{5.12}$$

where the capacitance acquires a correction factor $-\Delta\sigma/\sigma_0$ which is 1 for the vacuum gap but decreases to 0 for a uniform sheet where $\sigma = \sigma_0$. The additional integral represents the Shockley-Ramo image current in the leads induced by the current in the junction.[RSS03] The field in the junction can be expanded in the basis of resonant eigenmodes

$$E\,(|x| < a) = \frac{V}{2a}\left(1 + \sum_{n=1}^{\infty} b_n f_n(x)\right). \tag{5.13}$$

The coefficients

$$b_n = \frac{2a}{1 + \frac{\sigma}{\Delta\sigma}\frac{q_n}{q_0}} \frac{\int_{-a}^{a} dx \frac{f_n(x)}{\pi\sqrt{a^2-x^2}}}{\int_{-a}^{a} dx\, f_n^2(x)} \tag{5.14}$$

diverge whenever the plasmon wavevector q inside the junction matches the resonant wavevector q_n, and the field E becomes dominated by the n-th resonant mode f_n. The wavevector q_n of the resonant field is again determined by the resonance condition $2q_n a = n\pi + \theta$ as in the case of wide junctions. For narrow junctions, all resonances occur at $\sigma_0/\sigma \gg 1$, so that $\theta \approx \pi/4$ and

$q_n = \frac{\pi}{2a}(n+\frac{1}{4})$. The exact forms of the dimensionless resonant fields f_n has to be calculated numerically, but they quickly approach $\cos(q_n x - \frac{1}{2}n\pi)$ as n is increased.[RSS03, RSK$^+$04] The first three eigenmodes $f_n(x)$, $n = 1,2,3$, are plotted in Figure 1 and also in Supporting Information, Figure S2.

The reflection coefficient can be found by substituting the field (5.13) into Eq. (5.12), then using the general form of r (5.8) to obtain

$$r^{-1} = 1 - i\frac{2\pi}{\kappa}C - i\frac{1}{2q_0 a}\frac{\sigma}{\Delta\sigma}\left(1 + \sum_{n=1}^{\infty} b_n \int_{-a}^{a} dx \frac{f_n(x)}{\pi\sqrt{a^2-x^2}}\right). \tag{5.15}$$

At a resonance, r^{-1} diverges for zero damping and has a maximum for finite damping, so the reflectance has a local minimum, similar to the case of wide junctions. The main difference from a wide junction comes from the capacitive coupling, which gives rise to the non-resonant terms in Eq. (5.15). These terms vary smoothly with σ and a, giving the resonances a Fano shape and causing r to deviate significantly from the F-P formula, as shown in Figure 5.3c and d. Another distinction from wide junctions is that even and odd resonances become very different in strength. Numerical results show that the odd resonances are more narrow and less strong than the even ones. This can be seen more clearly in the analytical formula for r calculated under the approximation

$$f_n = \cos(q_n x - \frac{1}{2}n\pi), \quad q_n = \frac{\pi}{2a}\left(n+\frac{1}{4}\right). \tag{5.16}$$

In this case the odd modes disappear completely due to the integral in (5.14) and (5.15) yielding $b_n = 0$ for an odd $f_n(x)$. The residual presence of the odd modes in the numerical results is due to the incident field $E_i = E_0 e^{iq_0 x}$ containing both even and odd components. Except for this difference, the analytical approximation agrees well with the numerical result, and the agreement gets better at larger n where the approximation for f_n becomes more accurate.

Equivalent circuits are again helpful in understanding Eq. (5.15). Consider the impedance of a $2a$-long piece of plasmonic waveguide $Z = -4iZ_0 \tan qa$,[MDP48] which can be expanded

into the resonant form $Z^{-1} = (-4iZ_0qa)^{-1} + i\frac{qa}{2Z_0}\sum_n (q^2a^2 - \pi^2n^2)^{-1}$. This inverse impedance has the same structure as a network of parallel *LC* circuits (Figure 5.2c), which is in fact how resonant cavities are conventionally described. The impedance of the *LC* network is $Z^{-1} = (-i\omega L_0)^{-1} + i\omega\sum_n(\omega^2 L_n - C_n^{-1})^{-1}$, which yields $L_0 = 2ia/\sigma\omega$, $L_n = \frac{1}{2}L_0$, and $C_n = qa\kappa/2\pi^3n^2$. The load impedance of the junction, including the capacitor, the *LC* network and the right half-plane is then $Z_L = ZZ_c/(Z+Z_c) + Z_0$, which yields the reflection coefficient

$$r^{-1} = 1 - i\frac{2\pi}{\kappa}C + \frac{i}{2qa} + iqa\sum_{n=1}^{\infty}\frac{1}{q^2a^2 - \pi^2n^2}\,. \tag{5.17}$$

The first three terms represent the low q or dc response and match those in Eq. (5.15) (under the limit $\sigma \ll \sigma_0$). The strength of the resonances does not match Eq. (5.15) but this is expected, as the effect of capacitive coupling on the resonant modes was not accounted for. To correctly describe the junction L_n and C_n will have to be modified, which is effectively accomplished by Eq. (5.14). The resonant condition $qa = \pi n$ is however qualitatively correct (apart from the phase shift θ) and predicts the disappearance of the odd modes.

As in the case of wide junctions, the high-order resonances quickly diminish in the presence of plasmonic damping (Figure 5.3d), so these resonances are not robust enough for plasmonics applications. However, there is a prominent peak in the reflectance that occurs apart from the resonances and is persistent under damping, making it a promising candidate for controlling a compact plasmonic switch. The physical origin of this peak is as follows. If the field in the junction is approximately uniform, the junction reduces to a parallel *LC* circuit (Figure 5.2d) with impedances $Z_C = i/\omega C$ and $Z_{L_0} = -i\omega L_0 = \frac{2a}{\sigma}$, the later resembling the usual dc formula for resistance. When the inverse impedances, i.e., admittances Z_C^{-1} and $Z_{L_0}^{-1}$ cancel, the total impedance of the junction diverges and the reflectance is unity. We call this situation the

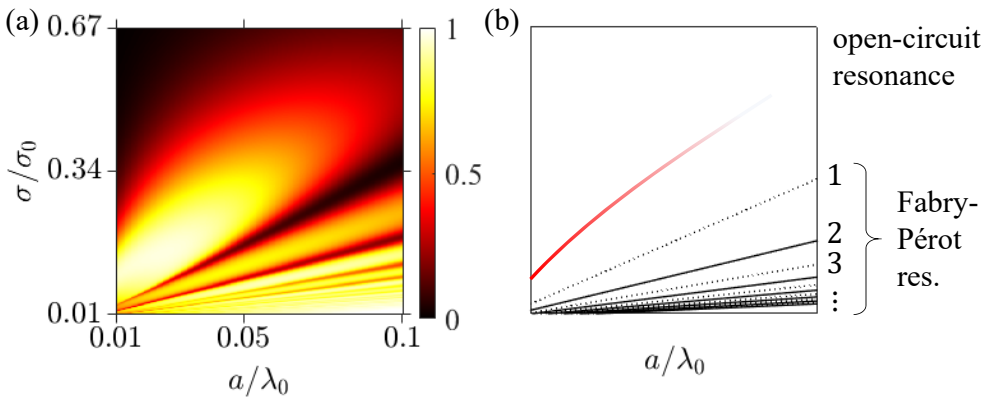

Figure 5.5: (a) False color plot of the reflectance of the junction for $\gamma = 0.05$. (b) Schematic diagram of the location of the open-circuit and the F-P resonances. The red curve is predicted by Eq. (5.18).

open-circuit resonance. It occurs when

$$\frac{\sigma}{\sigma_0 - \sigma}\frac{\lambda_0}{a} = 4\left(\log\frac{2}{q_0 a} - c\right). \tag{5.18}$$

The open-circuit resonance arises only for narrow junctions because it requires capacitive coupling of the leads. As the junction width increases, the capacitive coupling gradually weakens, and the junction evolves towards a regular F-P resonator, as shown in Figure 5.5.

5.4 Power absorption

. Low-conductivity junctions based on narrow slots in metallic films or 2DEGs under split-gates have also been proposed as basic units of terahertz detectors. They function by converting photons into plasmons, which are then detected as a dc current either through nonlinear optical effects[DS96, KDC$^+$09] or thermoelectric effects.[TMC$^+$15, LGW$^+$17, WPD$^+$17] The thermocurrent generated by the plasmons is proportional to the total power P absorbed by the system. Here we show that P is related to the plasmon reflectance R of the junction, so that the analytical formula (5.15) can be used to predict the detection efficiency. The photon to

plasmon conversion can be understood from our formulas above since the external terahertz field effectively replaces the field of the incident plasmon wave. For a narrow gap both fields are approximately constant over the junction, so that they launch plasmons from the junction with the same efficiency.

To quantify the extra absorbed power when the junction is present compared to a uniform 2DEG sheet, we define the dimensionless excess power absorption,

$$\Delta\bar{P} = \frac{\int_{-\infty}^{\infty} dx\,(P(x) - P_0)}{\lambda_0 \cdot P_{\text{inc}}} \tag{5.19}$$

where

$$P(x) = \text{Re}\,\sigma(x)|E(x)|^2 \tag{5.20}$$

is the local power absorption per unit area, $P_0 = \text{Re}\,\sigma_0|E_0|^2$ is the same quantity for a uniform system without a junction, and $P_{\text{inc}} = \frac{c}{4\pi}|E_0|^2$ is the incident power per unit area. We further divide $\Delta\bar{P}$ into two parts, absorption by the junction ($\Delta\bar{P}_j$) and by the leads ($\Delta\bar{P}_l$). The former is determined by the junction conductivity as well as the strength of the resonant fields, while the latter is determined by the strength of plasmon emission. The power absorption of the junction has a peak at each resonance where the resonant field is strong. It is also peaked at the open-circuit resonance, where the field in the junction is approximately constant and proportional to r, $E \simeq V/2a \propto r$, so that the peak in R is mirrored in $\Delta\bar{P}_j$. The power absorbed by the leads is determined by the scattered plasmon field which quickly approach $-re^{iq_0|x|}$ away from the junction. Hence $\Delta\bar{P}_l$ is approximately proportional to the reflectance. As shown in Figure 5.6a, $\Delta\bar{P}_l$ is very similar to the reflectance shown in Figure 5.3d. The resonances occur at the same locations but the odd modes disappear, due to the incident field being even in x.

A similar resonant structure is found when the frequency ω instead of the junction conductivity σ is varied, as shown in Figure 5.6b. Varying ω changes the background wavelength λ_0 and thus the ratio a/λ_0. The enhanced $\Delta\bar{P}_j$ at low frequencies is due to our assumption of a

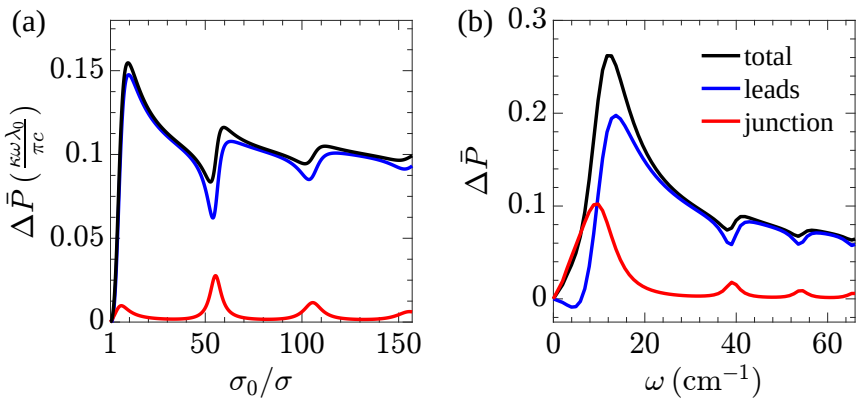

Figure 5.6: (a) Excess power absoption $\Delta\bar{P}$ for $a = 0.01\lambda_0$ and $\gamma = 0.05$. The power absorbed by the leads (blue) is proportional to the reflectance R. The power absorbed by the junction (red) has a maximum at the open-circuit and the F-P resonances. (b) Excess power absoption spectrum for a junction of width $2a = 1000\,\mathrm{nm}$ in graphene on hBN. The background conductivity is assumed to be Drude like, $\sigma_0 = \frac{iD_0}{\pi(\omega+i\nu)}$ with $\nu = 3\,\mathrm{cm}^{-1}$. The Drude weight $D_0 = \frac{e^2}{\pi\hbar^2}\mu$ is calculated at chemical potential $\mu = 0.12\mathrm{eV}$, the conductivity ratio is $D/D_0 = 0.03$. At low frequencies $\omega \sim \nu$ the conductivity is dominated by damping, resulting in a negative $\Delta\bar{P}_l$, i.e., the absorption in the leads is smaller than the background value.

Drude-like conductivity, $\sigma = iD/\pi(\omega + i\nu)$ where the damping rate ν is taken to be constant, so that $\mathrm{Re}\,\sigma \propto \frac{\nu}{\omega^2+\nu^2}$ has a maximum at $\omega = 0$. The Drude weight D is assumed to be proportional to the carrier density and independent of frequency. At low frequencies using the approximations $E \simeq V/2a$ and $E_s \simeq -re^{iq_0|x|}$, we obtain

$$\Delta\bar{P}_j \simeq \frac{2}{\pi\kappa g^2}\frac{D_0^2}{ca}\frac{\nu}{(\omega^2+\nu^2)^2}R\,, \quad \Delta\bar{P}_l \simeq \frac{4}{\pi}\frac{D_0}{c}\frac{\omega}{(\omega^2+\nu^2)}R\,, \tag{5.21}$$

where c is the speed of light and $g \equiv D/D_0 - 1$. Using these expressions, the optimal operating frequency can be found given the device parameters.

5.5 Discussions

Besides electrostatically gated graphene or 2DEGs, tunable conductivity profile needed for the plasmon-reflecting 1D junction can be also realized with layered high-T_c superconductors

by varying the temperature. Plasmon reflection in such a system would also yield information about local optical conductivity, which can be difficult to measure by other means. If the conductivity contrast is large, which is the case of our primary interest, the system becomes a weak link, i.e., a Josephson junction. Such a junction can be fabricated by lowering the T_c locally using focused ion beams or by etching and reducing the number of layers.[CCW+15] An estimate of the required parameters are as follows. Assume a BCS-like superconductor film has a thickness $d = 10\,\mathrm{nm}$, interlayer Josephson plasmon frequency $\omega_c = 50\,\mathrm{cm}^{-1}$, lattice dielectric constant $\varepsilon_\infty = 27$, anisotropy $\gamma = 17.5$, and is placed on a strontium titanate (STO) substrate. The low-frequency plasmon dispersion is then $\omega \simeq \gamma\omega_c\sqrt{(1+2\kappa/\varepsilon_\infty qd)^{-1}}$.[SWJ+14] At a frequency $\omega = 0.25\,\mathrm{THz}$, the effective dielectric constant of STO is $\kappa \simeq 1000$, and so the plasmon wavelength $\lambda_0 \simeq 10\,\mu\mathrm{m}$. If the conductivity inside the central region is $\sigma = 0.1\sigma_0$, then the strongest reflection/optimal THz detection occurs for a junction of width $2a \sim 200\,\mathrm{nm}$, which is quite practical.

We presented a theory of plasmonic interaction with a conductivity dip and showed that a narrow junction is a minimalistic yet robust plasmonic switch. Our analytical results and the associated physical insights into the plasmonic interaction with a 1D defect may be useful for the design of nanoscale plasmonic reflectors as well as terahertz detectors.

5.6 Fabry-Pérot formula for wide junctions

At the interface $x = 0$ of two conductive sheets with conductivities σ_1 and σ_2, a normally incident plasmon plane wave from sheet 1 has the following reflection and transmission coefficients[RK15]

$$r_{12} = e^{i\theta_{12}}\frac{\sigma_1-\sigma_2}{\sigma_1+\sigma_2}, \quad t_{12} = \frac{2\sqrt{\sigma_1\sigma_2}}{\sigma_1+\sigma_2}, \quad \theta_{12} = \frac{\pi}{4} - \frac{2}{\pi}\int_0^\infty du\,\frac{\tan^{-1}(\frac{\sigma_2}{\sigma_1}u)}{u^2+1}. \tag{5.22}$$

They have the properties $t_{12} = t_{21}$ and $\theta_{12} = -\theta_{21}$. Our junction has two interfaces between conductivities σ_0 and σ. Assigning $\sigma_1 = \sigma_0$ and $\sigma_2 = \sigma$, the total reflection and transmission from the left edge of the junction can be found by considering the multiple reflections within the junction,

$$r = r_{12} + \left(t_{12} r_{21} e^{i2\phi} t_{21} + t_{12} r_{21}^3 e^{i4\phi} t_{21} + ... \right) = r_{12} + \frac{t_{12} r_{21} e^{i2\phi} t_{21}}{1 - r_{21}^2 e^{i2\phi}}, \tag{5.23}$$

$$t = t_{12} e^{i\phi} t_{21} + t_{12} r_{21}^2 e^{i3\phi} t_{21} + ... = \frac{t_{12} e^{i\phi} t_{21}}{1 - r_{21}^2 e^{i2\phi}}, \tag{5.24}$$

where $\phi = 2qa$ is the phase the plasmon wave accumulates for crossing the junction once. The F-P formula

$$r_{\mathrm{FP}}^{-1} = \frac{\sigma_0^2 + \sigma^2}{\sigma_0^2 - \sigma^2} + i \frac{2\sigma_0 \sigma}{\sigma_0^2 - \sigma^2} \cot(\phi + \theta) \tag{5.25}$$

can be obtained by substituting (5.22) into (5.23). Note that the formula above was derived assuming the left interface to be the origin $x = 0$. If the origin is set at the center of the junction instead, which sits at a distance a from the left interface, the reflection and transmission coefficients (5.23), (5.24) and (5.25) will acquire an additional phase factor $e^{-i\phi}$.

5.7 Definition of impedances in the equivalent circuits

There is always some arbitrariness [MDP48] in the value of the characteristic impedance $Z_0 = v/i$ of a waveguide, which derives from the freedom of choosing the definition of "voltage" v and "current" i of a propagating wave. The only constraint is that the product of the two is proportional to the power P transmitted by such a wave, more precisely, that

$$P = \frac{1}{2} \mathrm{Re}\left(v^* i \right). \tag{5.26}$$

In our case the propagating wave is a plasmon. The plasmon travelling in the x-direction can be described by the field amplitudes

$$E_x = e^{iq_0 x} e^{-q_0|z|}, \quad E_z = iE_x, \quad H_y = \frac{2\pi}{c}\sigma_0 E_x, \tag{5.27}$$

which decay exponentially away from the plane, see Fig. 5.7(a). The power transmitted by the plasmon (per unit length in y) is given by the integrated Poynting vector,

$$P = \frac{1}{2}\mathrm{Re}\int_{-\infty}^{\infty} dz \frac{c}{4\pi} H_y^* E_z = \frac{1}{4}\mathrm{Re}(j^*\phi). \tag{5.28}$$

Here ϕ is the electric potential on the plane,

$$\phi = \int_0^{\infty} dz E_z, \tag{5.29}$$

and $j = \sigma_0 E_x$ is the current density. In this work we define our effective current to be equal to the physical current, $i = j$. This choice leads to a simple form of the current conservation in our equivalent circuits and fixes the effective voltage v to be one-half of the physical potential:

$$v = \frac{1}{2}\phi, \quad i = j. \tag{5.30}$$

The characteristic impedance of the sheet is then

$$Z_0 = \frac{1}{2}\frac{\phi}{j} = \frac{\pi}{\omega\kappa}. \tag{5.31}$$

One further application of this result is to a wide but finite strip. If this strip has the unperturbed sheet conductivity $\sigma = \sigma_0$ and the total length $2a \gg q_0^{-1}$ in the x-direction, then it is equivalent to

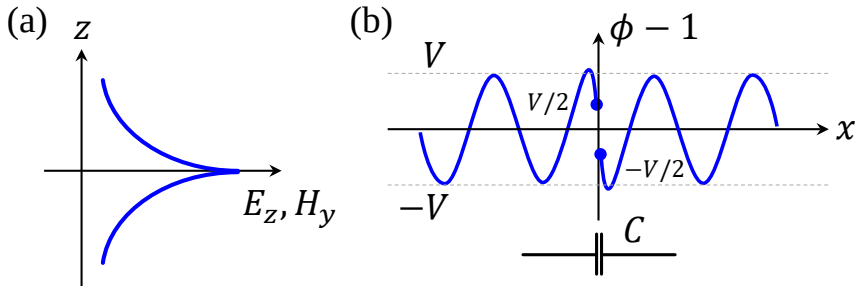

Figure 5.7: (a) Distribution of E_z and H_y outside the sheet. (b) Schematic diagram of the potential distribution for the case of a narrow vacuum gap. The large distance oscillations have amplitude V, while the potential right next to the gap oscillates with amplitude $V/2$.

a T-junction shown in Fig. 1(b) of the main text. The horizontal legs have the impedance [MDP48]

$$Z_1 = -iZ_0 \tan q_0 a \tag{5.32}$$

each and the vertical leg has the impedance

$$Z_2 = iZ_0 \csc q_0 a\,. \tag{5.33}$$

However, using Eqs. (5.32) and (5.33) for a narrow strip $2a \ll q_0^{-1}$ would be wrong because in that case the electric and magnetic fields have a much more complicated distribution than what is sketched in Fig. 5.7(a). Nevertheless, a simple result is obtained for a narrow vacuum gap, $\sigma = 0$, whose interaction with the plasmons is governed by the capacitive coupling of the two leads. As shown in the main text [see also Eq. (5.71) below], the correct plasmon reflection coefficient is obtained if we assume that the impedance of this type of junction is given by the usual circuit formula for the capacitor,

$$Z_C = \frac{1}{-i\omega C}\,. \tag{5.34}$$

Notably, it has no additional factor of one-half, which one would naively guess based on Eq. (5.30).

The reason for its absence is the difference between the "near-field" potential V and the "far-field" potential ϕ. Potential ϕ in our definiton of the impedance refers to the *asymptotic* amplitude of the plasmon waves. Our formula for the the reflection coefficient [Eq. (5.70) below] indicates that for the incident wave $\phi_i = e^{iq_0 x}$, the scattered potential at large distances oscillates with an amplitude V. On the other hand, since V is also the potential difference across the gap, its amplitude just outside the gap oscillates with an amplitude $V/2$, as shown in Fig. 5.7(b). This factor-of-two difference between the far-field and near-field amplitudes cancels the factor of one-half in the definition of the effective voltage $v = \phi/2$, leading to the standard form of Eq. (5.34).

The same cancellation occurs also if the short strip has a nonzero sheet conductivity σ. In other words, the correct result for the reflection coefficient is obtained if impedances Z_1 and Z_2 of the T-junction are multiplied by the factor of two. As far as the vertical leg is concerned, this modification has virtually no effect because it already has a large impedance $Z_2 \gg Z_1$, and so can be cut from the circuit. But the two horizontal legs now combine to the total impedance of $Z \simeq 4Z_1 = -4iZ_0 \tan qa$, as mentioned in the main text. The last expression can be further approximated by

$$Z \simeq -4iZ_0 qa = -4i \frac{\pi}{\omega\kappa} \frac{i\omega\kappa}{2\pi\sigma} a = \frac{2a}{\sigma}, \qquad qa \ll 1, \tag{5.35}$$

which is the usual formula for dc resistance of a conductor, similar to Eq. (5.34) being the standard formula for a capacitor.

5.8 Reflection coefficient of general 1D inhomogeneities

5.8.1 Plasmon equation

To make the paper self-contained, we begin this section with a derivation of the plasmon reflection coefficient for a 1D inhomogeneity. The quantity of interest is the electric potential ϕ on the sheet. Within the quasistatic approximation, which is valid when all distances involved are

much smaller than the radian length c/ω, the potential is given by the Coulomb law,

$$\phi(\mathbf{r}) = \phi_{\text{ext}}(\mathbf{r}) + (V * \delta\rho)(\mathbf{r}), \quad \mathbf{r} = (x, y). \tag{5.36}$$

Here and below the time dependence enters via the factor $e^{-i\omega t}$, which is implicit. An external potential ϕ_{ext} induces a charge distribution $\delta\rho$ on the sheet, which in turn creates a potential $V * \delta\rho$. Here $V = 1/\kappa r$ is the Coulomb kernel, κ is the dielectric constant of the environment exterior to the sheet, and $*$ denotes convolution,

$$(A * B)(\mathbf{r}) \equiv \int d^2 r' A(\mathbf{r} - \mathbf{r}') B(\mathbf{r}'). \tag{5.37}$$

Using the continuity equation $\partial_t \delta\rho + \nabla \cdot \mathbf{j} = 0$, which is equivalent to

$$-i\omega\delta\rho + \nabla \cdot (\sigma \mathbf{E}) = 0, \tag{5.38}$$

where $\sigma = \sigma(\mathbf{r})$ is the local sheet conductivity, we recast Eq. (5.36) into

$$\phi(\mathbf{r}) = \phi_{\text{ext}}(\mathbf{r}) - V(\mathbf{r}) * \nabla \cdot \left(\frac{\sigma(\mathbf{r})}{i\omega} \nabla\phi(\mathbf{r}) \right). \tag{5.39}$$

This is the principal equation governing the propagation of plasmons in a sheet. In general, it has to be solved numerically.

For the case of a uniform sheet with conductivity $\sigma(\mathbf{r}) = \sigma_0$, Eq. (5.39) can be solved analytically in the momentum space,

$$\tilde{\phi}(\mathbf{k}) = \tilde{\phi}_{\text{ext}}(\mathbf{k}) + \tilde{V}(\mathbf{k}) \cdot q^2 \frac{\sigma}{i\omega} \tilde{\phi}(\mathbf{k}), \tag{5.40}$$

where $\tilde{V}(\mathbf{k}) = 2\pi/\kappa|\mathbf{k}|$ and the Fourier transform is defined as

$$\tilde{f}(\mathbf{k}) = \int d\mathbf{r}\, f(\mathbf{r}) e^{-i\mathbf{k}\cdot\mathbf{r}}, \quad f(\mathbf{r}) = \int \frac{d\mathbf{k}}{(2\pi)^2}\, \tilde{f}(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{r}}. \tag{5.41}$$

Eq. (5.40) has the form

$$\tilde{\phi}(\mathbf{k}) = \frac{\tilde{\phi}_{\text{ext}}(\mathbf{k})}{\varepsilon(\mathbf{k})}, \tag{5.42}$$

where the dielectric function is define by

$$\varepsilon(\mathbf{k}) = 1 - \frac{|\mathbf{k}|}{q_0}. \tag{5.43}$$

The plasmon momentum q_0 is found by setting ε to zero:

$$q_0 = \frac{i\kappa\omega}{2\pi\sigma_0}. \tag{5.44}$$

This quantity is complex when a dissipation is present in the system, e.g., when the dielectric constant of the environment κ has a nonzero imaginary part, or when the conductivity σ_0 has a nonzero real part. The imaginary part $\operatorname{Im} q_0 > 0$ has the physical meaning of the inverse propagation length. For convenience, we parametrize q_0 as

$$q_0 = \frac{2\pi}{\lambda_0}(1 + i\gamma), \tag{5.45}$$

where $\lambda_0 = 2\pi/\operatorname{Re} q_0$ is the plasmon wavelength and $\gamma = \operatorname{Im} q_0/\operatorname{Re} q_0$ is the dimensionless damping. In the absence of the external potential, one can find (unbounded) solutions $\phi = e^{iq_x x + iq_y y}$ with real q_y and complex $q_x = \sqrt{q_0^2 - q_y^2}$, $\operatorname{Im} q_x > 0$, which can be thought of as decaying plane waves that are incident from the far left at some oblique angle.

5.8.2 1D inhomogeneities

In our problem the plasmon wave $\phi_i = e^{iq_x x + iq_y y}$ impinges upon a 1D inhomogeneity in the conductivity of the sheet localized around the y-axis. We parametrize the inhomogeneity as

$$\sigma(x) = \sigma_0 \left[1 + g(x)\right] \tag{5.46}$$

or

$$g(x) = \frac{\Delta\sigma(x)}{\sigma_0}, \quad \Delta\sigma(x) \equiv \sigma(x) - \sigma_0 . \tag{5.47}$$

As the conductivity and the plasmon wavevector are inversely proportional to each other [Eq. (5.44)], the inhomogeneity can also be parametrized in terms of the plasmon momentum,

$$\frac{1}{q(x)} = \frac{1 + g(x)}{q_0}, \quad q_0 \equiv q(\infty) . \tag{5.48}$$

A spatial variation in the momentum scatters an incident plasmon wave ϕ_i, so that the solution ϕ to the plasmon equation (5.39) contains both the incident ϕ_i and the scattered (reflected plus transmitted) waves ψ. Setting $\phi_{\text{ext}}(\mathbf{r}) \to 0$ and $\phi(\mathbf{r}) \to \phi(x)e^{iq_y y}$ in (5.39), where the $e^{iq_y y}$ factor is retained by virtue of translational invariance in the y-direction, we obtain the 1D version of the plasmon equation,

$$\phi(x) = V_1 * \left(\frac{1 + g(x)}{q_0} q_y^2 \phi(x) - \partial_x \frac{1 + g(x)}{q_0} \partial_x \phi(x) \right) . \tag{5.49}$$

Here $V_1(x)$ is the 1D Coulomb kernel divided by 2π,

$$V_1(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dy \frac{e^{iq_y y}}{\sqrt{x^2 + y^2}} = \frac{K_0(|q_y x|)}{\pi} , \tag{5.50}$$

where $K_0(z)$ is the modified Bessel function of the second kind. Note that $\tilde{V}_1(\mathbf{k}) = 1/|\mathbf{k}|$.

To find the expression for the scattered wave ψ we again go to the momentum space,

where Eq. (5.49) becomes

$$\left(\tilde{\phi}_i + \tilde{\psi}\right)\varepsilon = \frac{1}{q_0}\tilde{V}_1\left(q_y^2(\widetilde{g\phi}) - (\widetilde{\partial_x g \partial_x \phi})\right). \tag{5.51}$$

The incident field ϕ_i is a solution of the homogeneous equation (5.42), so $\tilde{\phi}_i\varepsilon = 0$, while the inverse dielectric function ε^{-1} has the meaning of the Fourier transformed Green's function, $\tilde{G} = \varepsilon^{-1}$. The equation for the scattered wave in real space is then

$$\psi(x) = \frac{1}{q_0}(G * V_1) * \left[q_y^2 g(x)\phi(x) - \partial_x g(x)\partial_x \phi(x)\right]. \tag{5.52}$$

where the Green's function is

$$G(x, q_y) = \int_{-\infty}^{\infty} \frac{dk}{2\pi} e^{ikx} \varepsilon^{-1}\left(\sqrt{k^2 + q_y^2}\right). \tag{5.53}$$

The integrand of the Green's function has two poles at $k = \pm q_x$ and two branch cuts on the imaginary axis, one from iq_y to $i\infty$ and the other from $-iq_y$ to $-i\infty$. Using contour integration techniques, we find

$$G(x, q_y) = -i\frac{q_0^2}{q_x} e^{iq_x|x|} - \frac{q_0}{\pi}\int_0^{\infty} \frac{dt}{\sqrt{t^2 + q_y^2}} \frac{t^2}{t^2 + q_0^2} e^{-\sqrt{t^2+q_y^2}|x|}. \tag{5.54}$$

5.8.3 Normal incidence

For normally incident waves $q_y = 0$ and $q_x = q_0$, the incident wave becomes $\phi_i = e^{iq_0 x}$, while the 1D plasmon equation (5.52) is reduced to

$$\psi(x) = \frac{1}{q_0}(G_1 * V_1) * [-\partial_x g(x)\partial_x \phi(x)] \,, \tag{5.55}$$

where $G_1(x) = G(x,0)$ is found to be

$$\begin{aligned} G_1(x) &= -iq_0 e^{iq_0|x|} + \frac{q_0}{\pi}\left\{ \mathrm{Ci}(q_0|x|)\cos(q_0|x|) + \left[\mathrm{Si}(q_0|x|) - \frac{\pi}{2}\right]\sin(q_0|x|)\right\} \\ &= -iq_0 e^{iq_0|x|} - \frac{q_0}{2\pi}\left[e^{iq_0|x|}E_1(iq_0|x|) + e^{-iq_0|x|}E_1(-iq_0|x|)\right] . \end{aligned} \tag{5.56}$$

Here $\mathrm{Ci}(z)$, $\mathrm{Si}(z)$ and $E_1(z)$ are the cosine, sine and exponential integrals,

$$\mathrm{Ci}(z) = -\int_z^\infty dt\,\frac{\cos t}{t}\,, \quad \mathrm{Si}(z) = \int_0^z dt\,\frac{\sin t}{t}\,, \quad E_1(z) = \int_z^\infty dt\,\frac{e^{-t}}{t}\,, \tag{5.57}$$

with the branch cut taken on the negative real axis. Equation (5.55) can be further simplified in momentum space in terms of the total electric field $E = -\partial_x \phi$ and the scattered field $E_s = -\partial_x \psi$,

$$\tilde{E}_s(k) = -ik\tilde{\psi}(k) = \left(\frac{1}{\varepsilon(k)} - 1\right)(\widetilde{gE})(k)\,, \quad (\widetilde{gE}) = \frac{1}{2\pi}\tilde{g} * \tilde{E}\,. \tag{5.58}$$

In the real space this reads,

$$E_s(x) = \int_{-\infty}^{\infty} dx' \left(G_1(x-x') - \delta(x-x')\right) g(x')E(x')\,. \tag{5.59}$$

This is the equation for the scattered field used in the main text. It indicates the scattered field E_s can be determined by the conductivity profile $g(x)$ and the local field $E(x)$ around the inhomogeneity using the Green's function G_1.

The reflection coefficient r of the normally incident plasmon wave can be obtained by

analyzing the scattered field E_s at large negative x, $E_s \simeq -rE_0 e^{-iq_0 x}$, where E_0 is the amplitude of the incident field $E_i = E_0 e^{iq_0 x}$. From Eq. (5.58), the long range behavior of E_s is determined by the pole of $\varepsilon^{-1}(k)$ at $k = -q_0$. Taking the residue, we get

$$r = \frac{iq_0}{E_0} \int_{-\infty}^{\infty} dx\, g(x) E(x) e^{iq_0 x}\,. \tag{5.60}$$

This formula is exact and can be used for any conductivity profile $g(x)$. The field $E(x)$ can be calculated numerically, cf. Sec. 5.11

An analytical expression for $E(x)$ (and thus r) can be found in the perturbative case, $r \ll 1$, where the current density is approximately constant across the inhomogeneity,

$$j = \sigma_0 E_0 = \sigma(x) E(x)\,, \tag{5.61}$$

so that $E(x) = \frac{\sigma_0}{\sigma(x)} E_0$ and

$$r = iq_0 \int_{-\infty}^{\infty} dx\, \frac{\sigma(x) - \sigma_0}{\sigma(x)} e^{iq_0 x} = iq_0 \int_{-\infty}^{\infty} dx\, \frac{g(x)}{g(x)+1} e^{iq_0 x}\,. \tag{5.62}$$

5.9 Reflection coefficient of narrow junctions

In this section we consider a particular type of conductivity profile,

$$g(x) = g\Theta(a - |x|)\,, \quad g = \frac{\Delta\sigma}{\sigma_0} = \frac{\sigma - \sigma_0}{\sigma_0}\,, \tag{5.63}$$

which describes a junction that has a constant conductivity σ and a width $2a$. Further, we assume the width is narrow compared to the background plasmon wavelength, $a \ll q_0^{-1}$. In this limit, the leads can be considered perfect metals of effectively infinite conductivity. This allows us to make simplifications and obtain analytical expressions for the reflection coefficient r.

5.9.1 Vacuum gap

Analytical solution of the field distribution $E_{\rm vac}$ for the special case of a vacuum gap, $\sigma = 0$ or $g = -1$, is well known,[Smy50]

$$E_{\rm vac}(x) = VF(x)\,, \quad F(x) = \frac{1}{\pi}\frac{\Theta(a-|x|)}{\sqrt{a^2-x^2}}\,, \tag{5.64}$$

where V is the voltage difference across the junction,

$$V = \int_{-a}^{a} dx E(x)\,. \tag{5.65}$$

Substituting $E_{\rm vac}(x)$ into Eq. (5.59), we get

$$\begin{aligned} E_i(x) \simeq E_0 &= V\int_{-a}^{a} dx' G_1(x-x')F(x) \\ &= V\frac{q_0}{\pi}\left[c + \log\left(\frac{q_0 a}{2}\right) - i\pi\right], \quad -a < x < a\,. \end{aligned} \tag{5.66}$$

Here we used the small-distance approximation for the Green's function,

$$G_1(x) \simeq \frac{q_0}{\pi}\left[c + \log\left(q_0|x|\right) - i\pi\right] + O(q_0|x|)\,, \quad q_0|x| \ll 1\,, \tag{5.67}$$

where $c \simeq 0.577$ is the Euler-Mascheroni constant, and the table integral

$$\int_{-a}^{a} dx' \frac{\log|x-x'|}{\pi\sqrt{a^2-x'^2}} = \log\frac{a}{2}\,. \tag{5.68}$$

(It can be evaluated by the change of variables $x' = a\cos\theta'$.) Identifying the capacitance of the junction (cf. Sec. 5.10)

$$C = \frac{\kappa}{2\pi^2}\log\frac{2e^{-c}}{q_0 a}\,, \tag{5.69}$$

and the reflection coefficient for a narrow junction [Eq. (5.60) with $e^{iq_0x} \to 1$]

$$r \simeq \frac{iq_0}{E_0} gV\,, \tag{5.70}$$

our Eq. (5.66) yields the reflection coefficient of a narrow vacuum gap,

$$r = \frac{i\pi}{\log\left(\frac{2}{q_0 a}\right) - c + i\pi} = \frac{i\kappa}{2\pi C + i\kappa}\,. \tag{5.71}$$

Note that as the gap gets wider, r approaches unity. This can be understood by considering an infinitely wide gap where the current $j(x)$ is completely reflected and is zero at the edge, so that the current reflection coefficient $r_j = -r = -1$.

5.9.2 General cases

We add E_i to both sides of Eq. (5.59) to get

$$E_i(x) \simeq E_0 = -\left[G(x) * gE(x)\right] + (1+g)E(x)\,, \quad -a < x < a\,. \tag{5.72}$$

We multiply this by $\sigma_0 F(x)$, where $F(x)$ is given by Eq. (5.64), and integrate it from $x = -a$ to a,

$$\sigma_0 E_0 = -\sigma_0 gV\frac{q_0}{\pi}\left[c + \log\left(\frac{q_0 a}{2}\right) - i\pi\right] + \sigma_0(1+g)\int_{-a}^{a} dx E(x)F(x)\,. \tag{5.73}$$

We arrive at the equation (mentioned in the main text)

$$\sigma_0 E_0(1-r) = -i\omega V C' + \int_{-a}^{a} dx\, \sigma E(x) F(x)\,, \quad C' \equiv -gC\,, \tag{5.74}$$

which represents the conservation of charge. Note that this equation has the correct limiting behavior, reproducing the reflection coefficient for the vacuum gap when $g = -1$ and $\sigma = 0$, and yielding $r = 0$ when $g = 0$ and $\sigma = \sigma_0$.

Equation (5.73) can be used to find the reflection coefficient r if the field $E(x)$ were known. We use the following ansatz for the field,

$$E(x) = d_0 + \sum_{n=1}^{\infty} d_n E_n(x)\,, \quad -a < x < a\,, \tag{5.75}$$

which is an expansion into the eigenmodes E_n of the system when there is no external field. The eigenmodes E_n have the following properties,

$$E_n(|x| > a) = 0\,, \quad V_n = \int_{-a}^{a} dx E_n(x) = 0\,, \tag{5.76}$$

as the system now consists of a junction with nonzero conductivity σ_n and two sheets of infinite conductivity. The eigenvalue equation governing $E_n(x)$ can be derived from (5.72) using the substitutions $g = -1$ and $1 + g = \sigma_n/\sigma_0$ and invoking (5.68) and (5.76),

$$E_n(x) = A - \frac{q_n}{\pi} \int_{-\infty}^{\infty} dx' \log \frac{|x - x'|}{L} E_n(x')\,, \tag{5.77}$$

where $q_n = i\kappa\omega/2\pi\sigma_n$, the constant L is arbitrary, and the constant A can be obtained by multiplying both sides of (5.77) by $F(x)$ and integrating over x,

$$A = \int_{-a}^{a} E_n(x) F(x)\,. \tag{5.78}$$

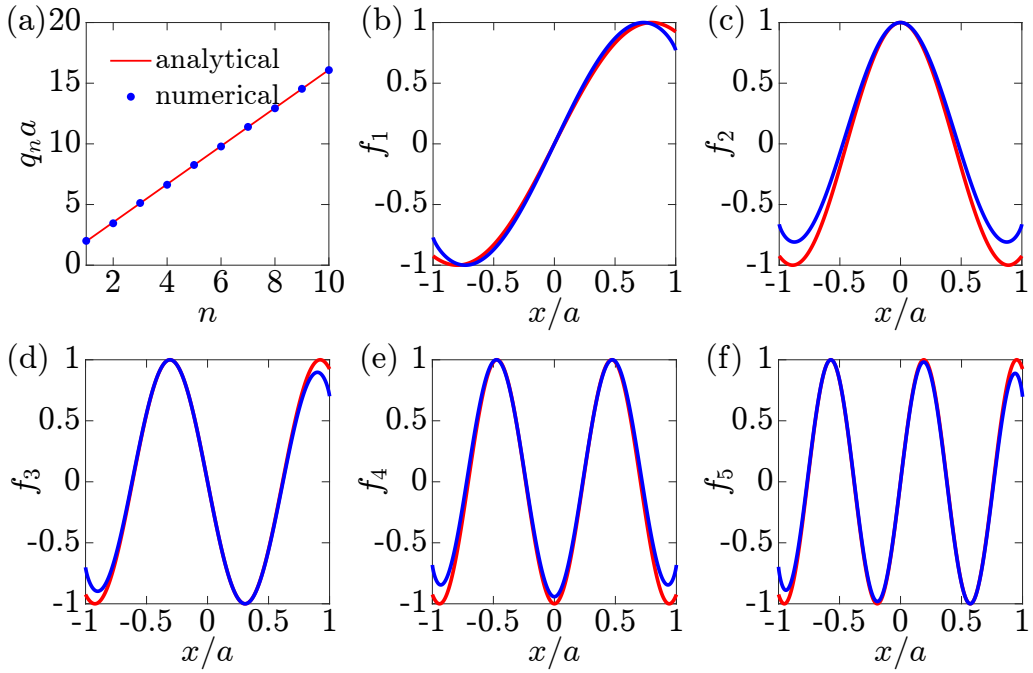

Figure 5.8: (a) Eigenvalues q_n and (b)-(f) the first five eigenmodes f_n of a junction in a perfect metal sheet. The analytical formula is (5.80), while the numerical results are obtained by solving (5.77).

Since the kernel in (5.77) is self-adjoint, the eigenmodes E_n are orthogonal,

$$\int_{-a}^{a} dx E_n(x) E_m(x) = 0\,, \quad n \neq m\,, \tag{5.79}$$

and the eigenvalues q_n are real. (5.77) can be solved numerically to obtain the eigenmodes. As shown in Fig. 5.8, they have the asymptotic forms

$$\frac{E_n}{E_0} \simeq \cos\left(q_n x - \frac{n\pi}{2}\right), \quad q_n \simeq \frac{\pi}{2a}\left(n + \frac{1}{4}\right), \tag{5.80}$$

as n is increased.

Next we determine the coefficients d_n. Under an external field the voltage across the

junction $V \neq 0$ but all $V_n = 0$. Thus V must be accounted for by d_0, i.e.,

$$d_0 = \frac{V}{2a}\,. \tag{5.81}$$

To find d_n, we rewrite (5.72) using (5.75) and (5.77) as

$$\log\frac{|x|}{L} * h(x) = 0\,, \tag{5.82}$$

with a solution

$$h(x) = V\left[\frac{1}{2a} - F(x)\right] + \sum_{n=1}^{\infty}\left(1 + \frac{g+1}{g}\frac{q_n}{q_0}\right) d_n E_n = 0\,. \tag{5.83}$$

Here (5.68) was used to convert constants into the convolution,

$$1 = \frac{1}{\log(a/2)}\,F(x) * \log\frac{|x|}{L}\,, \tag{5.84}$$

and (5.73) was used in the form

$$E_0 - (g+1)\int_{-a}^{a} dx E(x)F(x) = -g\frac{q_0}{\pi}\,V\log\frac{a}{2}\,. \tag{5.85}$$

Using the orthogonality of the eigenmodes, we find

$$d_n = \frac{V}{1 + \frac{g+1}{g}\frac{q_n}{q_0}}\,\frac{\int_{-a}^{a} dx E_n(x)F(x)}{\int_{-a}^{a} dx E_n^2(x)}\,. \tag{5.86}$$

Note that in the limit of a vacuum gap where $g = -1$, d_n reduces to coefficients of the expansion of $E_{\rm vac} = VF(x)$ in the E_n basis, as expected. In the manuscript, the dimensionless expansion coefficients b_0 and b_n are defined as

$$b_0 = \frac{d_0}{(V/2a)} = 1\,, \quad b_n = d_n\frac{E_0}{(V/2a)} = \frac{2a}{1 + \frac{g+1}{g}\frac{q_n}{q_0}}\,\frac{\int_{-a}^{a} dx f_n(x)F(x)}{\int_{-a}^{a} dx f_n^2(x)}\,, \tag{5.87}$$

and the field is expanded as

$$E(x) = \frac{V}{2a}\left(1 + \sum_{n=1}^{\infty} b_n f_n\right), \qquad f_n = \frac{E_n}{E_0}. \tag{5.88}$$

5.9.3 Analytic approximation

Having found the coefficients d_n, the expression for the reflection coefficient can now be written from (5.73) using (5.70),

$$r^{-1} = 1 - \frac{2\pi i}{\kappa}C - \frac{i}{q_0}\frac{g+1}{g}\left\{\frac{1}{2a} + \sum_{n=1}^{\infty}\frac{1}{1+\frac{g+1}{g}\frac{q_n}{q_0}}\frac{\left[\int_{-a}^{a} dx\, f_n(x)F(x)\right]^2}{\int_{-a}^{a} dx\, f_n^2(x)}\right\}. \tag{5.89}$$

An analytical expression for r can be found using the approximation (5.80), so that

$$\int_{-a}^{a} dx\, f_n^2(x) = a\left[1 + \frac{1}{\sqrt{2}\pi\left(n+\frac{1}{4}\right)}\right], \tag{5.90}$$

$$\begin{aligned}\left[\int_{-a}^{a} dx\, f_n(x)F(x)\right]^2 &= 0, \quad n \text{ odd} \\ &= J_0^2(q_n a) \simeq \frac{2+\sqrt{2}}{\pi^2}\frac{1}{n+\frac{1}{4}}, \quad n \gg 1 \text{ even},\end{aligned} \tag{5.91}$$

where $J_0(z)$ is the Bessel function of the first kind. The summation is then

$$\sum_{n=1}^{\infty}\frac{1}{1+\frac{g+1}{g}\frac{q_n}{q_0}}\frac{\left[\int_{-a}^{a} dx\, f_n(x)F(x)\right]^2}{\int_{-a}^{a} dx\, f_n^2(x)} \simeq \frac{2q_0}{\pi}\frac{g}{g+1}\frac{2+\sqrt{2}}{\pi^2}\sum_{n=1}^{\infty}\frac{1}{(2n+\alpha)(2n+\beta)}, \tag{5.92}$$

where $\alpha = \frac{2q_0 a}{\pi}\frac{g}{g+1} + \frac{1}{4}$ and $\beta = \frac{1}{\sqrt{2}\pi} + \frac{1}{4}$. Using the identity

$$\sum_{n=1}^{\infty}\frac{1}{(2n+\alpha)(2n+\beta)} = \frac{\Psi\left(1+\frac{\alpha}{2}\right) - \Psi\left(1+\frac{\beta}{2}\right)}{2(\alpha-\beta)}, \tag{5.93}$$

where $\Psi(z) = \Gamma'(z)/\Gamma(z)$ is the digamma function, the reflection coefficient is

$$r^{-1} \simeq 1 - \frac{2\pi i}{\kappa} C - \frac{i}{2q_0 a} \frac{g+1}{g} - i \frac{\sqrt{2}+1}{\pi^2} \frac{2(g+1)}{2\sqrt{2} q_0 a g - (g+1)} \left[\Psi\left(\frac{9}{8} + \frac{q_0 a}{\pi} \frac{g}{g+1} \right) - \Psi\left(\frac{9}{8} + \frac{\sqrt{2}}{4\pi} \right) \right], \tag{5.94}$$

in the limits $q_0 a \ll 1$ and $n \gg 1$.

5.10 Capacitance of a vacuum gap

The capacitance (5.69) found in the previous section can be simply derived by considering a vacuum gap in a perfect metal sheet. The induced charge distribution on the sheet is well-known,[Smy50]

$$\delta\rho(x) = -\frac{\kappa V}{2\pi^2} \frac{1}{\sqrt{x^2 - a^2}} \Theta(|x| - a) \operatorname{sgn}(x) . \tag{5.95}$$

The capacitance C is then

$$C = \frac{Q}{V} = \frac{1}{V} \int_{-L}^{-a} dx \, \delta\rho(x) = \frac{\kappa}{2\pi^2} \log\left(\frac{2L}{a} \right), \tag{5.96}$$

where L is the long-distance cutoff length. If σ_0 is finite, the charge distribution $\delta\rho(x)$ evolves into plasmonic waves at distance $\sim q_0^{-1}$. Therefore, we expect $L \sim q_0^{-1}$. From the expression for C in the previous section we get the exact relation $L = e^{-c} q_0^{-1}$.

5.11 Numerical calculation of the reflection coefficient

The reflection coefficient of a conductivity profile $g(x)$ of any shape and width can be calculated numerically by recasting Eq. (5.59) in the following form,

$$E = E_s + E_i = G * (gE) - gE + E_i \,, \tag{5.97}$$

or

$$E_i = \mathcal{K} E \,, \quad \mathcal{K} = 1 + g - G * g \,. \tag{5.98}$$

Inversion of $\mathcal{K}$ then yields the total field within the inhomogeneity $E = \mathcal{K}^{-1} E_i$ given the incident field E_i, and the reflection coefficient can be found using Eq. (5.60). The kernel $\mathcal{K}$ has elements

$$\mathcal{K}_{ij} = (1 + g(x_i))\,\delta_{ij} - G(x_i - x_j) g(x_j) \Delta x_j \,, \tag{5.99}$$

where x within the junction is discretized into x_i and the Green's function G is calculated using Eq. (5.56). For our calculation we chose the discretization

$$x_i = a \cos \frac{\pi}{N} \left(i - \frac{1}{2} \right) \,, \quad i = N, N-1, ..., 1 \,, \tag{5.100}$$

and

$$\Delta x_i = a \frac{\pi}{N} \sin \frac{\pi}{N} \left(i - \frac{1}{2} \right) \,, \tag{5.101}$$

with a number of grid points $N \sim 10^2$. The diverging diagonal elements $G(x_i - x_i)$ is regularized using $G(\Delta x_i / 2\pi)$, explained as follows. For small arguments, $G(x)$ is proportional to $\log|x|$. Assume we wish to calculate the integral $\int_{-L}^{L} dx \log|x|$ as a discrete sum on a uniform grid. The grid is taken to be $x_i = i\Delta x$ with $i = -N, ..., N$, which has $2N+1$ points separated by distance Δx and a total length $2L = (2N+1)\Delta x$. Writing L_0 as the regularized replacement for the diverging

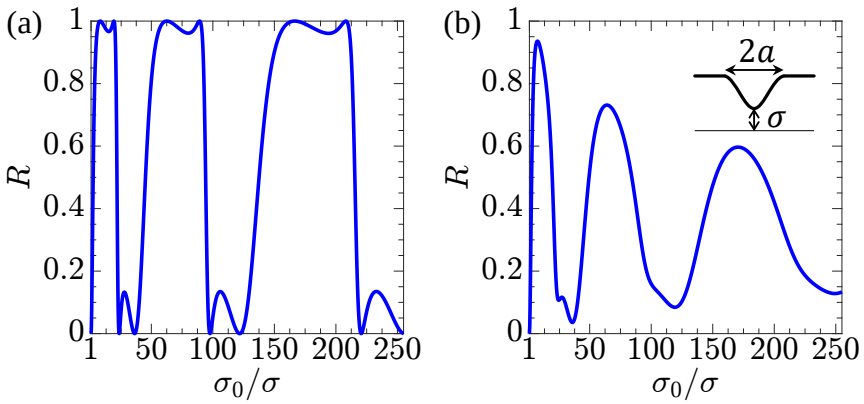

Figure 5.9: (a) Reflectance of a junction with a smooth conductivity profile for $a = 0.1\lambda_0$ and $\gamma = 0$. (b) Similar plot at $\gamma = 0.05$. (Inset) The conductivity profile σ_{sm}.

$\log 0$, we have

$$\int_{-L}^{L} dx \log|x| = \left(L_0 + \sum_{i \neq 0} \log|x_i| \right) \Delta x \,. \tag{5.102}$$

The summation can be simplified using Stirling's approximation for large N,

$$\sum_{i=1}^{N} \log|x_i| = \log\left(\Delta x^N N!\right) \simeq \log\left(\Delta x^N\right) + \log\left[\sqrt{2\pi N} \left(\frac{N}{e}\right)^N \right] . \tag{5.103}$$

After some algebra we find $L_0 = \log(\Delta x/2\pi)$, prescribing the regularization $G(\Delta x_i/2\pi)$.

5.12 Smooth conductivity profiles

In our model we assume sharp conductivity changes at the edges of the junction. This may be difficult to achieve in experiment, or yield unphysical results in theory. Here we show that the physics for a sharply-changing conductivity profile remains qualitatively the same for a smoothly-varying one. To demonstrate, we calculated numerically the reflectance of a smooth conductivity profile

$$\sigma_{sm}(|x| < a) = \sin^2\left(\frac{\pi}{2a}|x|\right) + \frac{\sigma}{\sigma_0}\cos^2\left(\frac{\pi}{2a}|x|\right) , \tag{5.104}$$

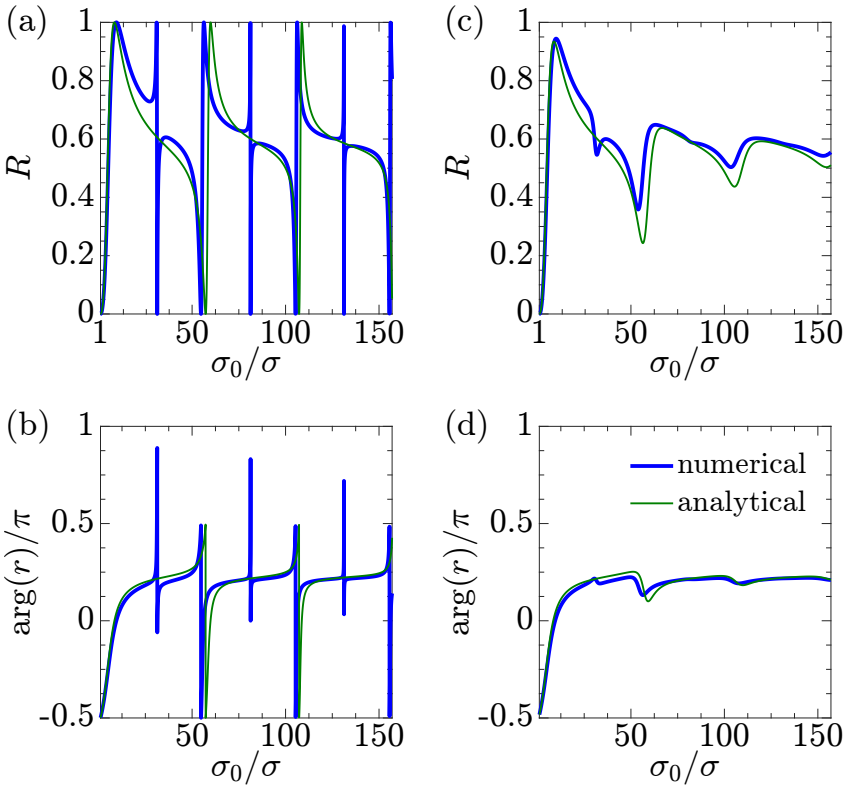

Figure 5.10: (a) Squared magnitude $R = |r|^2$ and (b) phase $\arg(r)$ of the reflection coefficient r for a junction of width $a = 0.01\lambda_0$ and damping $\gamma = 0$. (b) Similar quantities for $\gamma = 0.05$. The blue curves are calculated numerically as described in Sec. 5.11, while the green curves are from (5.94).

where σ now denotes the lowest value of conductivity occuring at $x = 0$. As shown in Figure 5.9, the reflectance retains the same features – capacitive open-circuit resonance and Fano-shaped cavity resonances with weak odd modes. The open-circuit resonance is still persistent under damping, while the cavity resonances are less sharp due to a larger $|t'|$ for smooth profiles. Note that to get an anti-resonance at the same location, $\sigma_0/\sigma \approx 10$, the junction width had to be increased by an order of magnitude, from $a/\lambda_0 = 0.01$ to 0.1.

5.13 Phase of the reflection coefficient

We show in Fig. 5.10 the phase of the reflection coefficient for a narrow junction $a = 0.01\lambda_0$. In the perturbative regime where $\sigma \lesssim \sigma_0$ the phase is $-\pi/2$, in agreement with Eq. (5.62). At the open-circuit resonance the phase is 0, $r > 0$ is real and $|r| \simeq 1$. For large conductivity contrasts the phase approaches the limiting value $\simeq 0.24\pi$ predicted by Eq. (5.71).

5.14 Effect of damping contrast

In the manuscript we only considered the case when the damping γ in the junction is the same as in the sheet, $\gamma = \gamma_0$. One might expect a stronger damping in the junction would further dampen the resonances and vice versa. This is indeed the case, as can be seen from the reflectances calculated under different values of γ shown in Fig. 5.11. The conductivities are parametrized as

$$\sigma_0 = \frac{S_0}{1 + i\gamma_0}, \quad \sigma = \frac{S}{1 + i\gamma}, \tag{5.105}$$

where $\gamma_0 = 0.05$ and γ takes the values 0.015, 0.05, and 0.15. The resonant features are enhanced if $\gamma < \gamma_0$ and diminished if $\gamma > \gamma_0$.

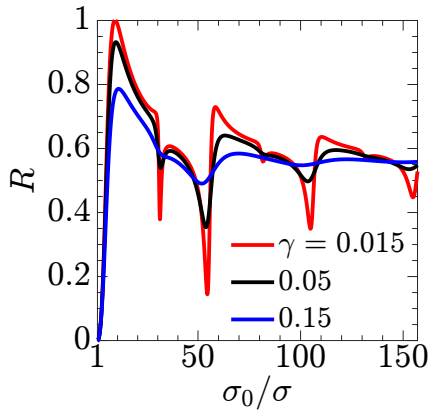

Figure 5.11: Reflectance of the junction when it has a damping γ different from or equal to the background value $\gamma_0 = 0.05$. Junction width $a = 0.01\lambda_0$.

Bibliography

[ABJR12] Joanna M. Atkin, Samuel Berweger, Andrew C. Jones, and Markus B. Raschke. Nano-optical imaging and spectroscopy of order, phases, and domains in complex solids. *Adv. Phys.*, 61(6):745–842, 2012.

[AGK09] Sergiu Amarie, Thomas Ganz, and Fritz Keilmann. Mid-infrared near-field spectroscopy. *Opt. Exp.*, 17(24):21794, 2009.

[AGS+06] G. Allison, E. A. Galaktionov, A. K. Savchenko, S. S. Safonov, M. M. Fogler, M. Y. Simmons, and D. A. Ritchie. Thermodynamic density of states of two-dimensional GaAs systems near the apparent metal-insulator transition. *Phys. Rev. Lett.*, 96(21):216407, 2006.

[AK11a] S. Amarie and F. Keilmann. Broadband-infrared assessment of phonon resonance in scattering-type near-field microscopy. *Phys. Rev. B*, 83(4):045404, Jan 2011.

[AK11b] S. Amarie and F. Keilmann. Erratum: Broadband-infrared assessment of phonon resonance in scattering-type near-field microscopy [Phys. Rev. B 83, 045404 (2011)]. *Phys. Rev. B*, 84:199904(E), 2011.

[AKSV99] M.S. Agranovich, B.Z. Katsenelenbaum, A.N. Sivov, and N.N. Voitovich. *Generalized method of eigenoscillations in diffraction theory*. Wiley-VCH, Berlin, 1999.

[AM76] N.W. Ashcroft and N.D. Mermin. *Solid State Physics*. Saunders College, Philadelphia, 1976.

[AM82] P. K. Aravind and Horia Metiu. Use of a perfectly conducting sphere to excite the plasmon of a flat surface. 1. Calculation of the local field with applications to surface-enhanced spectroscopy. *J. Phys. Chem.*, 86(26):5076–5084, 1982.

[AM83] P. K. Aravind and Horia Metiu. The effects of the interaction between resonances in the electromagnetic response of a sphere-plane structure; applications to surface enhanced spectroscopy. *Surf. Sci.*, 124(2-3):506 – 528, 1983.

[ATdA+08] Javier Aizpurua, Thomas Taubner, F. Javier García de Abajo, Markus Brehm, and Rainer Hillenbrand. Substrate-enhanced infrared near-field spectroscopy. *Opt. Express*, 16(3):1529–1545, 2008.

[ATH+13] J. S. Alden, A. W. Tsen, P. Y. Huang, R. Hovden, L. Brown, J. Park, D. A. Muller, and P. L. McEuen. Strain solitons and topological defects in bilayer graphene. *Proc. Natl Acad. Sci. USA*, 110(28):11256–11260, 2013.

[AZS02] Tsuneya Ando, Yisong Zheng, and Hidekatsu Suzuura. Dynamical Conductivity and Zero-Mode Anomaly in Honeycomb Lattices. *J. Phys. Soc. Jpn.*, 71(5):1318–1324, 2002.

[BDN+13] Benjamin Butz, Christian Dolle, Florian Niekiel, Konstantin Weber, Daniel Waldmann, Heiko B. Weber, Bernd Meyer, and Erdmann Spiecker. Dislocations in bilayer graphene. *Nature*, 505(7484):533–537, 2013.

[Ber78] David J. Bergman. The dielectric constant of a composite material: A problem in classical physics. *Phys. Rep.*, 43(9):377–407, 1978.

[Ber79] David J. Bergman. Dielectric constant of a two-component granular composite: A practical scheme for calculating the pole spectrum. *Phys. Rev. B*, 19:2359–2368, Feb 1979.

[BFL+14] D. N. Basov, M. M. Fogler, A. Lanzara, Feng Wang, and Yuanbo Zhang. Colloquium: Graphene spectroscopy. *Rev. Mod. Phys.*, 86(3):959–994, 2014.

[BH04] C. F. Bohren and D. R. Huffman. *Absorption and Scattering of Light by Small Particles*. Wiley, New York, 2004.

[BHH+12] Lola Brown, Robert Hovden, Pinshane Huang, Michal Wojcik, David A Muller, and Jiwoong Park. Twinning and twisting of tri-and bilayer graphene. *Nano Lett.*, 12(3):1609–1615, 2012.

[BL13] László P Biró and Philippe Lambin. Grain boundaries in graphene grown by chemical vapor deposition. *New J. Phys.*, 15(3):035024, 2013.

[BLD16] Arun Bansil, Hsin Lin, and Tanmoy Das. Colloquium: Topological band theory. *Rev. Mod. Phys.*, 88(2):021004, 2016.

[BSAT09] C. W. J. Beenakker, R. A. Sepkhanov, A. R. Akhmerov, and J. Tworzydło. Quantum Goos-Hänchen Effect in Graphene. *Phys. Rev. Lett.*, 102:146804, 2009.

[BV86] P.A. Bobbert and J. Vlieger. Light scattering by a sphere on a substrate. *Physica A*, 137:209 – 242, 1986.

[BV87] P.A. Bobbert and J. Vlieger. The polarizability of a spheroidal particle on a substrate. *Physica A*, 147(1):115 – 141, 1987.

[CB13] N. Candemir and O. Bayrak. Massive Dirac equation in asymmetric Hulthén potential. *J. Math. Phys.*, 54(4):042104, 2013.

[CBAG+12] Jianing Chen, Michela Badioli, Pablo Alonso-Gonzalez, Sukosin Thongrattanasiri, Florian Huth, Johann Osmond, Marko Spasenovic, Alba Centeno, Amaia Pesquera, Philippe Godignon, Amaia Zurutuza Elorza, Nicolas Camara, F. Javier Garcia de Abajo, Rainer Hillenbrand, and Frank H. L. Koppens. Optical nano-imaging of gate-tunable graphene plasmons. *Nature*, 487(7405):77–81, 2012.

[CCW+15] Shane A. Cybart, E. Y. Cho, T. J. Wong, Björn H. Wehlin, Meng K. Ma, Chuong Huynh, and R. C. Dynes. Nano josephson superconducting tunnel junctions in $yba_2cu_3o_{7V\delta}$ directly patterned with a focused helium ion beam. *Nature Nanotech.*, 10:598, 2015.

[CDN+14] Aron W. Cummings, Dinh Loc Duong, Van Luan Nguyen, Dinh Van Tuan, Jani Kotakoski, Jose Eduardo Barrios Vargas, Young Hee Lee, and Stephan Roche. Charge transport in polycrystalline graphene: Challenges and opportunities. *Adv. Mat.*, 26(30):5079, Jun 2014.

[CF06] Vadim V. Cheianov and Vladimir I. Fal'ko. Selective transmission of Dirac electrons and ballistic magnetoresistance of n-p junctions in graphene. *Phys. Rev. B*, 74(4):041403(R), 2006.

[CGS13] Baishali Chakraborty, Kumar S Gupta, and Siddhartha Sen. Topology, cosmic strings and quantum dynamics - a case study with graphene. *J. Phys. Conf. Ser.*, 442:012017, Jun 2013.

[CKC+14] Joshua D Caldwell, Andrey V Kretinin, Yiguo Chen, Vincenzo Giannini, Michael M Fogler, Yan Francescato, Chase T Ellis, Joseph G Tischler, Colin R Woods, Alexander J Giles, Minghui Hong, Kenji Watanabe, Takashi Taniguchi, Stefan A Maier, and Kostya S Novoselov. Sub-diffractional volume-confined polaritons in the natural hyperbolic material hexagonal boron nitride. *Nat. Comm.*, 5:5221, January 2014.

[COH07] A. Cvitkovic, N. Ocelic, and R. Hillenbrand. Analytical model for quantitative prediction of material contrasts inscattering-type near-field optical microscopy. *Opt. Express*, 15(14):8550–8565, 2007.

[CSM11] Nuno J. G. Couto, Benjamin Sacépé, and Alberto F. Morpurgo. Transport through graphene on $SrTiO_3$. *Phys. Rev. Lett.*, 107(22):225501, 2011.

[DEW99] A. Doicu, Yu.A. Eremin, and T. Wriedt. Convergence of the T-matrix method for light scattering from a particle on or near a surface. *Optics Commun.*, 159:266 – 277, 1999.

[DFM+14] S. Dai, Z. Fei, Q. Ma, A. S. Rodin, M. Wagner, A. S. McLeod, M. K. Liu, W. Gannett, W. Regan, M. Thiemens, G. Dominguez, A. H. Castro Neto, A. Zettl, F. Keilmann, P. Jarillo-Herrero, M. M. Fogler, and D. N. Basov. Tunable phonon polaritons in atomically thin van der Waals crystals of boron nitride. *Science*, 343:1125–1129, 2014.

[DMKME14] Alessandro De Martino, Denis Klöpfer, Davron Matrasulov, and Reinhold Egger. Electric-Dipole-Induced Universality for Dirac Fermions in Graphene. *Phys. Rev. Lett.*, 112:186603, 2014.

[DS96] M. Dyakonov and M. Shur. Detection, mixing, and frequency multiplication of terahertz radiation by two-dimensional electronic fluid. *IEEE Trans. Electron Devices*, 43(3):380–387, Mar 1996.

[DSAHR11] S. Das Sarma, Shaffique Adam, E. H. Hwang, and Enrico Rossi. Electronic transport in two-dimensional graphene. *Rev. Mod. Phys.*, 83(2):407, May 2011.

[DSL12] S. Das Sarma and Qiuzi Li. Graphene on $SrTiO_3$. *Solid State Commun.*, 152(19):1795–1799, 2012.

[DYM+10] C. R. Dean, A. F. Young, I. Meric, C. Lee, L. Wang, S. Sorgenfrei, K. Watanabe, T. Taniguchi, P. Kim, K. L. Shepard, and J. Hone. Boron nitride substrates for high-quality graphene electronics. *Nature Nano.*, 5:722–726, 2010.

[EVK06] R Esteban, R Vogelgesang, and K Kern. Simulation of optical near and far fields of dielectric apertureless scanning probes. *Nanotechnology*, 17(2):475, 2006.

[EVK09] R. Esteban, R. Vogelgesang, and K. Kern. Full simulations of the apertureless scanning near field optical microscopy signal: achievable resolution and contrast. *Opt. Express*, 17(4):2518–2529, 2009.

[FAA+16] S. Farajollahi, S. AbdollahRamezani, K. Arik, B. Rejae, and A. Khavasi. Circuit model for plasmons on graphene with one-dimensional conductivity profile. *IEEE Photon. Technol. Lett.*, 28(3):355–358, Feb 2016.

[FAB+11] Zhe Fei, Gregory O. Andreev, Wenzhong Bao, Lingfeng M. Zhang, Alexander S. McLeod, Chen Wang, Margaret K. Stewart, Zeng Zhao, Gerardo Dominguez, Mark Thiemens, Michael M. Fogler, Michael J. Tauber, Antonio H. Castro Neto, Chun Ning Lau, Fritz Keilmann, and Dimitri N. Basov. Infrared Nanoscopy of Dirac Plasmons at the Graphene-SiO_2 Interface. *Nano Lett.*, 11(11):4701–4705, Oct 2011.

[FNGS08] M. M. Fogler, D. S. Novikov, L. I. Glazman, and B. I. Shklovskii. Effect of disorder on a graphene p-n junction. *Phys. Rev. B*, 77(7):075420, 2008.

[FNS07] M. M. Fogler, D. S. Novikov, and B. I. Shklovskii. Screening of a hypercritical charge in graphene. *Phys. Rev. B*, 76:233402, 2007.

[Fog04] Michael M. Fogler. Nonlinear screening and percolative transition in a two-dimensional electron liquid. *Phys. Rev. B*, 69(12):121409(R), 2004.

[Fog13] M. M. Fogler. Electrostatic polarizability of a conical metallic particle in a contact with a dielectric half-space. unpublished, 2013.

[FQB+09] A. Frenzel, M. M. Qazilbash, M. Brehm, Byung-Gyu Chae, Bong-Jun Kim, Hyun-Tak Kim, A. V. Balatsky, F. Keilmann, and D. N. Basov. Inhomogeneous electronic state near the insulator-to-metal transition in the correlated oxide VO_2. *Phys. Rev. B*, 80(11):115115, 2009.

[FRA+12] Z. Fei, A. S. Rodin, G. O. Andreev, W. Bao, A. S. McLeod, M. Wagner, L. M. Zhang, Z. Zhao, M. Thiemens, G. Dominguez, M. M. Fogler, A. H. Castro Neto, C. N. Lau, F. Keilmann, and D. N. Basov. Gate-tuning of graphene plasmons revealed by infrared nano-imaging. *Nature*, 487:82–85, 2012.

[FRG+13] Z. Fei, A. S. Rodin, W. Gannett, S. Dai, W. Regan, M. Wagner, M. K. Liu, A. S. McLeod, G. Dominguez, M. Thiemens, Antonio H. Castro Neto, F. Keilmann, A. Zettl, R. Hillenbrand, M. M. Fogler, and D. N. Basov. Electronic and plasmonic phenomena at graphene grain boundaries. *Nature Nanotech.*, 8(11):821–825, 2013.

[FRS] Han Fu, K. V. Reich, and B. I. Shklovskii. Collapse of electrons to a donor cluster in $SrTiO_3$. arXiv:1504.03675 (unpublished).

[FSG+] Z Fei, M Scott, D. J. Gosztola, J. J. Foley, J. Yan, D. G. Mandrus, H. Wen, P. Zhou, D. W. Zhang, Y. Sun, J. R. Guest, S. K. Gray, W. Bao, G. P. Wiederrecht, and X. Xu. Nano-optical imaging of exciton polaritons inside WSe_2 waveguides. arXiv:1601.02133.

[FW81] G.W. Ford and W.H. Weber. Electromagnetic effects on a molecule at a metal surface: I. Effects of nonlocality and finite molecular size. *Surface Science*, 109(2):451 – 481, 1981.

[FW84] G.W. Ford and W.H. Weber. Electromagnetic interactions of molecules with metal surfaces. *Physics Reports*, 113(4):195 – 287, 1984.

[FXT+11] Aires Ferreira, Xiangfan Xu, Chang-Lin Tan, Su-Kang Bae, N. M. R. Peres, Byung-Hee Hong, Barbaros Özyilmaz, and A. H. Castro Neto. Transport properties of graphene with one-dimensional charge defects. *Europhys. Lett. (EPL)*, 94(2):28003, 2011.

[GBOR14] Justin A. Gerber, Samuel Berweger, Brian T. O'Callahan, and Markus B. Raschke. Phase-Resolved Surface Plasmon Interferometry of Graphene. *Phys. Rev, Lett.*, 113(5):055502, 2014.

[GdA14] F. Javier García de Abajo. Graphene plasmonics: Challenges and opportunities. *ACS Photon.*, 1(3):135–152, 2014.

[GDG+16] Alexander J. Giles, Siyuan Dai, Orest J. Glembocki, Andrey V. Kretinin, Zhiyuan Sun, Chase T. Ellis, Joseph G. Tischler, Takashi Taniguchi, Kenji Watanabe, Michael M. Fogler, Kostya S. Novoselov, Dimitri. N. Basov, and Joshua D. Caldwell. Imaging of anomalous internal reflections of hyperbolic phonon-polaritons in hexagonal boron nitride. *Nano Lett.*, 16(6):3858–3865, jun 2016.

[GDPC13] J. S. Gómez-Díaz and J. Perruisseau-Carrier. Graphene-based plasmonic switches at near infrared frequencies. *Opt. Express*, 21(13):15490–15504, Jul 2013.

[GGS15] E. V. Gorbar, V. P. Gusynin, and O. O. Sobol. Supercriticality of novel type induced by electric dipole in gapped graphene. *Phys. Rev. B*, 92(23):235417, dec 2015.

[GNP+15] Michael D. Goldflam, Guang-Xin Ni, Kirk W. Post, Zhe Fei, Yuting Yeo, Jun You Tan, Aleksandr S. Rodin, Brian C. Chapler, Barbaros Ozyilmaz, Antonio H. Castro Neto, Michael M. Fogler, and D. N. Basov. Tuning and persistent switching of graphene plasmons on a ferroelectric substrate. *Nano Lett.*, 15(8):4859–4864, 2015.

[GPN12] A. N. Grigorenko, M. Polini, and K. S. Novoselov. Graphene plasmonics. *Nature Photon.*, 6:749–758, 2012.

[GPNMM13] Juan Luis Garcia-Pomar, Alexey Yu. Nikitin, and Luis Martin-Moreno. Scattering of Graphene Plasmons by Defects in the Graphene Sheet. *ACS Nano*, 7(6):4988–4994, jun 2013.

[GR00] I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series, and Products*. Academic, San Diego, 6th edition, 2000.

[HDS07] E. H. Hwang and S. Das Sarma. Dielectric Function, Screening, and Plasmons in Two-Dimensional Graphene. *Phys. Rev. B*, 75:205418, May 2007.

[HET12] Benedikt Hauer, Andreas P. Engelhardt, and Thomas Taubner. Quasi-analytical model for scattering infrared near-field microscopy on layered systems. *Opt. Express*, 20(12):13173–13188, 2012.

[HK00] R. Hillenbrand and F. Keilmann. Complex Optical Constants on a Subwavelength Scale. *Phys. Rev, Lett.*, 85(14):3029–3032, 2000.

[HNA94] Alang Kasim Harman, Susumu Ninomiya, and Sadao Adachi. Optical constants of sapphire (α-Al_2O_3) single crystals. *J. Appl. Phys.*, 76(12):8032–8036, 1994.

[Hob29] E. W. Hobson. The Addition Theorems for General Legendre Functions. *Proc. London Math. Soc.*, s2-29(1):355–381, 1929.

[HP14] R. R. Hartmann and M. E. Portnoi. Quasi-exact solution to the Dirac equation for the hyperbolic-secant potential. *Phys. Rev. A*, 89:012101, 2014.

[HPT+15] Jhao-Wun Huang, Cheng Pan, Son Tran, Bin Cheng, Kenji Watanabe, Takashi Taniguchi, Chun Ning Lau, and Marc Bockrath. Superior current carrying capacity of boron nitride encapsulated carbon nanotubes with zero-dimensional contacts. *Nano Lett.*, 15(10):6836, oct 2015.

[HS17] Eddwi H. Hasdeo and Justin C. W. Song. Long-lived domain wall plasmons in gapped bilayer graphene. *arXiv:1706.00790*, 2017.

[HSW+11] F. Huth, M. Schnell, J. Wittborn, N. Ocelic, and R. Hillenbrand. Infrared-spectroscopic nanoimaging with a thermal source. *Nature Mater.*, 10:352–356, 2011.

[IZ13] S. Ihnatsenka and I. V. Zozoulenko. Electron interaction, charging, and screening at grain boundaries in graphene. *Phys. Rev. B*, 88(8):085436, 2013.

[JBS09] Marinko Jablan, Hrvoje Buljan, and Marin Soljačić. Plasmonics in graphene at infrared frequencies. *Phys. Rev. B*, 80:245435, 2009.

[JBW+10] Andrew C. Jones, Samuel Berweger, Jiang Wei, David Cobden, and Markus B. Raschke. Nano-optical Investigations of the Metal-Insulator Phase Behavior of Individual VO2 Microcrystals. *Nano Lett.*, 10(5):1574–1581, 2010.

[JF15] B.-Y. Jiang and M. M. Fogler. Electronic response of graphene to linelike charge perturbations. *Phys. Rev. B*, 91:235422, Jun 2015.

[JNP+16] B.-Y. Jiang, G. X. Ni, C. Pan, Z. Fei, B. Cheng, C. N. Lau, M. Bockrath, D. N. Basov, and M. M. Fogler. Tunable plasmonic reflection by bound 1D electron states in a 2D Dirac metal. *Phys. Rev. Lett.*, 117(8):086801, 2016.

[JSN+15] Long Ju, Zhiwen Shi, Nityan Nair, Yinchuan Lv, Chenhao Jin, Jairo Velasco, Claudia Ojeda-Aristizabal, Hans A. Bechtel, Michael C. Martin, Alex Zettl, James Analytis, and Feng Wang. Topological valley transport at bilayer graphene domain walls. *Nature*, 520(7549):650, Apr 2015.

[JSZ+16] Lili Jiang, Zhiwen Shi, Bo Zeng, Sheng Wang, Ji-Hun Kang, Trinity Joshi, Chenhao Jin, Long Ju, Jonghwan Kim, Tairu Lyu, Yuen-Ron Shen, Michael Crommie, Hong-Jun Gao, and Feng Wang. Soliton-dependent plasmon reflection at bilayer graphene domain walls. *Nature Mater.*, 15(8):840–844, 2016.

[JZCN+16] B-Y Jiang, LM Zhang, AH Castro Neto, DN Basov, and MM Fogler. Generalized spectral method for near-field optical microscopy. *J. Appl. Phys.*, 119(5):054305, 2016.

[Kat06] M. I. Katsnelson. Nonlinear screening of charge impurities in graphene. *Phys. Rev. B*, 74:201401, 2006.

[KBF97] Charles Kane, Leon Balents, and Matthew P. A. Fisher. Coulomb interactions and mesoscopic effects in carbon nanotubes. *Phys. Rev. Lett.*, 79:5086–5089, Dec 1997.

[KDC+09] Wojciech Knap, Mikhail Dyakonov, Dominique Coquillat, Frederic Teppe, Nina Dyakonova, Jerzy Łusakowski, Krzysztof Karpierz, Maciej Sakowicz, Gintaras Valusis, Dalius Seliuta, Irmantas Kasalynas, Abdelouahad El Fatimy, Y. M. Meziani, and Taiichi Otsuji. Field effect transistors for terahertz detection: Physics and first imaging applications. *J. Infrared Millim. Terahertz Waves*, 30(12):1319–1337, Dec 2009.

[KDS13] K. Kechedzhi and S. Das Sarma. Plasmon anomaly in the dynamical optical conductivity of graphene. *Phys. Rev. B*, 88:085403, 2013.

[Ken02] Piers Kennedy. The Woods-Saxon potential in the Dirac equation. *J. Phys. A*, 35(3):689, 2002.

[KH04] Fritz Keilmann and Rainer Hillenbrand. Near-field microscopy by elastic light scattering from a tip. *Phil. Trans. Roy. Soc. London, Ser. A*, 362(1817):787–805, 2004.

[KH09] F. Keilmann and R. Hillenbrand. *Nano-optics and Near-field Optical Microscopy*, chapter 11: Near-Field Nanoscopy by Elastic Light Scattering from a Tip, pages 235–266. Artech House, Norwood, 2009. edited by A. Zayats and D. Richards.

[KN94] A. Kučírková and K. Navrátil. Interpretation of Infrared Transmittance Spectra of SiO2 Thin Films. *Appl. Spectrosc.*, 48(1):113–120, 1994.

[KN07] M.I. Katsnelson and K.S. Novoselov. Graphene: New bridge between condensed matter physics and quantum electrodynamics. *Solid State Commun.*, 143:3–13, 2007.

[Kos13] Mikito Koshino. Electronic transmission through AB-BA domain boundary in bilayer graphene. *Phys. Rev. B*, 88(11):115409, 2013.

[KSZ13] Eugene B. Kolomeisky, Joseph P. Straley, and Hussain Zaidi. Fermion space charge in narrow band-gap semiconductors, weyl semimetals, and around highly charged nuclei. *Phys. Rev. B*, 88(16):165428, 2013.

[KUP+12] Valeri N. Kotov, Bruno Uchoa, Vitor M. Pereira, F. Guinea, and A. H. Castro Neto. Electron-Electron Interactions in Graphene: Current Status and Perspectives. *Rev. Mod. Phys.*, 84(3):1067–1125, 2012.

[KWE+13] Justin C. Koepke, Joshua D. Wood, David Estrada, Zhun-Yong Ong, Kevin T. He, Eric Pop, and Joseph W. Lyding. Atomic-Scale Evidence for Potential Barriers and Strong Carrier Scattering at Graphene Grain Boundaries: A Scanning Tunneling Microscopy Study. *ACS Nano*, 7(1):75, Jan 2013.

[KWS+17] Ji-Hun Kang, Sheng Wang, Zhiwen Shi, Wenyu Zhao, Eli Yablonovitch, and Feng Wang. Goos-hänchen shift and even-odd peak oscillations in edge-reflections of surface polaritons in atomically thin crystals. *Nano Letters*, 17(3):1768–1774, 2017.

[LFZ+13] Junhao Lin, Wenjing Fang, Wu Zhou, Andrew R. Lupini, Juan Carlos Idrobo, Jing Kong, Stephen J. Pennycook, and Sokrates T. Pantelides. AC/AB stacking boundaries in bilayer graphene. *Nano Lett.*, 13(7):3262–3268, 2013.

[LGW+17] Mark B. Lundeberg, Yuanda Gao, Achim Woessner, Cheng Tan, Pablo Alonso-Gonzalez, Kenji Watanabe, Takashi Taniguchi, James Hone, Rainer Hillenbrand, and Frank H. L. Koppens. Thermoelectric detection and imaging of propagating graphene plasmons. *Nature Mater.*, 16:204–207, 2017.

[LL91] L. D. Landau and E. M. Lifshitz. *Quantum Mechanics: Non-Relativistic Theory*. Butterworth-Heinemann, Amsterdam, 3rd ed. edition, 1991.

[LLPK16] Irina V. Lebedeva, Alexander V. Lebedev, Andrey M. Popov, and Andrey A. Knizhnik. Dislocations in stacking and commensurate-incommensurate phase transition in bilayer graphene and hexagonal boron nitride. *Phys. Rev. B*, 93(23):235414, 2016.

[LNK+10] Keji Lai, Masao Nakamura, Worasom Kundhikanjana, Masashi Kawasaki, Yoshinori Tokura, Michael A. Kelly, and Zhi-Xun Shen. Mesoscopic Percolating Resistance Network in a Strained Manganite Thin Film. *Science*, 329(5988):190–193, 2010.

[LWM+16] Jing Li, Ke Wang, Kenton J. McFaul, Zachary Zern, Yafei Ren, Kenji Watanabe, Takashi Taniguchi, Zhenhua Qiao, and Jun Zhu. Gate-controlled topological conducting channels in bilayer graphene. *Nature Nanotech.*, 11:1060–1065, 2016.

[Man69] Gerald S. Manning. Limiting laws and counterion condensation in polyelectrolyte solutions i. colligative properties. *J. Chem. Phys.*, 51(3):924, 1969.

[MAT] MATLAB Release 2012b, The MathWorks, Inc., Natick, Massachusetts.

[MBM08] Ivar Martin, Ya. M. Blanter, and A. F. Morpurgo. Topological confinement in bilayer graphene. *Phys. Rev. Lett.*, 100(3):036804, 2008.

[MDP48] C. G. Montgomery, R. H. Dicke, and E. M. Purcell. *Principles of Microwave Circuits*, volume 8 of *MIT Radiation Laboratory Series*. McGraw-Hill, 1948.

[MF81] Philip McCord Morse and Herman Feshbach. *Methods Of Theoretical Physics*. Feshbach Publishing, Minneapolis, 1981.

[MK13] Edward McCann and Mikito Koshino. The electronic properties of bilayer graphene. *Rep. Prog. Phys.*, 76(5):056503, 2013.

[MKG+14] Alexander S. McLeod, P. Kelly, M. D. Goldflam, Z. Gainsforth, A. J. Westphal, Gerardo Dominguez, Mark H. Thiemens, Michael M. Fogler, and D. N. Basov. Model for quantitative tip-enhanced spectroscopy and the extraction of nanoscale-resolved optical constants. *Phys. Rev. B*, 90:085136, 2014.

[ML06] Hongzhou Ma and Jeremy Levy. GHz Apertureless Near-Field Scanning Optical Microscopy of Ferroelectric Nanodomain Dynamics. *Nano Lett.*, 6(3):341–344, 2006.

[Mor09] Alexander Moroz. Depolarization field of spheroidal particles. *J. Opt. Soc. Am. B*, 26(3):517–527, 2009.

[NCV99] J. R. S. Nascimento, Inyong Cho, and Alexander Vilenkin. Charged vacuum condensate near a superconducting cosmic string. *Phys. Rev. D*, 60:083505, 1999.

[NIVO+13] Péter Nemes-Incze, Péter Vancsó, Zoltán Osváth, Géza I. Márk, Xiaozhan Jin, Yong-Sung Kim, Chanyong Hwang, Philippe Lambin, Claude Chapelier, and László Péter Biró. Electronic states of disordered grain boundaries in graphene prepared by chemical vapor deposition. *Carbon*, 64:178, Nov 2013.

[NJB+15] Guangxin Ni, Bor-Yuan Jiang, Marc Bockrath, Chen-Ning Lau, M. M. Fogler, and D. N. Basov. Tunable plasmonic reflectors in graphene nanostructures with carbon nanotube gates. unpublished, 2015.

[NKP+09] Maxim P. Nikiforov, Susanne C. Kehr, Tae-Hong Park, Peter Milde, Ulrich Zerweck, Christian Loppacher, Lukas M. Eng, Michael J. Therien, Nader Engheta, and Dawn Bonnell. Probing polarization and dielectric function of molecules with higher order harmonics in scattering-near-field scanning optical microscopy. *J. Appl. Phys.*, 106(11):114307, 2009.

[NLMM14] A. Yu. Nikitin, T. Low, and L. Martin-Moreno. Anomalous reflection phase of graphene plasmons and its influence on resonators. *Phys. Rev. B*, 90:041407, Jul 2014.

[Nov07] D. S. Novikov. Elastic scattering theory and transport in graphene. *Phys. Rev. B*, 76(24):245435, 2007.

[NWG+16] G. X. Ni, L. Wang, M. D. Goldflam, M. Wagner, Z. Fei, A. S. McLeod, M. K. Liu, F. Keilmann, B. Özyilmaz, A. H. Castro Neto, J. Hone, M. M. Fogler, and D. N. Basov. Ultrafast optical switching of infrared plasmon polaritons in high-mobility graphene. *Nature Photon.*, 10(4):244–247, 2016.

[NWW+15] G. X. Ni, H. Wang, J. S. Wu, Z. Fei, M. D. Goldflam, F. Keilmann, B. Özyilmaz, A. H. Castro Neto, X. M. Xie, M. M. Fogler, and D. N. Basov. Plasmons in graphene moiré superlattices. *Nature Mater.*, 14(12):1217–1222, 2015.

[Oce07] Nenad Ocelić. *Quantitative Near-Field Phonon-Polariton Spectroscopy*, volume 89 of *Selected Topics of Semiconductor Physics and Technology*. Walter Schottky Instituts der Technischen Universität München, 2007.

[Oos68] Fumio Oosawa. A theory on the effect of low molecular salts on the dissociation of linear polyacids. *Biopolymers*, 6(1):135–144, 1968.

[Ozb06] Ekmel Ozbay. Plasmonics: Merging photonics and electronics at nanoscale dimensions. *Science*, 311(5758):189–193, 2006.

[PCL+14] Alessandro Principi, Matteo Carrega, Mark B. Lundeberg, Achim Woessner, Frank H. L. Koppens, Giovanni Vignale, and Marco Polini. Plasmon losses due to electron-phonon scattering: The case of graphene encapsulated in hexagonal boron nitride. *Phys. Rev. B*, 90(16):165408, 2014.

[PCNLM10] Vitor M. Pereira, A. H. Castro Neto, H. Y. Liang, and L. Mahadevan. Geometry, Mechanics, and Electronics of Singular Structures and Wrinkles in Graphene. *Phys. Rev. Lett.*, 105:156603, 2010.

[PFM13] J. Polanco, R. M. Fitzgerald, and A. A. Maradudin. Scattering of surface plasmon polaritons by one-dimensional surface defects. *Phys. Rev. B*, 87:155417, Apr 2013.

[PJALR03] J. A. Porto, P. Johansson, S. P. Apell, and T. López-Ríos. Resonance shift effects in apertureless scanning near-field optical microscopy. *Phys. Rev. B*, 67(8):085409, 2003.

[PNCN07] Vitor M. Pereira, Johan Nilsson, and A. H. Castro Neto. Coulomb Impurity Problem in Graphene. *Phys. Rev. Lett.*, 99(16):166802, 2007.

[PPN12] V. V. Popov, O. V. Polishchuk, and S. A. Nikitov. Electromagnetic renormalization of the plasmon spectrum in a laterally screened two-dimensional electron system. *JETP Lett.*, 95(2):85–90, Mar 2012.

[PRL+05] A. Passian, R. H. Ritchie, A. L. Lereu, T. Thundat, and T. L. Ferrell. Curvature effects in surface plasmon dispersion and coupling. *Phys. Rev. B*, 71:115425, 2005.

[PS45] I. Ya. Pomeranchuk and Ya. A. Smorodinsky. On energy levels in systems with $Z > 137$. *J. Phys. USSR*, 9:97, 1945.

[PTSK05] V. V. Popov, G. M. Tsymbalov, M. S. Shur, and W. Knap. The resonant terahertz response of a slot diode with a two-dimensional electron channel. *Semiconductors*, 39(1):142–146, Jan 2005.

[PWM+01] A. Passian, A. Wig, F. Meriaudeau, M. Buncick, T. Thundat, and T. L. Ferrell. Electrostatic force density for a scanned probe above a charged surface. *J. Appl. Phys.*, 90(2):1011, 2001.

[QBA+09] M. M. Qazilbash, M. Brehm, G. O. Andreev, A. Frenzel, P.-C. Ho, Byung-Gyu Chae, Bong-Jun Kim, Sun Jin Yun, Hyun-Tak Kim, A. V. Balatsky, O. G. Shpyrko, M. B. Maple, F. Keilmann, and D. N. Basov. Infrared spectroscopy and nano-imaging of the insulator-to-metal transition in vanadium dioxide. *Phys. Rev. B*, 79(7):075107, Feb 2009.

[QBC+07] M. M. Qazilbash, M. Brehm, Byung-Gyu Chae, P.-C. Ho, G. O. Andreev, Bong-Jun Kim, Sun Jin Yun, A. V. Balatsky, M. B. Maple, F. Keilmann, Hyun-Tak Kim, and D. N. Basov. Mott Transition in VO2 Revealed by Infrared Spectroscopy and Nano-Imaging. *Science*, 318(5857):1750–1753, 2007.

[QJNM11] Zhenhua Qiao, Jeil Jung, Qian Niu, and Allan H. MacDonald. Electronic highways in bilayer graphene. *Nano Lett.*, 11:3453–3459, 2011.

[RGEH05] Jan Renger, Stefan Grafström, Lukas M. Eng, and Rainer Hillenbrand. Resonant light scattering by near-field-induced phonon polaritons. *Phys. Rev. B*, 71(7):075410, 2005.

[RK15] Behzad Rejaei and Amin Khavasi. Scattering of surface plasmons on graphene by a discontinuity in surface conductivity. *J. Opt.*, 17:075002, 2015.

[RS81] R. W. Rendell and D. J. Scalapino. Surface plasmons confined by microstructures on tunnel junctions. *Phys. Rev. B*, 24(6):3276–3294, 1981.

[RSK+04] V. Ryzhii, A. Satou, I. Khmyrova, A. Chaplik, and M. S. Shur. Plasma oscillations in a slot diode structure with a two-dimensional electron channel. *J. Appl. Phys.*, 96:7625, 2004.

[RSS03] V. Ryzhii, A. Satou, and M. S. Shur. Admittance of a slot diode with a two-dimensional electron channel. *J. Appl. Phys.*, 93(12):10041–10045, 2003.

[RSZF13] T. M. Radchenko, A. A. Shylau, I. V. Zozoulenko, and Aires Ferreira. Effect of charged line defects on conductivity in graphene: Numerical Kubo and analytical Boltzmann approaches. *Phys. Rev. B*, 87:195448, 2013.

[SCH14] M. Schnell, P. S. Carney, and R. Hillenbrand. Synthetic optical holography for rapid nanoimaging. *Nature Comm.*, 5:3499, Mar 2014.

[SF12] I. Sodemann and M. M. Fogler. Interaction corrections to the polarization function of graphene. *Phys. Rev. B*, 86:115408, 2012.

[SGRBF15] Z. Sun, A. Gutiérrez-Rubio, D. N. Basov, and M. M. Fogler. Hamiltonian optics of hyperbolic polaritons in nanogranules. *Nano Letters*, 15:4455–4460, 2015.

[SHB+15] Zhiwen Shi, Xiaoping Hong, Hans A. Bechtel, Bo Zeng, Michael C. Martin, Kenji Watanabe, Takashi Taniguchi, Yuen-Ron Shen, and Feng Wang. Observation of a Luttinger-liquid plasmon in metallic single-walled carbon nanotubes. *Nature Photon.*, 9(8):515–519, jul 2015.

[SJGG+14] Pablo San-Jose, RV Gorbachev, AK Geim, KS Novoselov, and F Guinea. Stacking boundaries and transport in bilayer graphene. *Nano Lett.*, 14(4):2052–2057, 2014.

[SKJ+14] Surajit Saha, Orhan Kahya, Manu Jaiswal, Amar Srivastava, Anil Annadi, Jayakumar Balakrishnan, Alexandre Pachoud, Chee-Tat Toh, Byung-Hee Hong, Jong-Hyun Ahn, T. Venkatesan, and Barbaros Özyilmaz. Unconventional Transport through Graphene on $SrTiO_3$: A Plausible Effect of $SrTiO_3$ Phase-Transitions. *Sci. Rep.*, 4:6173, 2014.

[SKL07a] A. V. Shytov, M. I. Katsnelson, and L. S. Levitov. Atomic Collapse and Quasi-Rydberg States in Graphene. *Phys. Rev. Lett.*, 99:246802, 2007.

[SKL07b] A. V. Shytov, M. I. Katsnelson, and L. S. Levitov. Vacuum Polarization and Screening of Supercritical Impurities in Graphene. *Phys. Rev. Lett.*, 99:236801, 2007.

[SLS14] Raymond Sachs, Zhisheng Lin, and Jing Shi. Ferroelectric-like $SrTiO_3$ surface dipoles probed by graphene. *Sci. Rep.*, 4:3657, 2014.

[Smy50] William Smythe. *Static and Dynamic Electricity*. McGraw-Hill, 2 edition, 1950.

[SRC05] A. Satou, V. Ryzhii, and A. Chaplik. Plasma oscillations in two-dimensional electron channel with nonideally conducting side contacts. *J. Appl. Phys.*, 98(3):034502, 2005.

[SWJ+14] H. T. Stinson, J. S. Wu, B. Y. Jiang, Z. Fei, A. S. Rodin, B. C. Chapler, A. S. McLeod, A. Castro Neto, Y. S. Lee, M. M. Fogler, and D. N. Basov. Infrared nanospectroscopy and imaging of collective superfluid excitations in anisotropic superconductors. *Phys. Rev. B*, 90:014502, Jul 2014.

[SYB+15] Y. Shimazaki, M. Yamamoto, I. V. Borzenets, K. Watanabe, T. Taniguchi, and S. Tarucha. Generation and detection of pure valley current by electrically induced berry curvature in bilayer graphene. *Nature Phys.*, 11(12):1032–1036, 2015.

[TBL+12] Adam W. Tsen, Lola Brown, Mark P. Levendorf, Fereshte Ghahari, Pinshane Y. Huang, Robin W. Havener, Carlos S. Ruiz-Vargas, David A. Muller, Philip Kim, and Jiwoong Park. Tailoring Electrical Transport Across Grain Boundaries in Polycrystalline Graphene. *Science*, 336(6085):1143–1146, 2012.

[TKH04] T. Taubner, F. Keilmann, and R. Hillenbrand. Nanomechanical Resonance Tuning and Phase Effects in Optical Near-Field Interaction. *Nano Lett.*, 4(9):1669–1672, 2004.

[TMC+15] Jiayue Tong, Martin Muthee, Shao-Yu Chen, Sigfrid K. Yngvesson, and Jun Yan. Antenna enhanced graphene thz emitter and detector. *Nano Letters*, 15(8):5295–5301, 2015. PMID: 26218887.

[VE11] Ashkan Vakil and Nader Engheta. Transformation Optics Using Graphene. *Science*, 332(6035):1291–1294, 2011.

[VG03] Víctor M. Villalba and Walter Greiner. Transmission resonances and supercritical states in a one-dimensional cusp potential. *Phys. Rev. A*, 67:052707, 2003.

[VLN+13] Abolhassan Vaezi, Yufeng Liang, Darryl H. Ngai, Li Yang, and Eun-Ah Kim. Topological edge states at a tilt boundary in gated multilayer graphene. *Phys. Rev. X*, 3(2):021018, 2013.

[WAGL+15] Achim Woessner, Pablo Alonso-Gonzalez, Mark B. Lundeberg, Gabriele Navickaite, Yuanda Gao, Qiong Ma, Davide Janner, Kenji Watanabe, Takashi Taniguchi, Valerio Pruneri, Pablo Jarillo-Herrero, James Hone, Rainer Hillenbrand, and Frank H.L. Koppens. Non-invasive photocurrent nanoscopy on bare and hexagonal boron nitride encapsulated graphene. unpublished, 2015.

[WD98] Thomas Wriedt and Adrian Doicu. Light scattering from a particle on or near a surface. *Optics Commun.*, 152(4):376 – 384, 1998.

[WGL82] A. Wokaun, J. P. Gordon, and P. F. Liao. Radiation Damping in Surface-Enhanced Raman Scattering. *Phys. Rev. Lett.*, 48:957–960, 1982.

[WGT+17] Achim Woessner, Yuanda Gao, Iacopo Torre, Mark B. Lundeberg, Cheng Tan, Kenji Watanabe, Takashi Taniguchi, Rainer Hillenbrand, James Hone, Marco Polini, and Frank H. L. Koppens. Electrical 2π phase control of infrared light in a 350-nm footprint using graphene plasmons. *Nature Photon.*, 11:421, 2017.

[Whe59] Harold Wheeler. The Radiansphere around a Small Antenna. *Proc. IRE*, 47(8):1325–1331, 1959.

[WLG+14] Achim Woessner, Mark B. Lundeberg, Yuanda Gao, Alessandro Principi, Pablo Alonso-González, Matteo Carrega, Kenji Watanabe, Takashi Taniguchi, Giovanni Vignale, Marco Polini, James Hone, Rainer Hillenbrand, and Frank H. L. Koppens. Highly confined low-loss plasmons in graphene-boron nitride heterostructures. *Nature Mater.*, 14(4):421–425, 2014.

[WPD+17] Achim Woessner, Romain Parret, Diana Davydovskaya, Yuanda Gao, Jhih-Sheng Wu, Mark B. Lundeberg, Sébastien Nanot, Pablo Alonso-González, Kenji Watanabe, Takashi Taniguchi, Rainer Hillenbrand, Michael M. Fogler, James Hone, and Frank H. L. Koppens. Electrical detection of hyperbolic phonon-polaritons in heterostructures of graphene and boron nitride. *NPJ 2D Mater. Appl.*, 1:25, 2017.

[WSSG06] B. Wunsch, T. Stauber, F. Sols, and F. Guinea. Dynamical Polarization of Graphene at Finite Doping. *New J. Phys.*, 8:318, Dec 2006.

[WWS+13] Yang Wang, Dillon Wong, Andrey V. Shytov, Victor W. Brar, Sangkook Choi, Qiong Wu, Hsin-Zon Tsai, William Regan, Alex Zettl, Roland K. Kawakami, Steven G. Louie, Leonid S. Levitov, and Michael F. Crommie. Observing Atomic Collapse Resonances in Artificial Nuclei on Graphene. *Science*, 340(6133):734–737, 2013.

[YC14] Oleg V. Yazyev and Yong P. Chen. Polycrystalline graphene and other two-dimensional materials. *Nature Nano.*, 9(10):755, Aug 2014.

[YJB+13] G. L. Yu, R. Jalil, Branson Belle, Alexander S. Mayorov, Peter Blake, Frederick Schedin, Sergey V. Morozov, Leonid A. Ponomarenko, F. Chiappini, S. Wiedmann, Uli Zeitler, Mikhail I. Katsnelson, A. K. Geim, Kostya S. Novoselov, and Daniel C. Elias. Interaction phenomena in graphene seen through quantum capacitance. *Proc. Natl. Acad. Sci. USA*, 110(9):3282–3286, 2013.

[YJQH16] Long-Jing Yin, Hua Jiang, Jia-Bin Qiao, and Lin He. Direct imaging of topological edge states at a bilayer graphene domain wall. *Nature Commun.*, 7:11760, 2016.

[YK09] Andrea F. Young and Philip Kim. Quantum interference and Klein tunnelling in graphene heterojunctions. *Nature Phys.*, 5:222 – 226, 2009.

[ZAF+12] L. M. Zhang, G. O. Andreev, Z. Fei, A. S. McLeod, G. Dominguez, M. Thiemens, A. H. Castro-Neto, D. N. Basov, and M. M. Fogler. Near-field spectroscopy of silicon dioxide thin films. *Phys. Rev. B*, 85:075419, 2012.

[ZAH+07] Hui Zhan, Victoria Astley, Michael Hvasta, Jason A. Deibel, Daniel M. Mittleman, and Yong-Sik Lim. The metal-insulator transition in VO_2 studied using terahertz apertureless near-field microscopy. *Appl. Phys. Lett.*, 91(16):162110, 2007.

[ZF08] L. M. Zhang and M. M. Fogler. Nonlinear Screening and Ballistic Transport in a Graphene *p-n* Junction. *Phys. Rev. Lett.*, 100:116804, Mar 2008.

[ZLB+08] L. M. Zhang, Z. Q. Li, D. N. Basov, M. M. Fogler, Z. Hao, and M. C. Martin. Determination of the electronic structure of bilayer graphene from infrared spectroscopy. *Phys. Rev. B*, 78(23):235408, 2008.

[ZMM13] F. Zhang, A. H. MacDonald, and E. J. Mele. Valley chern numbers and boundary modes in gapped bilayer graphene. *Proc. Natl Acad. Sci. USA*, 110(26):10546–10551, 2013.

[ZP72] Ya. B. Zeldovich and Valentin S. Popov. Electronic structure of superheavy atoms. *Sov. Phys. Uspekhi*, 14(6):673, 1972.

[ZPM13] Zeidler, S., Posch, Th., and Mutschke, H. Optical constants of refractory oxides at high temperatures. *Astron. Astrophys.*, 553:A81, 2013.