ECE 458 Lecture Notes on

Applications of Radio Wave Propagation

Erhan Kudeki
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

January 2006
(Edited: Jan. 2010 & Nov. 2012 & Nov. 2019)

ECE 458 is a follow-on course to ECE 350 concerned with the complexities of EM wave propagation on and around our planet with a special focus on near-Earth space environment including the troposphere, stratosphere, mesosphere, ionosphere, magnetosphere, and solar wind, the last three occupied by magnetized plasmas. We will assume ECE 350 level familiarity with Maxwell's equations, antenna fundamentals, and plane wave concepts, as well as familiarity with linear system theory tools (ECE 210) and random signals and thermal noise (ECE 313, ECE 350). Very broadly, the first one third of the course is focused on radiowave communications links near the ground and within the troposphere, the middle third on ionospheric propagation above 50 km altitude and sky-wave communication links, and the last third on radio remote sensing of the near Earth space environment using reflective and scattering based techniques.

Outline

1. Introduction and overview
2. Space-wave links above PEC and good conducting ground - Sections 1.5, 2.5
3. Plane wave decomposition of spherical waves and Sommerfeld integrals - App 5, App 6
4. Method of steepest descent and ground-wave links - App 6, App 7, 2.5
5. Huygens principle and Fourier optics - App 7
6. Fresnel zones in space-wave links - App 7
7. Knife edge diffraction and edge waves - App 7
8. Diffracted radio links, scintillations - App 7
9. Scintillations, partial reflections spaced antenna drifts - App 7
10. Ray tracing in plane stratified troposphere - 4.1, 4.2
11. Ray tracing and ducting in spherical stratified troposphere - 4.4, 4.5
12. Ionospheric propagation and dispersion - 3.1, 3.2
    - TAKE HOME 1
13. Phase and group delays, TEC - 3.3, 3.4
14. Ionospheric reflections at normal incidence - 3.5, 3.6
15. WKB and full-wave solutions - 3.6
16. Ionosondes and ionograms - 3.7 - Travel
17. Sky-wave links in plane stratified ionosphere - 4.6 - Travel

18. Sky-wave links in spherical geometry - 4.7, 4.8
19. Ionospheric absorption - 3.8
    - TAKE HOME 2
20. Radar equation, RCS, antenna noise, SNR - 5.1, 5.2
21. Soft-target radar equation, coherent and incoherent scatter radars - 5.2, 5.3
22. ISR spectrum measurements and estimation 5.3, 5.4
23. ISR spectral theory and spectral fitting - App 8
24. Anisotropic propagation and magneto-ionic theory - 6.1, 6.2
25. TEM waves, Appleton-Hartree equation, O- and X-modes - 6.2
26. Cutoffs, resonances, vertical sounding - 6.2, App 9
27. Ionogram inversions - App 9
28. QL- and QT-propagation and Faraday rotation - 6.3, 6.4
29. Radar differential phase measurements - 6.3, 6.4
    - TAKE HOME FINAL

Contents

1 Plane wave solutions of Maxwell's equations in material media

1.1 Electromagnetic fields and Maxwell's equations in phasor and plane wave forms

Interactions of charged particles q_j constituting space-time charge and current densities $\rho(\mathbf{r},t)$ and $\mathbf{J}(\mathbf{r},t)$ are described by *Lorentz force*

$$\mathbf{F} = q_j(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \tag{1.1}$$

in terms electric and magnetic fields $\mathbf{E}(\mathbf{r},t)$ and $\mathbf{B}(\mathbf{r},t)$ obeying Maxwell's equations

$$\begin{aligned}
\nabla \cdot \mathbf{E} &= \rho/\epsilon_o, && \textit{Gauss' law} && (1.2)\\
\nabla \cdot \mathbf{B} &= 0, && \textit{No magnetic charge} && (1.3)\\
\nabla \times \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t}, && \textit{Faraday's law} && (1.4)\\
\nabla \times \mathbf{B} &= \mu_o \mathbf{J} + \mu_o \epsilon_o \frac{\partial \mathbf{E}}{\partial t}. && \textit{Ampere's law} && (1.5)
\end{aligned}$$

Here $\mathbf{v} \equiv \frac{d\mathbf{r}}{dt}$ is the charged particle velocity in the reference frame where $\mathbf{E}$ and $\mathbf{B}$ are measured, $\mathbf{r} \equiv (x,y,x) \equiv x\hat{x} + y\hat{y} + z\hat{z}$ is the position vector of the particle, $\nabla \equiv (\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z})$ is partial differential *del* operator used in divergence, curl, and gradient calculations in Cartesian coordinates (with a right-handed convention) and ϵ_o and μ_o are *permittivity* and *permeability* of free space with numerical values of[1]

$$\epsilon_o \approx \frac{1}{36\pi 10^9}\,\frac{\mathrm{F}}{\mathrm{m}} \quad \text{and} \quad \mu_o = 4\pi 10^{-7}\,\frac{\mathrm{H}}{\mathrm{m}}.$$

In sinusoidal steady-state we can use the *Fourier transforms* of these equations given by

$$\begin{aligned}
\nabla \cdot \tilde{\mathbf{E}} &= \tilde{\rho}/\epsilon_o, && \textit{Gauss' law} && (1.6)\\
\nabla \cdot \tilde{\mathbf{B}} &= 0, && && (1.7)\\
\nabla \times \tilde{\mathbf{E}} &= -j\omega\tilde{\mathbf{B}}, && \textit{Faraday's law} && (1.8)\\
\nabla \times \tilde{\mathbf{B}} &= \mu_o \tilde{\mathbf{J}} + j\omega\mu_o\epsilon_o\tilde{\mathbf{E}}, && \textit{Ampere's law} && (1.9)
\end{aligned}$$

[1] In MKS system $\mathbf{E}[=]\frac{\mathrm{V}}{\mathrm{m}}$, $\mathbf{B}[=]\mathrm{T}$, $\rho[=]\frac{\mathrm{C}}{\mathrm{m}^3}$, $\mathbf{J}[=]\frac{\mathrm{A}}{\mathrm{m}^2}$ where T stands for teslas and V, A, C, F, and H are volts, amperes, coulombs, farads, and henries, as usual.

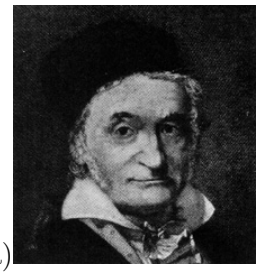
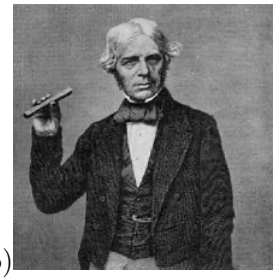
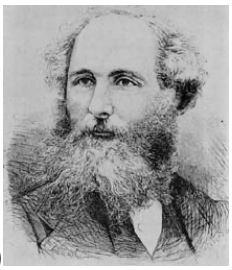
(a)  (b)  (c)  (d) 

Figure 1.1: (a) Johann Carl Friedrich Gauss (1771-1855), (b) Michael Faraday (1791-1867), (c) Andre Marie Ampere (1775-1836), and (d) James Clerk Maxwell (1831-1879)

where Fourier transform pair relations between $\mathbf{E}$ and $\tilde{\mathbf{E}}$ are of the form

$$\tilde{\mathbf{E}}(\mathbf{r},\omega) = \int_{-\infty}^{\infty} \mathbf{E}(\mathbf{r},t)e^{-j\omega t}dt \text{ and } \mathbf{E}(\mathbf{r},t) = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi}\tilde{\mathbf{E}}(\mathbf{r},\omega)e^{j\omega t}.$$

It is useful to think of frequency samples of $\tilde{\mathbf{E}}$ as phasors and thus also think of this form of Maxwell's equations as the "phasor form".

Furthermore, in stationary material media it is convenient to use macroscopic versions of the same equations written as

$$\nabla\cdot\tilde{\mathbf{D}} = \tilde{\rho}, \qquad \textit{Gauss' law} \tag{1.10}$$

$$\nabla\cdot\tilde{\mathbf{B}} = 0, \tag{1.11}$$

$$\nabla\times\tilde{\mathbf{E}} = -j\omega\tilde{\mathbf{B}}, \qquad \textit{Faraday's law} \tag{1.12}$$

$$\nabla\times\tilde{\mathbf{H}} = \tilde{\mathbf{J}} + j\omega\tilde{\mathbf{D}}, \qquad \textit{Ampere's law} \tag{1.13}$$

in terms of a "displacement" field

$$\tilde{\mathbf{D}} = \epsilon\tilde{\mathbf{E}}, \tag{1.14}$$

"magnetizing" field

$$\tilde{\mathbf{H}} = \mu^{-1}\tilde{\mathbf{B}}, \tag{1.15}$$

as well as $\tilde{\rho}$ and $\tilde{\mathbf{J}}$ consisting of only free charge carriers (as opposed to bound charge carriers entrapped in neutral atoms and molecules). Here, in general,

$$\epsilon = \epsilon_r\epsilon_o = (1+\chi_e)\epsilon_o$$

and

$$\mu = \mu_r\mu_o = (1+\chi_m)\mu_o$$

in terms of a suitable dielectric constant ϵ_r and susceptibilities χ_e and χ_m that may in general depend on frequency ω and/or position $\mathbf{x}$. The stationarity clause mentioned above refers to an assumption that bound charge distributions are time independent.

Finally, when phasors $\tilde{\mathbf{E}}$, $\tilde{\mathbf{D}}$ are assumed to have a *plane wave form*

$$\tilde{\mathbf{E}} \propto e^{-j\mathbf{k}\cdot\mathbf{r}},\ \ \tilde{\mathbf{D}} \propto e^{-j\mathbf{k}\cdot\mathbf{r}},\ \text{ etc.},$$

where vector $\mathbf{k}$, called *wavevector*, is specified in units of radians per unit length, the phasor form Maxwell's equations simplify as

$$\begin{aligned} -j\mathbf{k}\cdot\tilde{\mathbf{D}} &= \tilde{\rho}, && \textit{Gauss' law} && (1.16)\\ -j\mathbf{k}\cdot\tilde{\mathbf{B}} &= 0, && && (1.17)\\ \mathbf{k}\times\tilde{\mathbf{E}} &= \omega\tilde{\mathbf{B}}, && \textit{Faraday's law} && (1.18)\\ -j\mathbf{k}\times\tilde{\mathbf{H}} &= \tilde{\mathbf{J}} + j\omega\tilde{\mathbf{D}}, && \textit{Ampere's law} && (1.19) \end{aligned}$$

after replacing all occurrences of ∇ with $-j\mathbf{k}$. Notice that plane-wave form phasors $\tilde{\mathbf{E}}$ and $\tilde{\mathbf{D}} = \epsilon_o(1+\chi_e)\tilde{\mathbf{E}}$ would only be possible (or, permissible) if χ_e were position independent, i.e., when the medium is *homogeneous*. The adaptation of plane-wave solutions to inhomogeneous media — i.e., refraction effects — will be considered later on, starting in Chapter 3.

Exercise 1: Given that $\tilde{\mathbf{E}}(\mathbf{r}) = \mathbf{E_o}e^{-j\mathbf{k}\cdot\mathbf{r}}$, where $\mathbf{E}_o$ and $\mathbf{k}$ are independent of $\mathbf{r}$, show that $\nabla\cdot\tilde{\mathbf{E}} = -j\mathbf{k}\cdot\tilde{\mathbf{E}}$. Hint: write $\mathbf{k} = (k_x, k_y, k_z)$, $\mathbf{r} = (x, y, z)$ to express the dot product $\mathbf{k}\cdot\mathbf{r}$.

Wavevector $\mathbf{k}$ of plane-wave fields $\propto e^{\pm j(\omega t - \mathbf{k}\cdot\mathbf{r})}$ specifies both the propagation direction of the waves and their wavelength. To see this, it is useful to write the propagation vector as $\mathbf{k} = k\hat{k}$, where $\hat{k}$ is a unit vector such that $\hat{k}\cdot\hat{k} = 1$. With that notation, e.g.,

$$e^{j(\omega t - \mathbf{k}\cdot\mathbf{r})} = e^{-jk(s-\frac{\omega}{k}t)}, \tag{1.20}$$

where $s \equiv \hat{k}\cdot\mathbf{r}$ is the projection of position vector $\mathbf{r}$ onto $\hat{k}$. Constant *phase surfaces* of (1.20), identified by the constraint

$$s - \frac{\omega}{k}t = \text{const.}, \tag{1.21}$$

are clearly displaced in time with a velocity $\frac{ds}{dt}$ along $\hat{k}$, obtained by differentiating (1.21) as

$$\frac{ds}{dt} = \frac{\omega}{k} \equiv v_p. \tag{1.22}$$

This velocity v_p is known as *phase velocity* and it plays a central role — along with the *group velocity* v_g to be defined later on — in propagation studies. As for the relation between $\mathbf{k}$ and wavelength λ, it arises from the fact that (1.20) is periodic in s (along $\hat{k}$) at any instant t over displacements $\Delta s = \lambda$ such that $k\lambda = 2\pi$, i.e.,

$$\lambda = \frac{2\pi}{k}.$$

The analogy of this relation between λ and k to the relation

$$T = \frac{2\pi}{\omega}$$

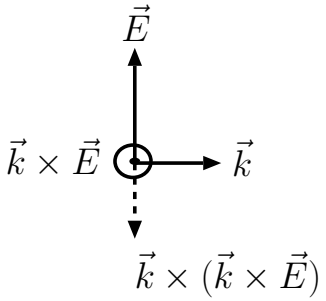


Figure 1.2: When $\mathbf{k}\cdot\mathbf{E} = 0$, $\mathbf{k}$ and $\mathbf{E}$ are perpendicular as shown above, and, furthermore, $\mathbf{k}\times\mathbf{E}$ is perpendicular to both $\mathbf{k}$ and $\mathbf{E}$. Thus, $\mathbf{k}\times(\mathbf{k}\times\mathbf{E})$ points opposite to $\mathbf{E}$ as shown and has a magnitude $\mathbf{k}\cdot\mathbf{k}$ times the magnitude of $\mathbf{E}$. Consequently, $\mathbf{k}\times(\mathbf{k}\times\mathbf{E}) = -(\mathbf{k}\cdot\mathbf{k})\mathbf{E}$ if $\mathbf{k}\cdot\mathbf{E} = 0$. If $\mathbf{k}\cdot\mathbf{E} \neq 0$, then $\mathbf{k}\times(\mathbf{k}\times\mathbf{E}) = (\mathbf{k}\cdot\mathbf{E})\mathbf{k} - (\mathbf{k}\cdot\mathbf{k})\mathbf{E}$.

between wave-period T and wave-frequency ω should be kept in mind.

The plane-wave concepts reviewed above — specifically wavevector $\mathbf{k}$, wavenumber k, phase velocity $\frac{\omega}{k}$, and wavelengths and periods $\lambda = \frac{2\pi}{k}$ and $T = \frac{2\pi}{\omega}$ — are *universal* to all types of wave phenomena in diverse physical settings and for diverse types of *dispersion relations* between the wave frequencies ω and propagation vectors $\mathbf{k}$. Much of our concern in this course will be about finding the correct dispersion relations describing radiowave propagation in Earth's ionosphere.

Plane electromagnetic waves in homogeneous propagation media satisfy the plane wave form of Maxwell's equations given above. To obtain the pertinent dispersion relation we proceed as follows:

According to Faraday's law

$$\mathbf{k}\times\tilde{\mathbf{E}} = \omega\tilde{\mathbf{B}} = \mu_o\omega\tilde{\mathbf{H}} \tag{1.23}$$

whereas according to Ampere's Law we have, with $\mathbf{J} = 0$,

$$-j\mathbf{k}\times\tilde{\mathbf{H}} = j\omega\epsilon_o(1+\chi_e)\tilde{\mathbf{E}}. \tag{1.24}$$

Therefore,

$$\mathbf{k}\times\mathbf{k}\times\tilde{\mathbf{E}} = \mu_o\omega\mathbf{k}\times\tilde{\mathbf{H}} = -\omega^2\mu_o\epsilon_o(1+\chi)\tilde{\mathbf{E}}. \tag{1.25}$$

Now, if $\mathbf{k}$ and $\tilde{\mathbf{E}}$ are orthogonal, that is, if $\mathbf{k}\cdot\tilde{\mathbf{E}} = 0$, then $\mathbf{k}\times\mathbf{k}\times\tilde{\mathbf{E}} = -k^2\tilde{\mathbf{E}}$ (see Figure 1.2), where $k^2 \equiv \mathbf{k}\cdot\mathbf{k}$, and, therefore, for the non-trivial case of $\tilde{\mathbf{E}} \neq 0$,

$$k^2 = \omega^2\mu_o\epsilon_o(1+\chi). \tag{1.26}$$

The above dispersion relation implies that

$$k = \pm\frac{\omega}{c}n, \tag{1.27}$$

where

$$c \equiv \frac{1}{\sqrt{\mu_o \epsilon_o}} \tag{1.28}$$

is the speed of light in free-space (3×10^8 m/s=300 km/ms) and

$$n \equiv \sqrt{1+\chi} \tag{1.29}$$

is known as *refractive index* of the medium for *transverse* electromagnetic (TEM) waves satisfying the $\mathbf{k} \cdot \tilde{\mathbf{E}} = 0$ condition.

It is possible to envision the existence of some type of *longitudinal waves* in addition to the TEM wave identified above. For such waves $\mathbf{k}$ and $\tilde{\mathbf{E}}$ are parallel, and, therefore, $\mathbf{k} \times \tilde{\mathbf{E}} = 0$. Therefore, according to (1.25) such waves can exist in the absence of a conduction current $\tilde{\mathbf{J}}$ only if $n^2 = 1 + \chi = 0$; for $\chi \geq 0$, that is not a possibility.

If the propagation medium contains free charge carriers a conduction current $\tilde{\mathbf{J}} = \sigma \tilde{\mathbf{E}}$ may flow in response wave field $\tilde{\mathbf{E}}$, where σ is said to be the medium *conductivity*. Allowing for a non-zero $\tilde{\mathbf{J}} = \sigma \tilde{\mathbf{E}}$, Ampere's Law used above is modified as:

$$-j\mathbf{k} \times \tilde{\mathbf{H}} = j\omega\epsilon_o(1+\chi)\tilde{\mathbf{E}} \;\; \Rightarrow \;\; \sigma\mathbf{E} + j\omega\epsilon_o(1+\chi)\tilde{\mathbf{E}} = j\omega\epsilon_o(1+\chi+\frac{\sigma}{j\omega\epsilon_o})\tilde{\mathbf{E}}.$$

The modification amounts to invoking an *effective dielectric constant*

$$1+\chi \;\; \Rightarrow \;\; 1+\chi+\frac{\sigma}{j\omega\epsilon_o}$$

that contains a conductivity term, and an overall dispersion relation modified as

$$k^2 = \omega^2 \mu_o \epsilon_o (1+\chi+\frac{\sigma}{j\omega\epsilon_o}). \tag{1.30}$$

Clearly, we can still write

$$k = \pm \frac{\omega}{c} n, \tag{1.31}$$

with

$$n = \sqrt{1+\chi+\frac{\sigma}{j\omega\epsilon_o}}, \tag{1.32}$$

but we need to take into account the real and imaginary parts of n which may result as a consequence of a real valued σ. Actually, such concerns would also be applicable to complex valued χ which arise in non-perfect dielectrics.

Writing a complex valued n as $n = n' - jn"$, it follows that $k = \frac{\omega}{c} n = k' - jk"$, so that

$$e^{j(\omega t - \mathbf{k}\cdot\mathbf{r})} = e^{j(\omega t - k's + jk"s)} = e^{-k"s} e^{j(\omega t - k's)} = e^{-k"s} e^{-jk'(s - \frac{\omega}{k'}t)}. \tag{1.33}$$

This indicates, in analogy to (1.20), that the real part k' of k must be used in TEM wave phase velocity and wavelength calculations — i.e., $v_p = \frac{\omega}{k'}$ and $\lambda = \frac{2\pi}{k'}$ — whereas the imaginary part $k" \propto n"$ describes wave damping with propagation distance s. In fact

$$\delta \equiv \frac{1}{k"} = \frac{c}{\omega n"}$$

is a characteristic distance, known as *penetration depth*, over which the wave amplitude $e^{-k"s}$ is reduced by one e-fold (i.e., from 1 to e^{-1}).

Longitudinal waves satisfying the condition $n^2 = 0$ also become possible with the modified refractive index $n = \sqrt{1 + \chi + \frac{\sigma}{j\omega\epsilon_o}}$ for a medium with an imaginary conductivity σ. The characteristics of such modes in ionospheric plasmas — known as plasma oscillations — will be examined in the next chapter. Note that with an imaginary σ — implying a non-Ohmic material — n should be real and therefore no wave damping would be expected in that case.

Returning back to TEM wave solutions and re-writing Faraday's and Ampere's Laws in plane-wave form as

$$\mathbf{k} \times \tilde{\mathbf{E}} = \omega\mu_o \tilde{\mathbf{H}} \quad \Rightarrow \quad \tilde{\mathbf{H}} = \frac{(\omega/c)n}{\omega\mu_o}\hat{k} \times \tilde{\mathbf{E}} = \frac{n}{c\mu_o}\hat{k} \times \tilde{\mathbf{E}}$$

$$-\mathbf{k} \times \tilde{\mathbf{H}} = \omega\epsilon_o(1 + \chi + \frac{\sigma}{j\omega\epsilon_o})\tilde{\mathbf{E}} = \omega\epsilon_o n^2 \tilde{\mathbf{E}} \quad \Rightarrow \quad \tilde{\mathbf{E}} = \frac{(\omega/c)n}{\omega\epsilon_o n^2}\tilde{\mathbf{H}} \times \hat{k} = \frac{1}{c\epsilon_o n}\tilde{\mathbf{H}} \times \hat{k}$$

Hence, we can write

$$\tilde{\mathbf{H}} = \frac{\hat{k} \times \tilde{\mathbf{E}}}{\eta} \quad \text{and} \quad \tilde{\mathbf{E}} = \eta\tilde{\mathbf{H}} \times \hat{k}, \tag{1.34}$$

where

$$\eta = \frac{1}{c\epsilon_o n} = \frac{c\mu_o}{n} \quad \Rightarrow \quad \eta = \frac{\sqrt{\mu_o/\epsilon_o}}{n} = \frac{\eta_o}{n} \tag{1.35}$$

is the *intrinsic impedance* of the medium and $\eta_o \equiv \sqrt{\frac{\mu_o}{\epsilon_o}} \approx 120\pi\,\Omega$ is the intrinsic impedance of free space. The so-called polarization relations (1.34) clearly show the reason why TEM waves are labeled as TEM.

We close this section with a sequence of examples. Starting in these examples, we drop the tildes we have been placing on phasor quantities in order to distinguish them from their time-domain counterparts. Since we will mostly be working with phasors this is hardly necessary and in the long run it should be obvious to recognize time and frequency domain quantities from the context.

Example 1: Consider a plane TEM wave in free space with a field phasor $\mathbf{E}(0) = \hat{y}3$ V/m at $\mathbf{r} = 0$, propagation direction $\hat{k} = \frac{\hat{x}+\hat{z}}{\sqrt{2}}$ and a wave frequency of $\omega = 2\pi \times 10^8$ rad/s. Determine the field phasor $\mathbf{E}(\mathbf{r})$ and the wave field $\mathbf{E}(\mathbf{r}, t)$.

Solution: Since the propagation medium is free space $\epsilon = \epsilon_o$ and $\mu = \mu_o$ and therefore

$$k = \omega\sqrt{\mu\epsilon} = 2\pi \times 10^8\sqrt{\mu_o\epsilon_o} = \frac{2\pi \times 10^8}{c} = \frac{2\pi \times 10^8}{3 \times 10^8} = \frac{2\pi}{3}\ \frac{\text{rad}}{\text{m}}.$$

Hence

$$\mathbf{k} = k\hat{k} = \frac{2\pi}{3}\frac{\hat{x}+\hat{z}}{\sqrt{2}} = \frac{\sqrt{2}\pi}{3}(\hat{x}+\hat{z})\ \frac{\text{rad}}{\text{m}}$$

and

$$\mathbf{E}(\mathbf{r}) = \mathbf{E}(0)e^{-j\mathbf{k}\cdot\mathbf{r}} = \hat{y}3e^{-j\frac{\sqrt{2}\pi}{3}(x+z)}\ \frac{\text{V}}{\text{m}}.$$

The wave field is

$$\mathbf{E}(\mathbf{r},t) = Re\{\mathbf{E}(\mathbf{r})e^{j\omega t}\} = \hat{y}3\cos(2\pi\times 10^8 t - \frac{\sqrt{2}\pi}{3}(x+z))\ \frac{\mathrm{V}}{\mathrm{m}}.$$

Example 2: Determine the $\mathbf{H}(\mathbf{r})$ phasor that accompanies $\mathbf{E}(\mathbf{r})$ of Example 1.

Solution: Since the propagation medium is free space the intrinsic impedance

$$\eta = \sqrt{\frac{\mu}{\epsilon}} = \sqrt{\frac{\mu_o}{\epsilon_o}} \approx 120\pi\ \Omega.$$

Thus,

$$\begin{aligned}\mathbf{H}(\mathbf{r}) &= \frac{\hat{k}\times\mathbf{E}(\mathbf{r})}{\eta} = \frac{(\hat{x}+\hat{z})\times\hat{y}3e^{-j\frac{\sqrt{2}\pi}{3}(x+z)}}{\sqrt{2}120\pi} \\ &= \frac{\hat{z}-\hat{x}}{\sqrt{2}}\frac{1}{40\pi}e^{-j\frac{\sqrt{2}\pi}{3}(x+z)}\ \frac{\mathrm{A}}{\mathrm{m}}.\end{aligned}$$

Example 3: What is the wavelength λ and Hertzian frequency $f = \frac{\omega}{2\pi}$ of the plane wave field discussed in Example 1?

Solution: Since $k = \frac{2\pi}{3}$ rad/m from Example 1 it follows that $\lambda = \frac{2\pi}{k} = 3$ m. Also $f = \frac{\omega}{2\pi} = \frac{2\pi\times 10^8}{2\pi} = 10^8$ Hz, or, equivalently, 100 MHz.

Example 4: A plane TEM wave in a medium with $n = 2$ has the phasor

$$\mathbf{E}(\mathbf{r}) = 4e^{-j2\pi y}\hat{x}\ \frac{\mathrm{V}}{\mathrm{m}}.$$

Assuming that $\mu_r = 1$, determine the wave frequency $f = \frac{\omega}{2\pi}$, and the corresponding phasor $\mathbf{H}(\mathbf{r})$. In which direction the wave is propagating.

Solution: Clearly $k = k_y = 2\pi$ rad/m, while $k_x = k_z = 0$. Therefore, $k = \frac{\omega}{c}n = 2\pi$, and with $n = 2$, we have $\frac{2\pi f}{c}2 = 2\pi$, or $f = \frac{c}{2} = 1.5\times 10^8 = 150\times 10^6$ Hz. The wave frequency is therefore 150 MHz. Since $\mu_r = 1$ and $n = 2$, $n = \sqrt{\mu_r\epsilon_r}$ implies that $\epsilon_r = 4$. Therefore, $\eta = \sqrt{\frac{\mu}{\epsilon}} = \frac{\eta_o}{2} \approx 60\pi\,\Omega$. Also, the wave propagation direction in the wave normal $\hat{k} = \hat{y}$. Thus,

$$\mathbf{H} = \hat{k}\times\mathbf{E}/\eta \approx \hat{y}\times\hat{x}\frac{4e^{-2\pi y}}{60\pi} = -\frac{1}{15\pi}e^{-j2\pi y}\hat{z}\ \frac{\mathrm{A}}{\mathrm{m}}.$$

Exercise 2: Repeat the example above for a medium with $\mu_r = \sqrt{2}$ and $n = 2$.

1.2 Poynting vector

Differencing $\mathbf{H}\cdot\nabla\times\mathbf{E}$ and $\mathbf{E}\cdot\nabla\times\mathbf{H}$ and using the vector identity

$$\nabla\cdot(\mathbf{E}\times\mathbf{H}) = \mathbf{H}\cdot\nabla\times\mathbf{E} - \mathbf{E}\cdot\nabla\times\mathbf{H}$$

we can obtain the following equality from Faraday's and Ampere's laws:

$$\underbrace{\mathbf{E}\cdot\frac{\partial\mathbf{D}}{\partial t}+\mathbf{H}\cdot\frac{\partial\mathbf{B}}{\partial t}}_{\text{Storage term}} \;+\; \overbrace{\mathbf{J}\cdot\mathbf{E}}^{\text{Absorbtion term}} \;+\; \underbrace{\nabla\cdot(\mathbf{E}\times\mathbf{H})}_{\text{Transport term}} \;=\; 0$$

This result is known as *Poynting theorem* and its term-by-term interpretation leads to the recognition that vector $\mathbf{E}\times\mathbf{H}$ represents the direction and *area density* of power transported by the fields $\mathbf{E}$ and $\mathbf{H}$. We refer to $\mathbf{E}\times\mathbf{H}$ as *Poynting vector* which has the units of $\frac{\mathrm{V}}{\mathrm{m}}\frac{\mathrm{A}}{\mathrm{m}}=\frac{\mathrm{W}}{\mathrm{m}^2}$, power per unit area.

For sinusoidally varying fields with phasors $\mathbf{E}(\mathbf{r})$ and $\mathbf{H}(\mathbf{r})$, the Poynting vector

$$\begin{aligned}\mathbf{E}(\mathbf{r},t)\times\mathbf{H}(\mathbf{r},t) &= \mathrm{Re}\{\mathbf{E}(\mathbf{r})e^{j\omega t}\}\times\mathrm{Re}\{\mathbf{H}(\mathbf{r})e^{j\omega t}\}\\ &= \frac{\mathbf{E}e^{j\omega t}+\mathbf{E}^*e^{-j\omega t}}{2}\times\frac{\mathbf{H}e^{j\omega t}+\mathbf{H}^*e^{-j\omega t}}{2}\\ &= \frac{\mathbf{E}\times\mathbf{H}^*+\mathbf{E}^*\times\mathbf{H}}{4}+\frac{\mathbf{E}\times\mathbf{H}e^{j2\omega t}+\mathbf{E}^*\times\mathbf{H}^*e^{-j2\omega t}}{4},\end{aligned}$$

in which the second term on the right is time-varying has a zero mean. Therefore the time-independent first term, which can be denoted as

$$\mathbf{S}(\mathbf{r})\equiv\frac{1}{2}\mathrm{Re}\{\mathbf{E}(\mathbf{r})\times\mathbf{H}^*(\mathbf{r})\},$$

is the time-average value of Poynting vector of sinusoidal fields with $\mathbf{E}(\mathbf{r})$ and $\mathbf{H}(\mathbf{r})$ phasors. For plane TEM waves with mutually perpendicular phasors $\mathbf{E}$ and $\mathbf{H}=\frac{\hat{k}\times\mathbf{E}}{\eta}$, vector $\mathbf{S}$ points in wave normal direction $\hat{k}$ and its magnitude is

$$S=\frac{|\mathbf{E}|^2}{2\eta},$$

where

$$|\mathbf{E}|^2\equiv\mathbf{E}\cdot\mathbf{E}^*.$$

We will refer to S above as *Poynting flux* or *power density*. Also, we will always compute the magnitude square of complex valued vectors such as $\mathbf{E}$ as shown above, i.e., in terms of conjugated dot product $\mathbf{E}\cdot\mathbf{E}^*$— note that for complex vectors such as $\mathbf{E}=\hat{x}-j\hat{y}$ the dot product without conjugation $\mathbf{E}\cdot\mathbf{E}$ vanishes, failing to provide the correct measure of the phasor magnitude $|\mathbf{E}|$.

Example 1: Poynting flux of the field phasor

$$\mathbf{E}_1=\hat{y}2e^{-j3\hat{x}\cdot\mathbf{r}}$$

in free-space is

$$S=\frac{2^2}{2\times120\pi}=\frac{1}{60\pi}\;\frac{\mathrm{W}}{\mathrm{m}^2}$$

and the flux is in wavenormal direction $\hat{x}$.

Exercise 1: Determine the Poynting flux of plane wave $\mathbf{E}_2 = \hat{y}2e^{-j3\hat{x}\cdot\mathbf{r}} + \hat{z}2je^{-j3\hat{x}\cdot\mathbf{r}}$ propagating in a medium with $\mu_r = 1$ and $n = 2$.

Exercise 2: Show that the time-average $\frac{1}{T}\int_0^T dt\mathbf{J}(\mathbf{r},t)\cdot\mathbf{E}(\mathbf{r},t)$ of absorbed power $\mathbf{J}(\mathbf{r},t)\cdot\mathbf{E}(\mathbf{r},t)$ can be computed as

$$\frac{1}{2}Re\{\mathbf{J}(\mathbf{r})\cdot\mathbf{E}^*(\mathbf{r})\}$$

using $\mathbf{J}(\mathbf{r})$ and $\mathbf{E}(\mathbf{r})$ phasors when $\mathbf{J}(\mathbf{r},t)$ and $\mathbf{E}(\mathbf{r},t)$ vary with time as co-sinusoidals. Hint: the procedure is similar to the derivation of $\mathbf{S}(\mathbf{r}) \equiv \frac{1}{2}\mathrm{Re}\{\mathbf{E}(\mathbf{r})\times\mathbf{H}^*(\mathbf{r})\}$.

1.3 Wave polarizations and orthogonality

For a plane TEM wave

$$\mathbf{E}(\mathbf{r}) = \mathbf{E}_o e^{-j\mathbf{k}\cdot\mathbf{r}} = \mathbf{E}_o e^{-jk\hat{k}\cdot\mathbf{r}}$$

specified in a homogeneous and isotropic medium there is no constraint on $\mathbf{E}_o \equiv \mathbf{E}(0)$ other than

$$\hat{k}\cdot\mathbf{E}_o = 0.$$

However, different choices for $\mathbf{E}_o$ can be classified in terms of different types of wave polarizations as discussed in this section.

Field phasors

$$\mathbf{E}_1 = \hat{y}2e^{-j3\hat{x}\cdot\mathbf{r}} \text{ and } \mathbf{E}_2 = \hat{z}5je^{-j2\pi y}$$

for instance describe *linearly polarized* plane waves with $\hat{y}$ and $\hat{z}$ directed polarizations, respectively. Phasor

$$\mathbf{E}_3 = (\hat{x}+\hat{y})e^{j\pi z}$$

is also linear polarized, but in direction $\frac{\hat{x}+\hat{y}}{\sqrt{2}}$. Linear polarized wave phasors can always be expressed as a product of a real valued vector (e.g., $\hat{z}$, $\hat{x}+\hat{y}$, etc.) and a scalar function (e.g., $5je^{-j2\pi y}$ as in the case of phasor $\mathbf{E}_2$ above) which is of course always complex.

Linear polarization is a special case of *elliptic polarization* which includes all possible polarization types as special cases. Another special case of elliptic polarization is *circular polarization*. Circular polarized waves are the superposition of orthogonal linears with equal magnitudes but a 90^o phase difference as in the case of

$$\mathbf{E}_4 = 8(\hat{x}-j\hat{y})e^{-j3z}.$$

This particular superposition of $\hat{x}$ and $\hat{y}$ polarized linears is circular because the tip of field vector $\mathrm{Re}\{\mathbf{E}_4 e^{j\omega t}\}$ traces out a circle as a function of time, pointing out first in $\hat{x}$ direction at $t = 0$ and then in $\hat{y}$ direction a quarter of a wave period $T = \frac{2\pi}{\omega}$ later and so on. The phasor

$$\mathbf{E}_5 = 4(\hat{x}+j\hat{y})e^{-j3z}$$

is also circular polarized, but the corresponding field vector rotates[2] in the opposite direction as compared to the field described by $\mathrm{Re}\{\mathbf{E}_4 e^{j\omega t}\}$. Circular polarized waves

[2]Linear polarized vectors only shrink and grow, circular polarized vectors only rotate, while elliptical polarized vectors shrink and grow as they rotate. Thus, linear and circular polarizations are special cases of elliptical polarizations.

are in general labeled as *right-* or *left-handed* depending on propagation and rotation directions—if you point your right thumb in the propagation direction and see that your fingers curl in the rotation direction then the field is right-handed. Phasor $\mathbf{E}_4 = 8(\hat{x} - j\hat{y})e^{-j3z}$ is right-handed since it represents propagation in $\hat{z}$-direction and the rotation sense from $\hat{x}$ toward $\hat{y}$ is the direction the fingers curl when the *right* thumb is directed along $\hat{z}$. Phasor $\mathbf{E}_5 = 4(\hat{x}+j\hat{y})e^{-j3z}$ is left-handed, but phasor $\mathbf{E}_6 = 4(\hat{x}+j\hat{y})e^{j3z}$ is again right-handed.

As the examples above illustrate, circularly polarized phasors can always be expressed as some scalar times a *complex unit vector* $\hat{c}$ with the structure

$$\hat{c} = \frac{\hat{a} \pm j\hat{b}}{\sqrt{2}}$$

where $\hat{a} \cdot \hat{b} = 0$, $\hat{a} \cdot \hat{a} = 1$,and $\hat{b} \cdot \hat{b} = 1$. Note that $\hat{c} \cdot \hat{c}^* = 1$ but $\hat{c} \cdot \hat{c} = 0$. Since for a *real unit vector* $\hat{r}$ (such as $\hat{r} = \hat{x}$) both $\hat{r} \cdot \hat{r} = 1$ and $\hat{r} \cdot \hat{r}^* = 1$ are true, the generalized condition for $\hat{u}$ to be a unit vector must therefore be $\hat{u} \cdot \hat{u}^* = 1$. Likewise, the generalized definition of *orthogonality* between a pair of vectors $\mathbf{U}$ and $\mathbf{V}$ is $\mathbf{U} \cdot \mathbf{V}^* = 0$. According to this definition, linear polarized phasors

$$\hat{x}e^{-j6z} \text{ and } 2j\hat{y}e^{-j6z}$$

describe orthogonal wave fields (as expected). But so do circular polarized phasors

$$(\hat{x} - j\hat{y})e^{-j6z} \text{ and } 2(\hat{x} + j\hat{y})e^{-j6z}$$

since $(\hat{x} - j\hat{y}) \cdot (\hat{x} + j\hat{y})^* = (\hat{x} - j\hat{y}) \cdot (\hat{x} - j\hat{y}) = \hat{x} \cdot \hat{x} - \hat{y} \cdot \hat{y} = 0$. Circular polarized orthogonal field vectors always rotate in opposite directions as a function of time.

The Poynting flux formula

$$S = \frac{|\mathbf{E}|^2}{2\eta}$$

from last section is valid for all plane TEM waves whatever the polarization of $\mathbf{E}$ may be, as long as we remember that $|\mathbf{E}|^2$ stands for $\mathbf{E} \cdot \mathbf{E}^*$. For instance, the flux of circular polarized $\mathbf{E}_1 = (\hat{x} - j\hat{y})e^{-j3z}$ is twice the flux of linear polarized $\mathbf{E}_2 = \hat{x}e^{-j3z}$ because $\mathbf{E}_1 \cdot \mathbf{E}_1^* = (\hat{x} - j\hat{y}) \cdot (\hat{x} - j\hat{y})^* = 2$ while $\mathbf{E}_2 \cdot \mathbf{E}_2^* = \hat{x} \cdot \hat{x}^* = 1$. The factor of 2 difference makes physical sense because $\mathbf{E}_1$ maintains a $1 \frac{\text{V}}{\text{m}}$ magnitude electric field at every instant at every location while the strength of $\mathbf{E}_2$ fluctuates between maximum and minimum values of 1 and $0 \frac{\text{V}}{\text{m}}$ in time and space—hence $\mathbf{E}_1$ naturally carries more flux than $\mathbf{E}_2$.

Exercise 1: Suppose $\mathbf{E} = \mathbf{E_1} + \mathbf{E}_2$ where $\mathbf{E}_1$ and $\mathbf{E}_2$ are orthogonal in the sense described above. Show that $|\mathbf{E}|^2 = |\mathbf{E}_1|^2 + |\mathbf{E}_2|^2$.

In view of the discussion above it should be a straightforward matter to recognize linear as well as circular polarized field phasors. If a given field phasor is neither linear (easy to decide) nor circular (also easy to decide) it must be by default *elliptical* polarized. We will not examine elliptical polarized fields in detail here except for noting here that any elliptical polarized field phasor can be expressed as a superposition of an orthogonal pair of linears or circulars.

1.4 Non-uniform plane waves

Consider an arbitrary homogeneous and isotropic propagation medium where plane wave dispersion relation is

$$\mathbf{k} \cdot \mathbf{k} = \omega^2 \mu \tilde{\epsilon}$$

with

$$\tilde{\epsilon} \equiv \epsilon - j\frac{\sigma}{\omega} = \epsilon_o(1 + \chi - j\frac{\sigma}{\epsilon_o \omega}).$$

If we assume $\mathbf{k} = \hat{k}k$, where $\hat{k}$ is a real unit vector, we obtain from the constraint above the dispersion relation

$$k = \omega\sqrt{\mu\tilde{\epsilon}}$$

of uniform plane waves. However, with complex $\hat{c}$, such that $\hat{c} \cdot \hat{c} = 1$, the assumption $\mathbf{k} = \hat{c}k$ leads to non-uniform plane waves discussed next:

With no loss of generality a complex vector $\hat{c}$ can be expressed as $\hat{c} = \mathbf{R} + j\mathbf{I}$ in terms of arbitrary real vectors $\mathbf{R}$ and $\mathbf{I}$. The constraint $\hat{c} \cdot \hat{c} = 1$ then implies

$$(\mathbf{R} + j\mathbf{I}) \cdot (\mathbf{R} + j\mathbf{I}) = \mathbf{R} \cdot \mathbf{R} - \mathbf{I} \cdot \mathbf{I} + j2\mathbf{R} \cdot \mathbf{I} = 1 \quad \Rightarrow \quad R^2 - I^2 = 1 \ \text{ and } \ \mathbf{R} \cdot \mathbf{I} = 0,$$

where $R^2 \equiv \mathbf{R} \cdot \mathbf{R}$ and $I^2 \equiv \mathbf{I} \cdot \mathbf{I}$. Hence, plane waves with wave vectors

$$\mathbf{k} = \omega\sqrt{\mu\tilde{\epsilon}}(\mathbf{R} + j\mathbf{I})$$

are compatible with Maxwell's equations so long as $\mathbf{R} \cdot \mathbf{I} = 0$ and $R^2 - I^2 = 1$. A special case corresponding to $\mathbf{R} = \hat{k}$ and $\mathbf{I} = 0$ gives us back the familiar uniform plane waves, but all other cases correspond to *non-uniform* plane waves.

Consider, for instance, the case $\sigma = 0$, so that $\mathbf{k} = \omega\sqrt{\mu\epsilon}(\mathbf{R} + j\mathbf{I})$ and

$$\mathbf{E}_o e^{-j\mathbf{k}\cdot\mathbf{r}} = e^{\omega\sqrt{\mu\epsilon}\mathbf{I}\cdot\mathbf{r}}\mathbf{E}_o e^{-j\omega\sqrt{\mu\epsilon}\mathbf{R}\cdot\mathbf{r}}.$$

Clearly, the corresponding wave field propagates in direction of $\mathbf{R}$ with a propagation velocity

$$v_p = \frac{1}{R\sqrt{\mu\epsilon}}$$

while the field amplitude varies exponentially in the orthogonal direction of vector $\mathbf{I}$. Both directions are furthermore orthogonal to $\mathbf{E}_o$ since $\mathbf{k} \cdot \mathbf{E}$ condition still applies. An example of non-uniform plane waves is the evanescent waves formed in events of total internal reflection examined in the next section.

1.5 Plane wave reflection coefficients

The *differential form* of Maxwell's equations reviewed in this chapter requires the fields as well as medium parameters such as χ to be continuous functions of position. When

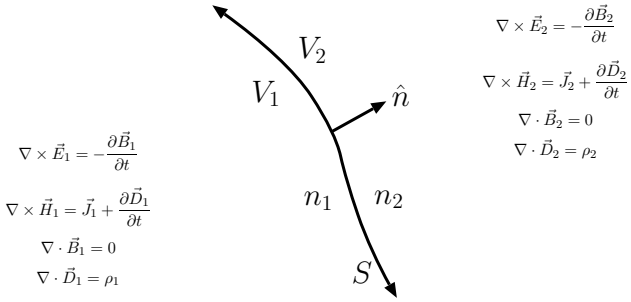


Figure 1.3: Unit vector $\hat{n}$ on surface S is *normal* to S and points from volume V_1 into volume V_2.

the medium parameters are discontinuous, Maxwell's equations must be replaced at discontinuity surfaces with *boundary condition* equations

$$\begin{aligned} \hat{n} \cdot (\mathbf{D}_2 - \mathbf{D}_1) &= \rho_s \\ \hat{n} \cdot (\mathbf{B}_2 - \mathbf{B}_1) &= 0 \\ \hat{n} \times (\mathbf{E}_2 - \mathbf{E}_1) &= 0 \\ \hat{n} \times (\mathbf{H}_2 - \mathbf{H}_1) &= \mathbf{J}_s, \end{aligned}$$

as explained in more detail in the caption of Figure 1.3. In essence, the boundary condition equations say that the *tangential* component of $\mathbf{E}$ and *normal* component of $\mathbf{B}$ have to be continuous across *any* surface; tangential $\mathbf{H}$ and normal $\mathbf{D}$, on the other hand, can *jump* across a surface if the surface holds a non-zero current and charge densities $\mathbf{J}_s$ and ρ_s, respectively. Such surface densities can arise naturally only on the surfaces of perfect conductors, a useful idealization of highly conducting metals.

The plane TEM waves described in this chapter are only valid in homogeneous volumes of space where the refractive index n does not vary with position. If two such regions with refractive constant indices n_1 and n_2 are adjacent to one another across some planar interface, then plane wave solutions valid for the two regions individually can be matched at the interface to describe wave reflection and transmission (see Figure 1.4) phenomena as discussed below:

Horizontal (TE) polarization:

Figure 1.4a depicts a horizontal polarized plane wave

$$\mathbf{E}_i = \hat{y} e^{-j(k_x x - k_z z)},$$

accompanied by a $\mathbf{H}_i$, propagating towards an interface on $z = 0$ plane separating distinct dielectrics with refractive indices n_1 and n_2. As a result of the discontinuity across the interface, $\mathbf{E}_i$ launches the reflected and transmitted fields

$$\mathbf{E}_r = R\hat{y} e^{-j(k_x x + k_z z)} \quad \text{and} \quad \mathbf{E}_t = T\hat{y} e^{-j(k_{xt} x - k_{zt} z)}$$

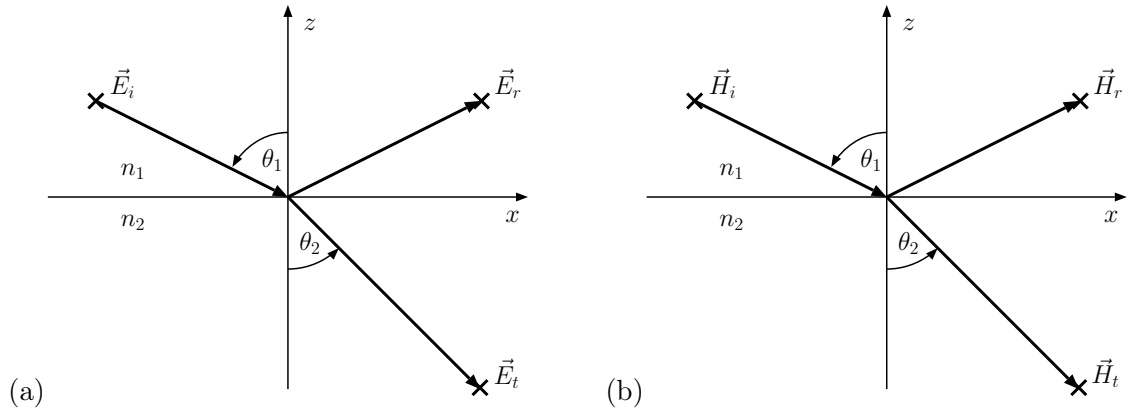


Figure 1.4: (a) Horizontal polarized TEM wave (TE) being reflected from $z = 0$ interface, and (b) vertically polarized TEM wave (TM) being reflected from $z = 0$ interface

accompanied by $\mathbf{H}_r$ and $\mathbf{H}_t$, respectively. All three wave fields share a common $k_x = k_{xt}$ because otherwise tangential components of total $\mathbf{E}$ and $\mathbf{H}$ on each side of the interface could not be matched for all x. Such a matching is required by Maxwell's boundary condition equations across any interface between any two dielectrics and leads to Snell's Law

$$k_x = k_{xt} \quad \Rightarrow \quad n_1 \sin\theta_1 = n_2 \sin\theta_2$$

since $k_x = k \sin\theta$ and $k = \frac{\omega}{c} n$ in terms of refractive index n.

Matching the tangential $\mathbf{E}$ components as required, we obtain a constraint

$$1 + R = T$$

between reflection coefficient R and transmission coefficient T describing $\mathbf{E}_r$ and $\mathbf{E}_t$, respectively.

Also using $\mathbf{H} = \frac{\hat{k} \times \mathbf{E}}{\eta}$ and $\eta = \eta_o / n$ (we will assume $\mu_t = \mu_i = \mu_o$, for simplicity), we find out that at $x = 0$ and $z = 0$

$$\hat{x} \cdot \mathbf{H}_i = -\frac{n_1 \cos\theta_1}{\eta_o}, \quad \hat{x} \cdot \mathbf{H}_r = R\frac{n_1 \cos\theta_1}{\eta_o}, \quad \text{and} \quad \hat{x} \cdot \mathbf{H}_t = -T\frac{n_2 \cos\theta_2}{\eta_o}.$$

Thus a second constraint obtained by matching the tangential $\mathbf{H}$ at the interface is

$$n_1 \cos\theta_1 (1 - R) = n_2 \cos\theta_2 T.$$

Solving the two constraints for R, we obtain

$$R = \frac{n_1 \cos\theta_1 - n_2 \cos\theta_2}{n_1 \cos\theta_1 + n_2 \cos\theta_2} \equiv R_h = \frac{E_{yr}}{E_{yi}} = \frac{H_{zr}}{H_{zi}},$$

where R_h is the *reflection coefficient for horizontal polarization.* The corresponding transmission coefficient is $T_h = 1 + R_h = \frac{E_{yt}}{E_{yi}}$.

If μ on both sides of the interface differ from μ_o, then the reflection coefficient is expressed in terms of $\eta_{1,2} = \sqrt{\mu_{1,2}/\epsilon_{1,2}}$ as

$$R_h = \frac{\eta_2 \cos\theta_1 - \eta_1 \cos\theta_2}{\eta_2 \cos\theta_1 + \eta_1 \cos\theta_2} = \frac{E_{yr}}{E_{yi}} = \frac{H_{zr}}{H_{zi}}$$

and the transmission coefficient is still $T_h = 1 + R_h$.

Vertical (TM) polarization:

Figure 1.4b depicts a horizontal polarized plane wave

$$\mathbf{H}_i = \hat{y} e^{-j(k_x x - k_z z)},$$

accompanied by a $\mathbf{E}_i$, propagating towards an interface on $z = 0$ plane separating distinct dielectrics with refractive indices n_i and n_t. To obtain the corresponding reflection coefficient $R_v \equiv H_{yr}/H_{yi}$ we take advantage of a "duality principle" that states that transformation of specific solutions of source free Maxwell's equations according to the recipe $\mathbf{E} \to \mathbf{H}$, $\mathbf{H} \to -\mathbf{E}$, and $\epsilon \leftrightarrow \mu$ furnishes additional solutions of the same set of equations. Applying this to R_h above expressed in terms of $\eta_{1,2} = \sqrt{\mu_{1,2}/\epsilon_{1,2}} \to 1/\eta_{1,2}$ we get

$$R_h = \frac{\eta_2 \cos\theta_1 - \eta_1 \cos\theta_2}{\eta_2 \cos\theta_1 + \eta_1 \cos\theta_2} = \frac{E_{yr}}{E_{yi}} = \frac{H_{zr}}{H_{zi}} \quad \to \quad R_v = \frac{\eta_1 \cos\theta_1 - \eta_2 \cos\theta_2}{\eta_1 \cos\theta_1 + \eta_2 \cos\theta_2} = \frac{H_{yr}}{H_{yi}} = \frac{E_{zr}}{E_{zi}}$$

as well as $T_v = 1 + R_v$ for *vertically polarized* waves. Furthermore, if $\mu_t = \mu_i = \mu_o$, then we have

$$R_v = \frac{n_2 \cos\theta_1 - n_1 \cos\theta_2}{n_2 \cos\theta_1 + n_1 \cos\theta_2} = \frac{H_{yr}}{H_{yi}} = \frac{E_{zr}}{E_{zi}}.$$

In summary then, we have

$$R_h = \frac{n_1 \cos\theta_1 - n_2 \cos\theta_2}{n_1 \cos\theta_1 + n_2 \cos\theta_2} = \frac{E_{yr}}{E_{yi}}$$

and

$$R_v = \frac{n_2 \cos\theta_1 - n_1 \cos\theta_2}{n_{2t} \cos\theta_1 + n_1 \cos\theta_2} = \frac{E_{zr}}{E_{zi}},$$

with

$$\cos\theta_2 = \sqrt{1 - \sin^2\theta_2} = \sqrt{1 - \frac{n_1^2}{n_2^2}\sin^2\theta_1},$$

which are useful in modeling the behavior of radiowave links operated near a reflecting ground plane. In lossy dielectrics with $\mu = \mu_o$, the results obtained above will apply using the complex refractive index

$$n = \sqrt{\epsilon_r - j\frac{\sigma}{\omega\epsilon_o}} = \sqrt{1 + \chi - j\frac{\sigma}{\omega\epsilon_o}}.$$

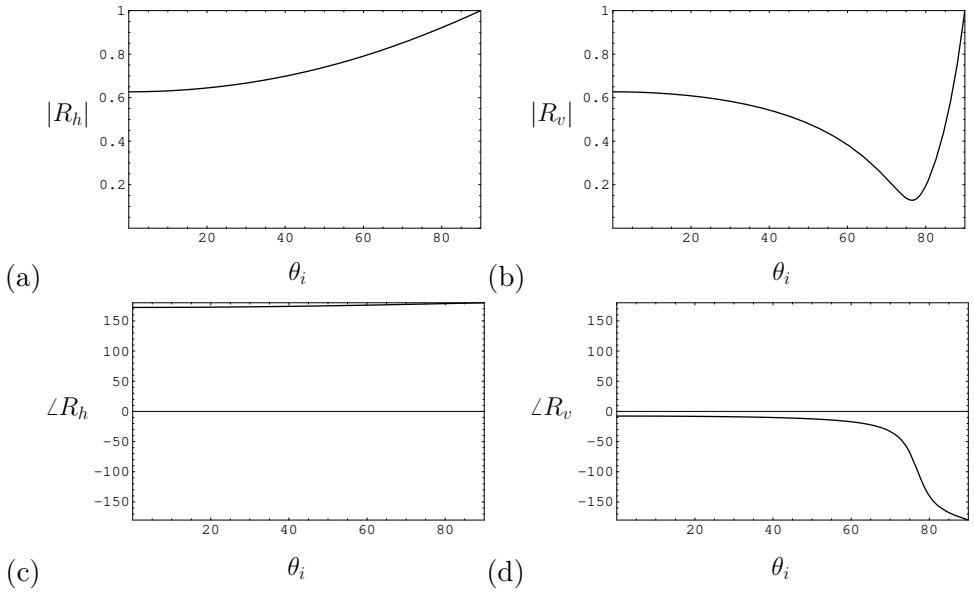


Figure 1.5: Magnitudes and angles of R_v and R_h at $f = 10$ MHz for $n_i = 1$, $\epsilon_t = 15\epsilon_o$, and $\sigma_t = 5 \times 10^{-3}\,\frac{\mathrm{S}}{\mathrm{m}}$. Note that R_h, $R_v \to -1$ as $\theta_i \to 90^o$. Angle θ_i where $|R_v|$ minimizes is known as *pseudo-Brewster angle*.

For instance, for $n_1 = 1$, $\epsilon_2 = 15\epsilon_o$, $\sigma_2 = 5 \times 10^{-3}\,\frac{\mathrm{S}}{\mathrm{m}}$, and $f = \frac{\omega}{2\pi} = 10$ MHz, R_h and R_v vary with θ_1 as shown in Figure 1.5.

Note that $R_h \to -1$ and $R_v \to 1$ for all angles θ_1 when the transmission medium is a perfect conductor, i.e., $|n_2| \to \infty$. This fact should be contrasted with the fact that $R_{h,v} \to -1$ as $\theta_i \to 90^o$ for all n except when $|n| \to \infty$.

For applications where $n_1 = 1$ for $z > 0$, $n_2 = n = \sqrt{\epsilon_r - j\frac{\sigma}{\omega\epsilon_o}}$ in $z < 0$ — that is, for air over a conducting ground — and with $\cos\theta_2 = \sqrt{1 - \sin^2\theta_1/n^2}$, we have

$$R_h = \frac{\cos\theta - n\sqrt{1 - \sin^2\theta/n^2}}{\cos\theta + n\sqrt{1 - \sin^2\theta/n^2}} \to \frac{\cos\theta - n}{\cos\theta + n} \text{ for } |n| \gg 1$$

after dropping the subscript on θ_1 — we will make use of this form of R_h in radiowave link modeling with antennas located in air in $z > 0$ half space above a reflecting ground with a refractive index n below $z = 0$. Likewise, under the same conditions,

$$R_v = \frac{n\cos\theta - \sqrt{1 - \sin^2\theta/n^2}}{n\cos\theta + \sqrt{1 - \sin^2\theta/n^2}} \to \frac{n\cos\theta - 1}{n\cos\theta + 1} \text{ for } |n| \gg 1.$$

2 Radiation, antennas, links, imaging

In this chapter we will examine the radiation process and transmission/reception properties of antennas. Radiation — by accelerated charges and time-varying currents — is fundamental not only to the operation of antennas but also to radiowave diffraction and scattering effects which will be discussed in subsequent chapters. The chapter includes a discussion of radiowave links as well as related concepts of antenna noise and antenna arrays. It concludes with a discussion of radio imaging.

2.1 Radiation and transmitting antennas

2.1.1 Vector and scalar potentials

We *define* a vector field $\mathbf{A}(\mathbf{r},t)$ with a curl

$$\nabla \times \mathbf{A} \equiv \mathbf{B}$$

— and a divergence $\nabla \cdot \mathbf{A}$ to be specified later on — where vector $\mathbf{B}$ denotes the magnetic flux density. Vector $\mathbf{A}(\mathbf{r},t)$, to be called *vector potential*, is useful for determining the radiation solutions of Maxwell's equations driven by *specified* current distributions $\mathbf{J}(\mathbf{r},t)$.

Substituting $\mathbf{B} = \nabla \times \mathbf{A}$ in Faraday's law, the latter can be rearranged as

$$\nabla \times (\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t}) = 0.$$

This has the form of a vector identity $\nabla \times (\nabla \Phi) = 0$ which is valid for any scalar function $\Phi(\mathbf{r},t)$. We can set $\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t} = -\nabla \Phi$ and obtain

$$\mathbf{E} = -\nabla \Phi - \frac{\partial \mathbf{A}}{\partial t},$$

which reduces to $\mathbf{E} = -\nabla \Phi$ in the special case of $\frac{\partial \mathbf{A}}{\partial t} = 0$; we call Φ the *scalar potential*, a generalization of electrostatic potential Φ that satisfies $\mathbf{E} = -\nabla \Phi$ when $\frac{\partial \mathbf{A}}{\partial t} = 0$.

We next substitute $\nabla \times \mathbf{A}$ and $-\nabla \Phi - \frac{\partial \mathbf{A}}{\partial \mathbf{t}}$ for $\mathbf{B}$ and $\mathbf{E}$ in Gauss' and Ampere's laws to obtain

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_o} \quad \Rightarrow \quad -\nabla^2 \Phi - \frac{\partial}{\partial t} \nabla \cdot \mathbf{A} = \frac{\rho}{\epsilon_o},$$

and

$$\nabla \times \frac{\mathbf{B}}{\mu_o} = \mathbf{J} + \epsilon_o \frac{\partial \mathbf{E}}{\partial t} \quad \Rightarrow \quad \nabla \times (\nabla \times \mathbf{A}) = \mu_o \mathbf{J} - \mu_o \epsilon_o \frac{\partial}{\partial t} (\nabla \Phi + \frac{\partial \mathbf{A}}{\partial t}).$$

Using the vector identity $\nabla \times (\nabla \times \mathbf{A}) = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A}$,

$$-\nabla^2 \Phi - \frac{\partial}{\partial t} \nabla \cdot \mathbf{A} = \frac{\rho}{\epsilon_o} \quad \text{and} \quad \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} = \mu_o \mathbf{J} - \mu_o \epsilon_o \frac{\partial}{\partial t}(\nabla \Phi + \frac{\partial \mathbf{A}}{\partial t}),$$

and thus *choosing*[1]

$$\nabla \cdot \mathbf{A} = -\mu_o \epsilon_o \frac{\partial \Phi}{\partial t},$$

we obtain the decoupled radiation equations

$$\nabla^2 \Phi - \mu_o \epsilon_o \frac{\partial^2 \Phi}{\partial t^2} = -\frac{\rho}{\epsilon_o} \quad \text{and} \quad \nabla^2 \mathbf{A} - \mu_o \epsilon_o \frac{\partial^2 \mathbf{A}}{\partial t^2} = -\mu_o \mathbf{J}.$$

The equations above for Φ and $\mathbf{A}$ are of the same form — known as *Helmholtz equation* — and they can be solved independently using identical procedures. We will next discuss the solution for Φ and then obtain by analogy the solution for $\mathbf{A}$ in terms of $\mathbf{J}$. Once $\mathbf{A}$ is known $\mathbf{B} = \nabla \times \mathbf{A}$, and $\mathbf{E}$ can be readily obtained from $\mathbf{B}$ as we will see.

2.1.2 Solutions of Helmholtz equation

For a time-independent charge density

$$\rho = q\delta(x)\delta(y)\delta(z)$$

the scalar potential equation simplifies to *Poisson's equation*

$$\nabla^2 \Phi = -\frac{q}{\epsilon_o} \delta(x)\delta(y)\delta(z)$$

describing the electrostatic potential Φ of point charge q at the origin. From *Coulomb's law* we know that the solution is

$$\Phi = \frac{q}{4\pi\epsilon_o r} \quad \text{with} \quad r = \sqrt{x^2 + y^2 + z^2},$$

which is singular at the origin and its Laplacian $\nabla^2 \Phi$ vanishes for all other locations.

For a sinusoidally varying $\rho(\mathbf{r}, t) = q\delta(x)\delta(y)\delta(z)\cos(\omega t)$ with a phasor

$$\rho(\mathbf{r}) = q\delta(x)\delta(y)\delta(z),$$

the phasor form scalar potential equation is

$$\nabla^2 \Phi(\mathbf{r}) + \omega^2 \mu_o \epsilon_o \Phi(\mathbf{r}) = -\frac{q}{\epsilon_o} \delta(x)\delta(y)\delta(z).$$

[1] Known as *Lorenz gauge*, named after Danish physicist Ludwig Lorenz (1829-1891), completes the specification of the vector potential. The gauge impacts $\mathbf{A}$, but not $\nabla \times \mathbf{A}$, which is of main interest here. A common mistake in the literature is to credit Lorenz gauge to Dutch physicist and Nobel Prize winner Hendrik A. Lorentz (1853-1928) after whom Lorentz force and Lorentz-Fitzgerald transformations are (appropriately) named.

For $\omega = 0$ this is the same as Poisson's equation and therefore its solution for $\omega = 0$ is once again the electrostatic solution

$$\Phi(\mathbf{r}) = \frac{q}{4\pi\epsilon_o r} \equiv \Phi_s.$$

However, with arbitrary ω the solution $\Phi(\mathbf{r})$ is required to have the Laplacian

$$\nabla^2\Phi(\mathbf{r}) = -\omega^2\mu_o\epsilon_o\Phi(\mathbf{r})$$

away from the origin ($r > 0$) and also reduce to Φ_s in $\omega \to 0$ limit. The solution of interest[2] is

$$\Phi(\mathbf{r}) = \Phi_s e^{-jkr}$$

with $k \equiv \omega\sqrt{\mu_o\epsilon_o}$. This can be easily verified by using spherical coordinates (see, for instance, *Rao*) to compute the Laplacian

$$\nabla^2\frac{e^{-jkr}}{r} = \frac{1}{r^2}\frac{\partial}{\partial r}(r^2\frac{\partial}{\partial r}(\frac{e^{-jkr}}{r})) = \frac{1}{r^2}\frac{\partial}{\partial r}(-jkre^{-jkr} - e^{-jkr}) = -k^2\frac{e^{-jkr}}{r}.$$

Thus, with the postulated solution $\Phi(\mathbf{r}) \propto \frac{e^{-jkr}}{r}$, the Laplacian

$$\nabla^2\Phi(\mathbf{r}) = -k^2\Phi(\mathbf{r}) = -\omega^2\mu_o\epsilon_o\Phi(\mathbf{r}),$$

as required.

In summary, then, the solution of the differential equation

$$\nabla^2\Phi(\mathbf{r}) + \omega^2\mu_o\epsilon_o\Phi(\mathbf{r}) = -\frac{\rho(r)}{\epsilon_o}$$

with an input $\rho(\mathbf{r}) = q\delta(x)\delta(y)\delta(z)$ is

$$\Phi(\mathbf{r}) = \frac{qe^{-jkr}}{4\pi\epsilon_o r} \equiv qg(\mathbf{r}),$$

where $g(\mathbf{r})$ is, in effect, the *impulse response* or *Green's function* of the differential equation. The solution of the differential equation with an arbitrary input $\rho(\mathbf{r}) = \rho(x, y, z)$ is therefore obtained with convolution

$$\Phi(\mathbf{r}) = \Phi(x, y, z) = \int \rho(x', y', z')g(x - x', y - y', z - z')dx'dy'dz',$$

i.e.,

$$\Phi(\mathbf{r}) = \int_{V'} \frac{\rho(x', y', z')}{\epsilon_o}\frac{e^{-jk|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r} - \mathbf{r}'|}dV',$$

[2] $k = -\omega\sqrt{\mu_o\epsilon_o}$ also yields a valid solution but it will be disregarded here because the fields obtained with $k = -\omega\sqrt{\mu_o\epsilon_o}$ correspond to energy transport towards the origin rather than away from origin. The solution of interest in radiation studies is therefore obtained with $k = \omega\sqrt{\mu_o\epsilon_o}$.

where

$$|\mathbf{r}-\mathbf{r}'| = \sqrt{(x-x')^2+(y-y')^2+(z-z')^2},$$

symbol $\int_{V'}$ stands for a volume integral over the primed coordinates x', y', and z', and $dV' \equiv dx'dy'dz'$.

Finally, our result above implies (by analogy) that for the vector potential, in phasor form[3],

$$\nabla^2\mathbf{A}(\mathbf{r}) + \omega^2\mu_o\epsilon_o\mathbf{A}(\mathbf{r}) = -\mu_o\mathbf{J}(\mathbf{r}) \quad \Rightarrow \quad \mathbf{A}(\mathbf{r}) = \int_{V'} \mu_o\mathbf{J}(\mathbf{r}')\frac{e^{-jk|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|}dV'.$$

2.1.3 Radiation fields

The next step in deriving the electromagnetic fields generated by a specified current phasor $\mathbf{J}(\mathbf{r})$ is to take the curl of vector potential solution obtained above — $\nabla\times\mathbf{A}(\mathbf{r})$ gives $\mathbf{B}(\mathbf{r})$, and subsequently $\mathbf{E}(\mathbf{r})$ is obtained from the curl of $\mathbf{H}(\mathbf{r}) = \mu_o^{-1}\mathbf{B}(\mathbf{r})$ using Ampere's law.

The procedure just described produces multi-term expressions for both $\mathbf{E}(\mathbf{r})$ and $\mathbf{H}(\mathbf{r})$ which include leading terms $\propto r^{-1}$ and addition to terms $\mathrm{O}(r^{-2})$. Only the leading terms of $\mathbf{E}(\mathbf{r})$ and $\mathbf{H}(\mathbf{r})$ which are $\mathrm{O}(r^{-1})$ contribute to the time average Poynting flux $\frac{1}{2}\mathrm{Re}\{\mathbf{E}(\mathbf{r})\times\mathbf{H}^*(\mathbf{r})\}$ and dominate at large distances r away from the origin where $\mathbf{J}(\mathbf{r})$ is assumed to be concentrated — it is these leading terms of $\mathbf{E}(\mathbf{r})$ and $\mathbf{H}(\mathbf{r})$, known as *radiation fields* of $\mathbf{J}(\mathbf{r})$, which is of interest for us in communication and remote sensing applications.

The radiation fields can also be obtained directly from $\mathbf{A}(\mathbf{r})$ by first approximating $\mathbf{A}(\mathbf{r})$ for $r \equiv |\mathbf{r}| \gg |\mathbf{r}'|$, and then applying curl to the approximation. The geometry pertinent to the far-field approximation of $\mathbf{A}(\mathbf{r})$ is shown in Figure 2.1.

As shown in the caption of Figure 2.1 $|\mathbf{r}-\mathbf{r}'| \approx \hat{r}\cdot(\mathbf{r}-\mathbf{r}')$ when $r \gg L$, where L is the characteristic size of the region near the origin where source current $\mathbf{J}(\mathbf{r})$. Hence for $r \gg L$ we can write

$$\mathbf{A}(\mathbf{r}) = \int_{V'} \mu_o\mathbf{J}(\mathbf{r}')\frac{e^{-jk|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|}dV' \approx \frac{\mu_o}{4\pi r}\int_{V'}\mathbf{J}(\mathbf{r}')e^{-jk\hat{r}\cdot(\mathbf{r}-\mathbf{r}')}dV'.$$

Next, when we take the curl of the approximation on the right using the identity

$$\nabla\times(\mathbf{V}s) = s\nabla\times\mathbf{V} + \nabla s\times\mathbf{V},$$

we will ignore the slow varying r^{-1} term out in the front and effectively replace

$$\nabla\times[\mathbf{J}(\mathbf{r}')e^{-jk\hat{r}\cdot(\mathbf{r}-\mathbf{r}')}]$$

[3]Which in turn implies a time-domain solution

$$\mathbf{A}(\mathbf{r},t) = \int_{V'}\frac{\mu_o\mathbf{J}(\mathbf{r}',t-\frac{|\mathbf{r}-\mathbf{r}'|}{c})}{4\pi|\mathbf{r}-\mathbf{r}'|}dV',$$

which can be confirmed by showing that its Fourier transform yields the phasor solution above under the constraint $\frac{\omega}{c}=k$. This result is known as *retarded potential solution*. The term retarded refers to the time delay $\frac{|\mathbf{r}-\mathbf{r}'|}{c}$ in the response of $A(\mathbf{r},t)$ to $\mathbf{J}(\mathbf{r}',t)$.

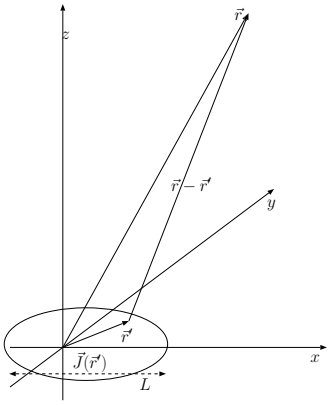


Figure 2.1: Radiation geometry pertaining a source current distribution $\mathbf{J}(\mathbf{r})$ near the origin. Note that for $r \equiv |\mathbf{r}| \gg L$, $|\mathbf{r}-\mathbf{r}'| \approx r - \hat{r}\cdot\mathbf{r}' = \hat{r}\cdot(\mathbf{r}-\mathbf{r}')$, where the unit vector $\hat{r} \equiv \frac{\mathbf{r}}{r}$.

by

$$\nabla e^{-jk\hat{r}\cdot(\mathbf{r}-\mathbf{r}')} \times \mathbf{J}(\mathbf{r}') \approx e^{-jk\hat{r}\cdot(\mathbf{r}-\mathbf{r}')}(-jk\hat{r}) \times \mathbf{J}(\mathbf{r}'),$$

retaining only the leading term on the right which dominates in $r \gg L$ limit. Hence,

$$\mathbf{B}(\mathbf{r}) = \mu_o \mathbf{H}(\mathbf{r}) = \nabla \times \mathbf{A}(\mathbf{r}) \approx \frac{\mu_o}{4\pi r}\int_{V'} e^{-jk\hat{r}\cdot(\mathbf{r}-\mathbf{r}')}(-jk\hat{r}) \times \mathbf{J}(\mathbf{r}')dV',$$

and, likewise, after taking the curl again, we obtain

$$\frac{\mathbf{E}(\mathbf{r})}{j\omega\epsilon_o\mu_o} = \nabla \times \mathbf{B}(\mathbf{r}) \approx \frac{\mu_o}{4\pi r}\int_{V'} e^{-jk\hat{r}\cdot(\mathbf{r}-\mathbf{r}')}(-jk\hat{r}) \times (-jk\hat{r}) \times \mathbf{J}(\mathbf{r}')dV'.$$

The last two expressions provide the radiation fields of the source $\mathbf{J}(\mathbf{r})$ discussed above. The fields can be expressed more succinctly as

$$\mathbf{H}(\mathbf{r}) = \frac{\hat{r} \times \mathbf{E}(\mathbf{r})}{\eta_o}$$

and

$$\mathbf{E}(\mathbf{r}) = j\eta_o I_o k \mathbf{f}(\hat{r})\frac{e^{-jkr}}{4\pi r},$$

where function

$$\mathbf{f}(\hat{r}) \equiv \hat{r} \times \hat{r} \times \int_{V'} \frac{\mathbf{J}(\mathbf{r}')}{I_o} e^{jk\hat{r}\cdot\mathbf{r}'}dV'$$

is a vector with the dimensions of length, provided that the normalization constant I_o introduced in the equations above is defined in units of electrical current (amps). In practice, I_o is often chosen as the input current of an antenna system described by current distribution $\mathbf{J}(\mathbf{r})$ at its input terminals — in that case $\mathbf{f}(\hat{r})$ is said to be *effective length* of the antenna.

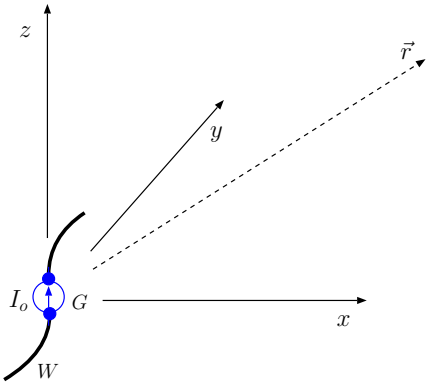


Figure 2.2: A wire antenna of located near the origin. An ideal current source I_o applied across a small gap (G) excites a current $I(\mathbf{r})$ along the antenna wires (W) in a reference direction parallel to infinitesimals $d\mathbf{l}$ along the wire.

The radiation fields $\mathbf{E}(\mathbf{r})$ and $\mathbf{H}(\mathbf{r})$ above describe a *spherical wave propagation* in radial direction $\hat{r}$. The mutual perpendicularity of vectors $\mathbf{E}(\mathbf{r})$, $\mathbf{H}(\mathbf{r})$, and $\hat{r}$ indicates that the wave has a TEM nature. Also, it can be easily shown that the time-average Poynting vector

$$\mathbf{S} = \frac{1}{2}\mathrm{Re}\{\mathbf{E}(\mathbf{r}) \times \mathbf{H}^*(\mathbf{r})\} = \frac{|\mathbf{E}(\mathbf{r})|^2}{2\eta_o}\hat{r},$$

indicating power flow in the direction of propagation as expected. Spherical wave field magnitudes vary as r^{-1} (remember that faster decaying field components were dropped from radiation fields) and the corresponding Poynting flux $S \equiv |\mathbf{S}|$ varies as r^{-2}.

2.1.4 Radiation fields of thin wire antennas

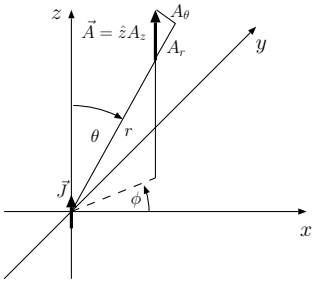
For antennas where currents $\mathbf{J}(\mathbf{r})$ are constrained to flow along thin wires as shown in Figure 2.2 the radiation field can be expressed as

$$\mathbf{E}(\mathbf{r}) = j\eta_o I_o k \mathbf{f}(\hat{r}) \frac{e^{-jkr}}{4\pi r}$$

where I_o is the antenna excitation current applied across the antenna terminals and vector

$$\mathbf{f}(\hat{r}) = \hat{r} \times \hat{r} \times \int_L d\mathbf{l} \frac{I(\mathbf{r}')}{I_o} e^{jk\hat{r}\cdot\mathbf{r}'}$$

is the *effective length* of the antenna. Note that the effective length determines both the *polarization* and angular *beam pattern* of the radiated field $\mathbf{E}(\mathbf{r})$. By beam pattern we refer to the variation of $|\mathbf{E}(\mathbf{r})|^2 \propto |\mathbf{f}(\hat{r})|^2$ as a function of radiation direction $\hat{r}$. We will next work out the effective length for some common types of thin wire antennas and discuss their beam patterns.

(a)  (b) 

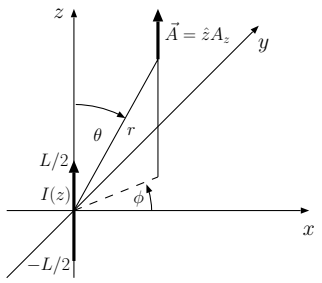
(c) 

Figure 2.3: (a) Heinrich Hertz (1857-1894), (b) a z-polarized Hertzian dipole at the origin, and (c) a z-polarized dipole antenna of length L centered about the origin. — Unit vectors $\hat{r}$, $\hat{\theta}$, and $\hat{\phi}$ in *spherical coordinates* point in the directions of increasing $r \equiv \sqrt{x^2+y^2+z^2}$, $\theta \equiv \cos^{-1}(\frac{z}{r})$, and $\phi \equiv \angle x + jy$, respectively, and they can be expressed in terms of Cartesian coordinates x, y, and z as follows: $\hat{r} = \frac{(x,y,z)}{r}$, $\hat{\phi} = \frac{(-y,x,0)}{\sqrt{x^2+y^2}}$, and $\hat{\theta} = \frac{(xz,yz,z^2-r^2)}{r\sqrt{x^2+y^2}}$. We refer to θ and ϕ as *zenith angle* and *azimuth*, respectively, and to r as radial distance, or, *radius*.

Hertzian dipole:

For a conceptual line antenna of an infinitesimal length dl

$$\mathbf{f}(\hat{r}) = \hat{r} \times \hat{r} \times \int_L d\mathbf{l} \frac{I(\mathbf{r}')}{I_o} e^{jk\hat{r}\cdot\mathbf{r}'} \quad \Rightarrow \quad \mathbf{f}(\hat{r}) = \hat{r} \times \hat{r} \times d\mathbf{l}$$

where the direction of infinitesimal vector $d\mathbf{l}$ is the reference direction of current phasor I_o on the element. This radiation element is known as *Hertzian dipole*. For a $\hat{z}$-polarized Hertzian dipole, that is, for $d\mathbf{l} = \hat{z}dz$,

$$\begin{aligned} \mathbf{f}(\hat{r}) &\equiv \hat{r} \times \hat{r} \times d\mathbf{l} = \hat{r} \times \hat{r} \times \hat{z}dz \\ &= -\hat{r} \times \hat{\phi} \sin\theta dz = \hat{\theta} \sin\theta dz, \end{aligned}$$

where $\hat{\phi} \equiv \hat{z} \times \hat{r}$ and $\hat{\theta} \equiv \hat{\phi} \times \hat{r}$ are as shown and discussed in the caption of Figure 2.3. Hence, the radiation field of a $\hat{z}$-polarized Hertzian dipole is the $\hat{\theta}$-polarized

$$\mathbf{E}(\mathbf{r}) = j\eta_o I_o k dl \sin\theta \frac{e^{-jkr}}{4\pi r} \hat{\theta}.$$

Radiation fields of finite sized antennas can in principle be calculated by using weighted superpositions of the radiation fields of shifted and rotated Hertzian dipoles.

Short dipole:

Figure 2.3c depicts a $\hat{z}$-polarized dipole antenna of some finite length L. Such an antenna is said to be a *short-dipole* if $L \ll \lambda$ and its normalized current distribution has a

triangular form

$$\frac{I(z)}{I_o} = \triangle(\frac{z}{L}).$$

For a short-dipole the effective length is

$$\mathbf{f}(\hat{r}) = \hat{r} \times \hat{r} \times \hat{z} \int dz' \triangle(\frac{z'}{L}) e^{jk\hat{r}\cdot\mathbf{r}'} \approx \hat{r} \times \hat{r} \times \hat{z} \int dz' \triangle(\frac{z'}{L}) = \frac{L}{2} \sin\theta\hat{\theta}$$

by analogy to a $\hat{z}$-polarized Hertzian dipole.

Half-wave dipole:

For a $\hat{z}$-polarized half-wave or $\frac{\lambda}{2}$-dipole, $L = \frac{\lambda}{2}$ *and* current distribution

$$\frac{I(z)}{I(0)} = \cos(kz),$$

so that

$$\begin{aligned} \int dz' \frac{I(z')}{I_i} e^{jk\hat{r}\cdot\mathbf{r}'} &= \int_{-\lambda/4=-\pi/2k}^{\lambda/4=\pi/2k} \cos(kz') e^{jk\cos\theta z'} dz' \\ &= \int_{-\pi/2k}^{\pi/2k} \cos(kz)\cos(k\cos\theta z) dz = \frac{\lambda}{\pi} \frac{\cos(\frac{1}{2}\pi\cos\theta)}{\sin^2\theta}. \end{aligned}$$

Thus, the effective length

$$\mathbf{f}(\hat{r}) \equiv \hat{r} \times \hat{r} \times \hat{z} \int dz' \frac{I(z')}{I_i} e^{jk\hat{r}\cdot\mathbf{r}'} = \frac{\lambda}{\pi} \frac{\cos(\frac{1}{2}\pi\cos\theta)}{\sin\theta} \hat{\theta}.$$

Figure 2.4a shows the polar diagrams of $|\mathbf{f}(\hat{r})|$ for a half-wave dipole and a short-dipole with $L = \frac{\lambda}{20}$, showing that the half-wave dipole will be radiating a higher intensity field than the short dipole (assuming equal input currents I_o). In Figure 2.4b the beam patterns $|\mathbf{f}(\hat{r})|^2$ of the same two antennas are shown after each one has been normalized by its peak value $|\mathbf{f}(\hat{r})|^2_{max}$. Clearly, the beam patterns of the two dipoles are quite similar — both dipoles have radiation nulls along the along the dipole axes ($\theta = 0^o$ and 180^o) and for both dipoles beam pattern peaks in the *broadside* direction $\theta = 90^o$.

A brief comment on antenna current distributions: In order to derive the effective length $\mathbf{f}$ for short- and $\frac{\lambda}{2}$-dipoles we postulated the form of $\frac{I(z)}{I(0)}$. The postulated current distributions are consistent with the following "rules of thumb" derived from measurements as well as numerical calculations: For thin wire antennas $I(z) \to 0$ at the endpoints while from the endpoints towards the antenna center $I(z)$ grows like a cosine function with a wavelength $\lambda = \frac{2\pi}{k}$; if $L \ll \lambda$, then cosine-like variation is practically linear. In numerical models $I(z)$ is solved for by enforcing tangential boundary conditions on the antenna surface for generated fields — including the storage components neglected in our radiation field results.

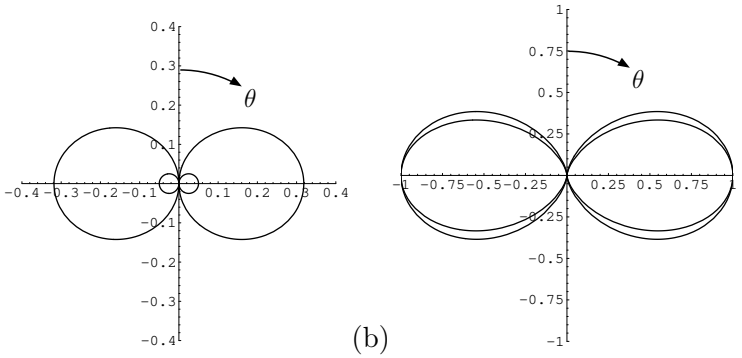


Figure 2.4: (a) Polar plots of $|\mathbf{f}|$ for a $\frac{\lambda}{2}$-dipole and a short-dipole with $L = \frac{\lambda}{20}$ and $\lambda = 1$ m, and (b) corresponding self-normalized beam patterns $|\mathbf{f}|^2/|\mathbf{f}|^2_{max}$. The outer curve in (a) is the $\frac{\lambda}{2}$-dipole while in (b) it is the short-dipole (with a slightly fatter beam pattern).

Small loops:

A small loop antenna is a wire loop of some radius $a \ll \lambda$ carrying an oscillating current with a constant distribution I_o around the loop. Thus, for a loop with a $\hat{\phi}$ directed current I_o centered about the origin on $z = 0$ plane, we find

$$\mathbf{f}(\hat{r}) = \hat{r} \times \hat{r} \times \int_L d\mathbf{l} \frac{I(\mathbf{r}')}{I_o} e^{jk\hat{r}\cdot\mathbf{r}'} = \hat{r} \times \hat{r} \times \int_0^{2\pi} \hat{\phi}' a \; d\phi' \; e^{jka\hat{r}\cdot\hat{r}'} = -jk\pi a^2 \sin\theta \hat{\phi}$$

(as verified in homework). The corresponding radiation field can be written as

$$\mathbf{E}(\mathbf{r}) = \eta_o I_o k^2 A_l \sin\theta \frac{e^{-jkr}}{4\pi r} \hat{\phi},$$

where $A_l = \pi a^2$ is the loop area. Note that the loop antenna has the same beam pattern $|\mathbf{f}|^2 \propto |\sin\theta|^2$ as a $\hat{z}$-polarized short dipole (but its polarization in $\hat{\phi}$ direction is different from the polarization $\hat{\theta}$ of the dipole antenna).

2.2 Antenna parameters — radiation resistance, gain, directivity, beam-width, solid angle

Interactions of antennas with lumped circuit elements can be described in terms of terminal voltage and currents V_o and I_o — see Figure 2.5 depicting a circuit including a transmitting antenna shown as a two-terminal element — and an impedance

$$Z_a = \frac{V_o}{I_o}$$

known as antenna impedance. An antenna with a terminal current I_o and impedance Z_a will absorb an average power of

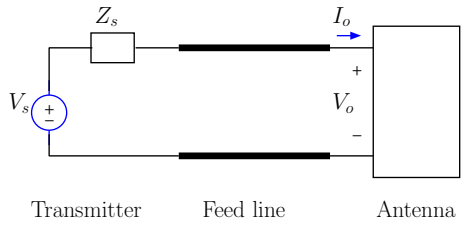


Figure 2.5: A simple circuit with a transmitting antenna. The feed line is typically a low-loss transmission line many wavelengths long. The transmitter supplies the average power P_t radiated by the antenna.

$$P = \frac{1}{2}\mathrm{Re}\{V_o I_o^*\} = \frac{1}{2}|I_o|^2 \mathrm{Re}\{Z_a\}$$

from its circuit. A fraction of P will account for Ohmic losses in antenna wires and the remainder P_t will account for the average power radiated by the antenna in the form of electromagnetic waves. In an ideal lossless antenna

$$P_t = \frac{1}{2}|I_o|^2 R_{rad},$$

where $R_{rad} = \mathrm{Re}\{Z_a\}$ is known as *radiation resistance*. For lossy antennas R_{rad} is some fraction of $\mathrm{Re}\{Z_a\}$.

An *isotropic radiator* by definition radiates equal power in all directions and maintains an average outgoing (i.e., $\hat{r}$ directed) Poynting flux of

$$S = \frac{P_t}{4\pi r^2}$$

$\frac{\mathrm{W}}{\mathrm{m}^2}$. For all other antennas, however,

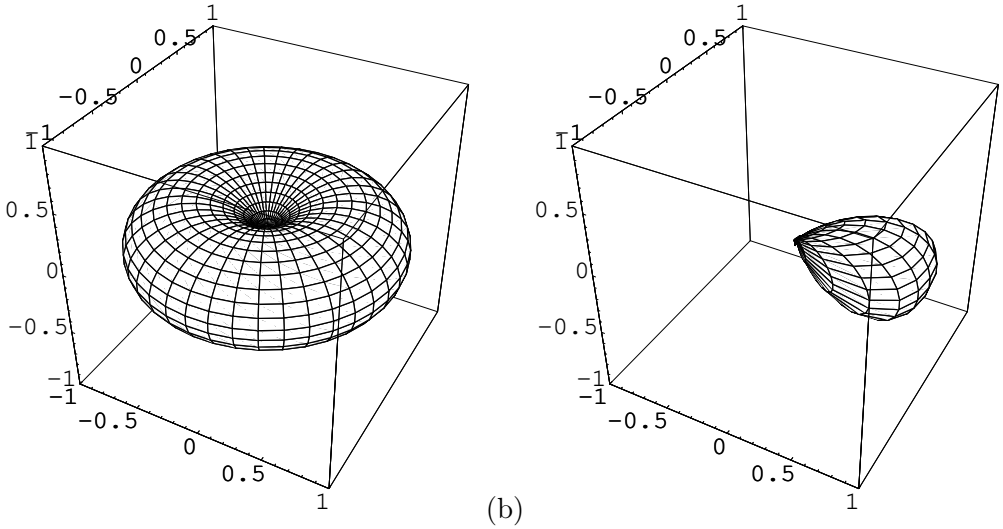
$$S = \frac{P_t}{4\pi r^2} G,$$

where $G = G(\theta, \phi)$ is called *antenna gain* over isotropic radiator.

The gain function $G(\theta, \phi)$ satisfies an important constraint imposed by energy conservation: For antennas radiating into lossless media the surface integral of their Poynting flux S at any fixed r has to match the radiated power P_t. The described surface integral operator is[4]

$$r^2 \int d\Omega \equiv r^2 \int_{\theta=0}^{\pi} d\theta \sin\theta \int_{\phi=0}^{2\pi} d\phi,$$

[4]The surface of a sphere of radius r consists of infinitesimal rectangular patches of dimensions $r d\theta$ and $r \sin\theta d\phi$. Hence an infinitesimal area element is $r^2 d\theta \sin\theta d\phi = r^2 d\Omega$, where $d\Omega \equiv d\theta \sin\theta d\phi$ is known as infinitesimal *solid angle* and is measured in units of steradians.



(a) (b)

Figure 2.6: Antenna beam pattern $\frac{G}{D} = \frac{|\mathbf{f}|^2}{|\mathbf{f}|^2_{max}}$ for (a) a z-polarized $\frac{\lambda}{2}$-dipole, and (b) an 8-element endfire antenna array with $\frac{\lambda}{4}$ element spacings and $\alpha = -90^o$ phase increments (see Section 6). The beam pattern depicted in (b) has a larger directivity D and smaller solid angle $\Omega_o = \frac{4\pi}{D}$ than the $\frac{\lambda}{2}$-dipole.

and, therefore,

$$r^2 \int d\Omega S = r^2 \int d\Omega \frac{P_t}{4\pi r^2} G = \frac{P_t}{4\pi} \int d\Omega G(\theta, \phi) = P_t.$$

This result implies that

$$\int d\Omega G(\theta, \phi) = 4\pi,$$

which is satisfied by the gain of any antenna. In the following discussion we will refer to the operation $\int d\Omega$ as *solid angle integration*.

The maximum value of gain $G = G(\theta, \phi)$ is defined to be the antenna *directivity* D, and satisfies a relation

$$\int d\Omega \frac{G}{D} = \frac{4\pi}{D} \equiv \Omega_o,$$

where Ω_o is said to be the *solid angle* of the gain function, or, the antenna beam. Radiation resistance R_{rad}, antenna gain G, directivity D, and beam solid angle Ω_o constitute the main system properties of radiating antennas. We will next obtain explicit expressions for R_{rad} and G in terms of effective length $\mathbf{f}$.

Since the Poynting flux

$$S = \frac{|\mathbf{E}|^2}{2\eta_o},$$

it follows that

$$\frac{|\mathbf{E}|^2}{2\eta_o} = \frac{P_t}{4\pi r^2} G = \frac{\frac{1}{2}|I_o|^2 R_{rad}}{4\pi r^2} G \quad \Rightarrow \quad G R_{rad} = \frac{4\pi r^2}{\eta_o} \frac{|\mathbf{E}|^2}{|I_o|^2}.$$

Taking the solid angle integral of both sides of this result we get

$$R_{rad} = \frac{1}{\eta_o}\int d\Omega \frac{r^2|\mathbf{E}|^2}{|I_o|^2} \quad \text{and subsequently} \quad G = \frac{4\pi|\mathbf{E}|^2}{\int d\Omega|\mathbf{E}|^2}.$$

With

$$\mathbf{E} = j\eta_o I_o k \mathbf{f}\frac{e^{-jkr}}{4\pi r} \quad \Rightarrow \quad |\mathbf{E}|^2 = \eta_o^2|I_o|^2k^2|\mathbf{f}|^2\frac{1}{(4\pi r)^2},$$

and, therefore,

$$R_{rad} = \frac{\eta_o}{4\lambda^2}\int d\Omega|\mathbf{f}|^2 \quad \text{and} \quad G = \frac{4\pi|\mathbf{f}|^2}{\int d\Omega|\mathbf{f}|^2}.$$

Note that R_{rad} and G also satisfy

$$GR_{rad} = \frac{\eta_o\pi|\mathbf{f}|^2}{\lambda^2}.$$

For a z-polarized short-dipole with $|\mathbf{f}(\hat{r})| = \frac{L}{2}|\sin\theta|$, for instance, these equations imply[5] that (using $\eta_o = 120\pi\,\Omega$)

$$R_{rad} = 20\pi^2(\frac{L}{\lambda})^2\,\Omega \quad \text{and} \quad G = 1.5\sin^2\theta.$$

Also $D = 1.5$ — the maximum value of antenna gain — is the directivity of short-dipole.

For a $\frac{\lambda}{2}$-dipole, on the other hand, it can be shown that

$$R_{rad} \approx 73\,\Omega \quad \text{and} \quad G = 1.64\frac{\cos^2(\frac{\pi}{2}\cos\theta)}{\sin^2\theta}.$$

The corresponding directivity is $D = 1.64$ (since G peaks at $\theta = 90^o$ like for the short-dipole).

While the short- and $\frac{\lambda}{2}$-dipoles have comparable directivities D, their radiation resistances are vastly different. For a short-dipole with $L = 0.1\lambda$, for instance, $R_{rad} \approx 2\,\Omega$, as compared to $73\,\Omega$ for $\frac{\lambda}{2}$-dipole. Even though Ohmic resistance of $\frac{\lambda}{2}$-dipole is 5 times larger than the Ohmic resistance of a $L = 0.1\lambda$ dipole (since resistance increases linearly with antenna length L), its radiation resistance is $\sim \frac{73}{2} = 36.5$ times larger! Thus $\frac{\lambda}{2}$-dipole radiates a larger fraction of its power input, and, therefore, it is a more efficient antenna than the short-dipole. The normalized gain pattern $\frac{G}{D} \propto |\mathbf{f}|^2$ of a z-polarized $\frac{\lambda}{2}$-dipole is plotted in Figure 2.6a.

2.3 Reciprocity and antenna reception

Linear and time-invariant (LTI) circuits composed of resistors, inductors, capacitors satisfy an important system property known as *reciprocity* which can be stated as follows:

[5] $\int d\Omega(\frac{L}{2})^2\sin^2\theta = (\frac{L}{2})^2\int_0^\pi d\theta\sin\theta\int_0^{2\pi}d\phi\sin^2\theta = \frac{8\pi}{3}(\frac{L}{2})^2$. Hence, $G = \frac{4\pi(\frac{L}{2})^2\sin^2\theta}{\int d\Omega(\frac{L}{2})^2\sin^2\theta} = \frac{4\pi(\frac{L}{2})^2\sin^2\theta}{\frac{8\pi}{3}(\frac{L}{2})^2} = \frac{3}{2}\sin^2\theta$, for instance.

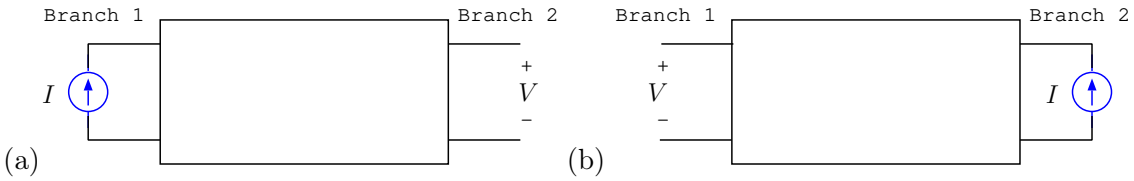


Figure 2.7: A reciprocal circuit excited by a source applied to (a) branch 1, and (b) to branch 2.

If an ideal current source I, applied through branch 1 *of a* **reciprocal circuit,** *causes a voltage response V measured across an open branch labeled as* 2*, applying I through branch* 2 *will induce an open-circuit voltage V across the terminals of branch* 1.

The situation just described is depicted in Figures 2.7a and b where the box in the middle hides all the branches of a reciprocal circuit except for branches 1 and 2. Clearly, the reciprocity property of the circuit illustrated in the figure must also belong to the *two-port network*, i.e., the box, bridging branches 1 and 2.

Examples of *non-reciprocal* circuits and networks include those containing additional sources and/or non-linear elements such as diodes or square-law devices. The proof of reciprocity for circuits and two-ports composed exclusively of resistors, inductors, capacitors is straightforward [see, e.g., Irwin, 1987, *Basic Engineering Circuit Analysis*], and, by extension, *any* network with an equivalent circuit that can be constructed exclusively in terms of the same elements satisfies the reciprocity property. For instance, linear networks containing transformers and/or transmission lines can be constructed in terms of equivalent R, L, and C's and thus satisfy the reciprocity property. Reciprocity property also applies to a pair of antennas coupled electromagnetically through free-space (as well as any any linear and passive propagation medium with the exception of anisotropic materials).

We will next make use of the reciprocity of antennas located in free-space — to be proven shortly — to relate their receiving properties to the already known transmitting properties studied in the previous section. Figure 2.8a depicts a pair of such antennas separated by some distance r while the equivalent circuit for the same system is depicted in Figure 2.8b in terms of a reciprocal two-port where the ports correspond to the antenna terminals. In terms of the equivalent circuit, the antenna terminal voltages can be expressed as

$$\begin{aligned} V_1 &= Z_1 I_1 + Z_c I_2 \\ V_2 &= Z_2 I_2 + Z_c I_1, \end{aligned}$$

where Z_1 and Z_2 stand for the input impedances of the individual antennas and Z_c is a coupling impedance that enables the exchange of signal between the two antennas. Note that if a current source I is applied to either port (as I_1 or I_2), the same open circuit voltage $V = Z_c I$ is obtained at the opposite port, ensuring the reciprocity of the shown network.

Proof of antenna reciprocity in free space: Since in the reciprocal network in Figure

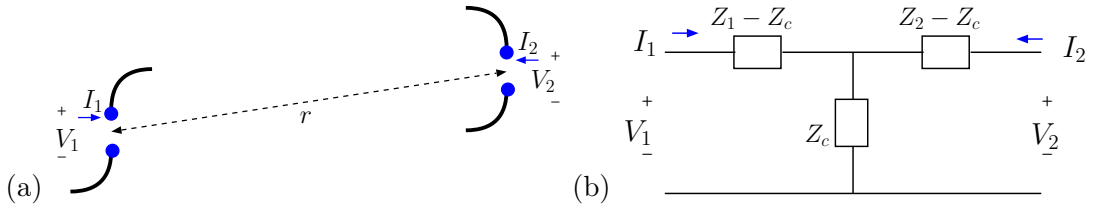


Figure 2.8: (a) A pair of arbitrary antennas separated by some distance r, and (b) a reciprocal two-port model describing the coupling of the antennas shown in (a).

2.8b

$$Z_c = \left.\frac{V_2}{I_1}\right|_{A:\, I_2=0} = \left.\frac{V_1}{I_2}\right|_{B:\, I_1=0},$$

it follows that reciprocity *requires*

$$V_A I_B = V_B I_A$$

in reference to possible states A and B where antennas 1 and 2 in Figure 2.8a are driven one at a time, with state A indicating antenna 2 receiving ($I_2 = 0$), and state B the reverse ($I_1 = 0$). Since $V_A I_B \equiv -\int dv\, \mathbf{E}_A \cdot \mathbf{J}_B$ and $V_B I_A \equiv -\int dv\, \mathbf{E}_B \cdot \mathbf{J}_A$ in terms of dot products of fields $\mathbf{E}$ and currents $\mathbf{J}$ *overlapping* spatially in the feed regions of the two antennas, the reciprocity requirement above becomes

$$\int dv\, \mathbf{E}_A \cdot \mathbf{J}_B = \int dv\, \mathbf{E}_B \cdot \mathbf{J}_A.$$

The fact that this last condition can be derived from Maxwell's equations for any *isotropic* region of space with no further conditions concerning *homogeneity* constitutes a proof of the reciprocity of a pair of antennas communicating through free space. The condition cannot be reproduced if the medium is *anisotropic*, in which case antenna coupling becomes non-reciprocal.

2.3.1 Effective antenna height and open-circuit voltage

Next we write the open circuit voltage of antenna 2 as

$$V_2 = \mathbf{E}_1 \cdot \mathbf{l}_2$$

where $\mathbf{E}_1$ denotes a TEM wave field produced by antenna 1 at the location of antenna 2 (in the absence of antenna 2) and $\mathbf{l}_2$ is a vector which we will call *effective height* of antenna 2. Our problem is to determine $\mathbf{l}_2$ in terms of transmission properties of antenna 2. Thus, we must next consider the case when antenna 2 is transmitting and antenna 1 is receiving; in that case open-circuit voltage of antenna 1 can be written as

$$V_1 = \mathbf{E}_2 \cdot \mathbf{l}_1$$

in terms of TEM field $\mathbf{E}_2$ from antenna 2. Now, *by reciprocity,* it is required that $V_2 = V_1$, i.e.,

$$\mathbf{E}_2 \cdot \mathbf{l}_1 = \mathbf{E}_1 \cdot \mathbf{l}_2$$

if $I_1 = I_2 = I$. However, with $I_1 = I_2 = I$, the fields

$$\mathbf{E}_{1,2} = j\eta_o I k \mathbf{f}_{1,2} \frac{e^{-jkr}}{4\pi r},$$

from which it follows that

$$\mathbf{f}_2 \cdot \mathbf{l}_1 = \mathbf{f}_1 \cdot \mathbf{l}_2$$

must be true for any pair of antennas.

The condition just obtained is readily satisfied by $\mathbf{l} = \mathbf{f}$, but it is not immediately obvious that that is the unique solution of our problem — after all, $\mathbf{l} = 2\mathbf{f}$ also fits the same constraint, for instance.

On the other hand, if it is independently known that $\mathbf{l}_1 = \mathbf{f}_1$, for example, by some means, then the constraint would demand $\mathbf{l}_2 = \mathbf{f}_2$ as well, and, by analogy, demand $\mathbf{l} = \mathbf{f}$, for *any* antenna. Since for the limiting case of a $\hat{z}$-polarized Hertzian dipole exposed to an incident TEM wave $\mathbf{E} = E\hat{\theta}$ we have $V = \mathbf{E} \cdot \mathbf{l} = E \sin\theta dz$, implying $\mathbf{l} = \hat{\theta} \sin\theta dz = \mathbf{f}$, it follows that $\mathbf{l} = \mathbf{f}$ for all possible antennas.

Thus, there is no distinction between antenna height $\mathbf{l}$ and antenna length $\mathbf{f}$, and we are free to use either term/symbol interchangeably in antenna transmission as well as reception problems. Finding the open-circuit voltage response of a receiving antenna to an incident TEM wave $\mathbf{E}$, in particular, is as easy as taking the dot product of $\mathbf{E}$ with $\mathbf{f}$ for the same direction[6].

2.3.2 Available power and effective area

We now turn our attention to average power exchange between the ports of the network shown in Figure 2.8. The open circuit voltage $V_2 = I_1 Z_c$ at port 2 can be interpreted as the Thevenin voltage of antenna 2 when it is used in a receiving role. The corresponding Thevenin impedance looking into port 2 is Z_2 (this can be easily seen by impedance combinations after suppressing I_1). Hence, the available average power of antenna 2, i.e., power that can be delivered to a load $Z_L = Z_2^*$ connected to port 2, can be expressed as

$$P_2 = \frac{|V_2|^2}{8\mathrm{Re}\{Z_2\}} = \frac{|I_1|^2 |Z_c|^2}{8\mathrm{Re}\{Z_2\}}$$

[6]The knowledge that $\mathbf{l} = \mathbf{f}$ can also be used as follows: consider the statement $V_A I_B \equiv -\int dv\, \mathbf{E}_A \cdot \mathbf{J}_B$ from the proof given above, where V_A is referring to antenna 2 voltage when antenna 1 is transmitting and $\mathbf{J}_B/I_B$ to the current distribution of antenna 2 when it is used for transmission. Hence, antenna-2 voltage in reception

$$V_2 = -\int dv \mathbf{E}_A \cdot \frac{\mathbf{J}_2}{I_2} = -\int dv \mathbf{E}_1 \cdot \frac{\mathbf{J}_2}{I_2} - \int dv \mathbf{E}_s \cdot \frac{\mathbf{J}_2}{I_2},$$

in terms of incident field $\mathbf{E}_1$ from antenna 1 and a field $\mathbf{E}_s$ scattered by antenna 2. It can be shown that the first term simplifies to $\mathbf{E}_1 \cdot \mathbf{f}_2$ and so the second term must vanish. This fact leaves us with a useful general formula

$$V = -\int dv \mathbf{E}_i \cdot \frac{\mathbf{J}}{I}$$

for the open circuit voltage of any antenna in terms of its current distribution $\mathbf{J}/I$ and any field $\mathbf{E}_i$ (not TEM, necessarily) incident on the antenna location.

using the usual formula for maximum power transfer. Also, the average power transmitted by antenna 1 to produce V_2 across antenna 2 can be expressed as

$$P_1 = \frac{1}{2}\mathrm{Re}\{V_1 I_1^*\} = \frac{1}{2}|I_1|^2\mathrm{Re}\{Z_1\}$$

since $V_1 = Z_1 I_1$ with port 2 open. Therefore, the available power to transmitted power ratio in the network can be expressed as

$$\frac{P_2}{P_1} = \frac{|Z_c|^2}{4\mathrm{Re}\{Z_1\}\mathrm{Re}\{Z_2\}} \equiv \frac{P_a}{P_t}.$$

Since the result is symmetric with respect to network parameters Z_1, Z_2, and Z_c, it follows that when transmitting and receiving roles of antennas 1 and 2 are exchanged, the power ratio P_a/P_t remains unchanged — this is another important property of reciprocal networks.

We will next exploit the power reciprocity property just derived for a pair of antennas known to be *co-polarized*. By definition, antennas 1 and 2 are said to be co-polarized when their respective effective lengths $\mathbf{f}_1$ and $\mathbf{f}_2$ evaluated for the path joining the antennas are parallel (or anti-parallel) to one another[7]. Now, with antenna 1 transmitting an average power P_t and a co-polarized antenna 2 receiving an average power $P_a \propto P_t$ into an impedance matched load $Z_L = Z_2^*$, we can write

$$P_a = \frac{P_t}{4\pi r^2} G_1 A_2,$$

where parameter A_2 is *defined* to be the *effective area* of antenna 2 for reception — also G_1 in the expression *denotes* the transmission gain of antenna 1 in the direction of antenna 2 and $\frac{P_t}{4\pi r^2}G_1$ stands for the average Poynting flux of the co-polarized TEM wave field from antenna 1 incident on antenna 2. When the antenna roles are reversed — that is, when antenna 2 is transmitting and antenna 1 receiving — we can write, likewise,

$$P_a = \frac{P_t}{4\pi r^2} G_2 A_1,$$

where A_1 is by definition the effective reception area of antenna 1. Since the P_a/P_t ratio cannot change with the role reversal just described, it follows that the products G_2A_1 and G_1A_2 must be equal, or, equivalently, but far more more significantly,

$$\frac{A_1}{G_1} = \frac{A_2}{G_2}.$$

[7] More specifically (a detail which becomes important if $\mathbf{f}$ is complex valued as for circular-polarized antennas), $\mathbf{f}_1$ and $\mathbf{f}_2$ need to satisfy

$$|\mathbf{f}_1 \cdot \mathbf{f}_2| = |\mathbf{f}_1||\mathbf{f}_2|,$$

where $|\mathbf{f}|^2 \equiv \mathbf{f} \cdot \mathbf{f}^*$. For instance $\mathbf{f}_1 = \mathbf{f}_2 = \hat{x}$ are co-polarized (linear, in x direction), as well as $\mathbf{f}_1 = \hat{x} - j\hat{y}$ and $\mathbf{f}_2 = \hat{x} + j\hat{y}$ (circular with the same sense of rotation given that $\mathbf{f}_1$ and $\mathbf{f}_2$ describe propagation in opposite directions). However, $\mathbf{f}_1 = \hat{x}$ and $\mathbf{f}_2 = \hat{x}+\hat{y}$, for instance, are not co-polarized.

The required equality of ratio A_1/G_1 to A_2/G_2 concluded above implies that A/G ratio cannot depend on antenna parameters such as $\mathbf{f}$ and/or $\mathbf{l}$. Instead, A/G must have a universal value applicable to all antennas, i.e., some number proportional to λ^2, the only pertinent physical parameter with the right units which is independent of any antenna specifics.

According to its general definition given above, the *effective area* A of an antenna corresponds to the ratio of its available power

$$P_a = \frac{|V|^2}{8R_{rad}} = \frac{|\mathbf{E}\cdot\mathbf{l}|^2}{8R_{rad}}$$

and incident Poynting flux

$$S = \frac{|\mathbf{E}|^2}{2\eta_o}$$

due to a co-polarized wave field $\mathbf{E} \propto \mathbf{f}^*$, where $R_{rad} \equiv \mathrm{Re}\{Z_a\}$ is the antenna radiation resistance and $\mathbf{f}$ and $\mathbf{l}$ the pertinent effective length and height, respectively. Consequently, we have

$$A \equiv \frac{P_a}{S} = \frac{\eta_o}{4R_{rad}}\frac{|\mathbf{E}\cdot\mathbf{l}|^2}{|\mathbf{E}|^2},$$

which is applicable for any antenna. Combining this result for the effective area A with

$$GR_{rad} = \frac{\eta_o\pi|\mathbf{f}|^2}{\lambda^2}$$

from Section 2, we obtain

$$\frac{A}{G} = \frac{\lambda^2}{4\pi}\frac{|\mathbf{E}\cdot\mathbf{l}|^2}{|\mathbf{E}|^2|\mathbf{f}|^2} = \frac{\lambda^2}{4\pi},$$

given that $\mathbf{l} = \mathbf{f}$ in general and $\mathbf{E} \propto \mathbf{f}^*$ by our original assumption. Note that A/G is independent of any antenna properties as required by power reciprocity discussed earlier. Furthermore, we could have deduced the fact that $\mathbf{l} = \mathbf{f}$ (had we not known it already) from above ratio using the power reciprocity constraint applied to the case of a Hertzian dipole.

Summarizing the key features of the results above:

1. The available power received from a *co-polarized* TEM wave field $\mathbf{E}$ carrying an average Poynting flux S is

   $$P_a = SA,$$

   where

   $$A = \frac{\lambda^2}{4\pi}G$$

   is the *effective area* of the antenna for the direction of arrival of TEM wave $\mathbf{E}$ and G denotes the transmission gain of the antenna in the same direction — the effective area A of any antenna should satisfy a universal constraint $\int A d\Omega = \lambda^2$, which follows from $\int G d\Omega = 4\pi$ and the A to G relation above.

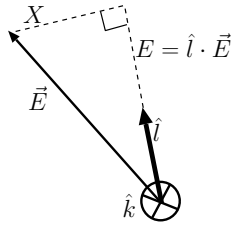


Figure 2.9: Only the projection $E \equiv \hat{l} \cdot \mathbf{E}$ of the incident TEM wave field $\mathbf{E}$ at the antenna location matters for the antenna voltage output $V \propto E$. The unit vector $\hat{l}$, such that $\hat{l} \cdot \hat{l}^* = 1$, describes the polarization of outgoing radiation field of the antenna in direction $\hat{k}$, same as the direction of arrival of TEM wave field $\mathbf{E}$.

2. The open-circuit voltage of any antenna responding to *any* (co-polarized or not) TEM wave field $\mathbf{E}$ can be expressed as

$$V = \mathbf{E} \cdot \mathbf{l},$$

where $\mathbf{l} = \mathbf{f}$ is the effective length of the same antenna for the direction of arrival of field $\mathbf{E}$. We will no longer distinguish between $\mathbf{l}(\hat{r})$ and $\mathbf{f}(\hat{r})$ in future usage and call them both the *effective length* (or height).

3. For a *cross-polarized* field $\mathbf{E}$, such that $\mathbf{E} \cdot \mathbf{l} \equiv 0$, the corresponding open-circuit voltage $V = 0$, and consequently $P_r = 0$ and co-polarized flux $S = 0$. Antennas can only respond to co-polarized fields in reception — that is one reason why re-orienting a radio antenna can help improve radio reception (what is another reason?).

4. A TEM wave field $\mathbf{E} = E_o\hat{e}$ of an arbitrary polarization $\hat{e}$ arriving from some direction $\hat{k}$ can be expressed (with no loss of generality) as $\mathbf{E} = E\hat{l}^* + X\hat{q}^*$, where $\hat{l} \equiv \mathbf{l}(\hat{k})/|\mathbf{l}(\hat{\mathbf{k}})|$, $\hat{l} \cdot \hat{q}^* = 0$, with

$$E \equiv \hat{l} \cdot \mathbf{E}$$

denoting the co-polarized component of $\mathbf{E}$, and X the cross-polarized component (see Figure 2.9). Since the corresponding open-circuit voltage is $V = \mathbf{E} \cdot \mathbf{l}(\hat{k}) = E|\mathbf{l}(\hat{k})|$, co-polarized Poynting flux S needed to calculate P_a in terms of effective area A is (see the examples below for how to use this formula)

$$S = \frac{|E|^2}{2\eta_o}.$$

Example 1: Figure 2.10 shows a z-polarized short-dipole antenna located at the origin exposed to an incident TEM wave field

$$\mathbf{E}(\mathbf{r}) = (\hat{x} - \hat{z})e^{j\pi(x+z)/\sqrt{2}}\text{V/m}.$$

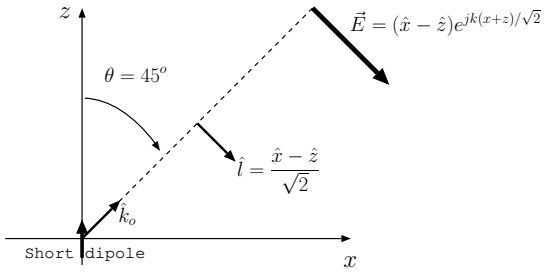


Figure 2.10: A wave field $\mathbf{E}$ incident on a z-polarized *short-dipole* at the origin. The fields arrives from $\theta = 45^o$ and $\phi = 0^o$ direction, and, therefore, the available power at the dipole will be proportional to $A(45^o) = \frac{\lambda^2}{4\pi}G(45^o)$. Unit vector $\hat{l}$ describes the polarization of the field the dipole would radiate in $\theta = 45^o$ and $\phi = 0^o$ direction.

Clearly $k = \pi$ rad/m, $\lambda = \frac{2\pi}{k} = 2$ m, $\mathbf{E} = (\hat{x} - \hat{z})$ V/m at the origin, and co-polarized

$$E = \hat{l} \cdot \mathbf{E} = \sqrt{2}\ \frac{\text{V}}{\text{m}}$$

since $\hat{l} = \frac{\hat{x}-\hat{z}}{\sqrt{2}}$. Therefore, using $G = 1.5\sin^2\theta$ for a short-dipole with $\theta = 45^o$ (as shown in Figure 2.10), we find that

$$P_r = \frac{|E|^2}{2\eta_o}A = \frac{|E|^2}{2\eta_o}\frac{\lambda^2}{4\pi}G(45^o) = \frac{\sqrt{2}^2}{240\pi}\frac{(2)^2}{4\pi}1.5\sin^2 45^o = \frac{3/2}{240\pi^2}\text{W}.$$

This is the power the dipole will deliver to a matched termination with an impedance $Z_L = Z_a^*$ where Z_a stands for the antenna input impedance.

Example 2: The available power of a z-polarized short-dipole antenna exposed to an incident field

$$\mathbf{E} = \hat{y}e^{j\pi(x+z)/\sqrt{2}}\text{V/m}$$

will be zero because the given $\mathbf{E}$ is entirely cross-polarized. Note that since $\hat{k} = \frac{\hat{x}+\hat{z}}{\sqrt{2}}$, polarization $\hat{l} = \frac{\hat{x}-\hat{z}}{\sqrt{2}}$ in direction $\hat{k}$, and, therefore,

$$E = \hat{l} \cdot \mathbf{E} = \frac{\hat{x} - \hat{z}}{\sqrt{2}} \cdot \hat{y} = 0.$$

Example 3: An x-polarized half-wave dipole at the origin is exposed to a circular polarized incident field

$$\mathbf{E} = j4(\hat{x} - j\hat{y})e^{j2\pi z}\ \text{V/m}.$$

What is the co-polarized $E = \hat{l} \cdot \mathbf{E}$ and what is the available power P_a?

Solution: Since $\hat{k} = \hat{z}$, for the given x-polarized dipole the transmitted wave polarization is $\hat{l} = -\hat{x}$ in direction $\hat{k}$. Consequently

$$E = \hat{l} \cdot \mathbf{E} = -\hat{x} \cdot j4(\hat{x} - j\hat{y}) = -j4\ \text{V/m}.$$

Given that $G = 1.64$ for the same dipole in $\hat{k} = \hat{z}$ direction, and since the relevant wavelength is $\lambda = \frac{2\pi}{k} = 1$ m, it follows that the pertinent effective area is $A = \frac{\lambda^2}{4\pi}G = \frac{1.64}{4\pi}$. Hence,

$$P_a = \frac{|-j4|^2}{240\pi}\frac{1.64}{4\pi} = \frac{1.64}{60\pi^2} \text{ W}.$$

Note that half the power density of $\mathbf{E}$ did not contribute to P_a since it was associated with the cross-polarized component of $\mathbf{E}$ in $\hat{y}$ direction. An antenna with $\hat{l} = \frac{\hat{x}+j\hat{y}}{\sqrt{2}}$ would have extracted twice the amount of P_a above from the same incident field.

When a receiving antenna is terminated by a load Z_L other than a properly matched Z_a^*, the average power P_r delivered to Z_L will be less than the available power P_a, namely $P_r = MP_a \leq 1$, where — as it can be easily derived — the so-called *impedance mismatch factor* M is given by

$$M = \frac{4\text{Re}\{Z_a\}\{Z_L\}}{|Z_a + Z_L|^2} \leq 1.$$

Also, it is useful to define a *polarization mismatch factor*

$$p \equiv \frac{|\mathbf{E}\cdot\mathbf{l}|^2}{|\mathbf{E}|^2|\mathbf{l}|^2} \leq 1$$

such that available power P_a can be calculated as

$$P_a = Ap\frac{|\mathbf{E}|^2}{2\eta_o}$$

in terms of the full Poynting flux $|\mathbf{E}|^2/2\eta_o$ of the incident field (co-polarized or not) $\mathbf{E}$.

Our final example here concerns the reception of a combination of circular polarized fields by a circular polarized antenna (typically constructed using crossed dipoles or helical elements).

Example 4: An antenna at the origin which transmits with polarization $\hat{l} = \frac{\hat{x}-j\hat{y}}{\sqrt{2}}$ towards direction $\hat{z}$ is being exposed to a field

$$\mathbf{E} = (R\frac{\hat{x}+j\hat{y}}{\sqrt{2}} + L\frac{\hat{x}-j\hat{y}}{\sqrt{2}})e^{j2\pi z}.$$

What is the co-polarized component E of field $\mathbf{E}$ incident on the antenna?

Solution: Clearly, the plane wave field $\mathbf{E}$ is coming from the z-direction, and, therefore,

$$E = \hat{l}\cdot\mathbf{E} = \frac{\hat{x}-j\hat{y}}{\sqrt{2}}\cdot(R\frac{\hat{x}+j\hat{z}}{\sqrt{2}} + L\frac{\hat{x}-j\hat{z}}{\sqrt{2}}) = R.$$

Clearly, the antenna is responding only to the right-circular component of $\mathbf{E}$ proportional to R — note that unit vector $\frac{\hat{x}+j\hat{y}}{\sqrt{2}}$ describes a right-circular wave propagating in $-z$ direction, which is co-polarized with with the right-circular wave $\propto \frac{\hat{x}-j\hat{z}}{\sqrt{2}}$ propagating in $+z$ direction.

As Example 4 has just shown, right-circular polarized antennas (which radiate away right-circular waves) are totally blind to incoming waves with left-circular polarization. Thus, a right-circular antenna cannot see or detect left-circular echoes (reflected or scattered) of its own transmissions, rotating in the same direction as the radiated wave but propagating (obviously) in reverse direction. In that regard, notice that an $\hat{e}$ polarized wave propagating in $\hat{k}$ direction and an $\hat{e}^*$ polarized wave propagating in $-\hat{k}$ direction have the *same* polarization — e.g., *right-circular* and co-polarized $\hat{e} = \frac{1}{\sqrt{2}}(\hat{x} - j\hat{y})$ and $\hat{e}^* = \frac{1}{\sqrt{2}}(\hat{x} + j\hat{y})$ propagating in $\hat{z}$ and $-\hat{z}$ directions, respectively.

2.4 Antenna noise

It is well known that resistors produce fluctuating voltages $v(t)$ at their terminals known as *noise*. The source of resistor noise is random thermal motions of electrical charge carriers (mainly electrons) within resistive materials. The rms value of noise voltage $v(t)$ and the associated available power P_n are related by (as for any AC voltage)

$$P_n = \frac{v_{rms}^2}{4R},$$

where R is the resistance value and

$$P_n = \int_0^\infty \frac{hf}{e^{hf/KT} - 1} df$$

for any resistor at temperature T [e.g., *Nyquist,* 1928] — here, K is Boltzmann's constant, 1.381×10^{-23} W/Hz/K, h is Planck's constant, 6.626×10^{-34} J.s, and $f = \omega/2\pi$ is frequency in Hz.

In view of the above we can write

$$dP_n = \frac{hf}{e^{hf/KT} - 1} df \quad \overset{hf \ll KT}{\longrightarrow} \quad KT df$$

for power available over an infinitesimal bandwidth df. Hence, over a finite bandwidth B in the frequency range $f \ll KT/h$, where dP_n is independent of f, the available noise power of a resistor is

$$P_n = KTB.$$

It is convenient to define a noise phasor V_n, with a mean-square denoted as $\langle |V_n|^2 \rangle$, such that

$$P_n = KTB \equiv \frac{\langle |V_n|^2 \rangle}{8R} B.$$

The noise phasor V_n represents a co-sinusoid that carries the same average power as noise voltage $v(t)$ over a unit bandwidth[8], and can be treated as a regular voltage phasor

[8] Any bandwidth unit can be employed here, e.g., Hz, kHz, or even μHz, as long as the same unit is used to express B in the formula for P_n.

in circuit calculations in superposition with phasors defined for the same frequency f. Noise phasors V_n and V_m due to different resistors in a circuit are assumed to have (by definition) independent random phases so that $\langle|V_n + V_m|^2\rangle = \langle|V_n|^2\rangle + \langle|V_m|^2\rangle$ is true (see Example 1 below).

So far we talked about resistor noise, avoiding the subject in the title of this section, namely, antenna noise. It turns out, however, that all the formulae above for P_n, dP_n, as well as noise phasor V_n, also apply to the noise output of antennas immersed in blackbody radiation.

To see how and why, assume that an antenna, terminated by a matched load $R-jX$, is in thermal equilibrium with a background described by Planck's blackbody distribution

$$\mathcal{E}(f,T)df = \frac{8\pi}{c^3}\frac{hf^3}{e^{hf/KT}-1}df \quad \overset{hf\ll KT}{\longrightarrow} \quad \frac{8\pi}{c^3}KTf^2df$$

for electromagnetic energy density $\mathcal{E}$. The antenna load resistance R at the blackbody temperature T will produce and deliver an average noise power KTB to its matched antenna over a bandwidth B, only be radiated out into the noisy background (assuming a lossless antenna for simplicity). This, of course, would lead to a cooling of the load *unless* the antenna picks up and delivers an equal amount of noise power, KTB, back to the load. Since a net heat exchange is not possible between systems at equal temperatures — the load and the background, in this case — it follows that the antenna must indeed pick up an available noise power of KTB from incident blackbody radiation over a bandwidth B [e.g., *Burgess*, 1941].

The available noise power of antennas within a bandwidth B is always expressed as

$$P_n = \frac{\langle|V_n|^2\rangle}{8R}B = KTB$$

whether or not the antenna is actually immersed in blackbody radiation. Hence, T in the formula above represents an *equivalent antenna temperature* which accounts for the radiation picked up from the environment, a sum of a large number independent random co-sinusoids with a mean square amplitude $\langle|V_n|^2\rangle$ in each unit bandwidth. We envision an open circuit voltage phasor

$$V_1 + V_2 + V_3 + \cdots$$

where each component is a random-phased voltage phasor due to an independent radiation source, sources like radio stars, galaxies, our galactic core, etc., (all of which constitute signals for a radio astronomer), as well as atmospheric lightning and thermal emissions from atmospheric gases, and so on. The mean-squared value of the large sum is obtained as

$$\langle|V_1 + V_2 + V_3 + \cdots|^2\rangle = \langle|V_1|^2\rangle + \langle|V_2|^2\rangle + \cdots \equiv \langle|V_n|^2\rangle,$$

leading to the P_n formula for any bandwidth B.

The practical antenna noise temperature T to be used in P_n formula turns out to be independent of antenna properties when noise sources are uniformly distributed across the

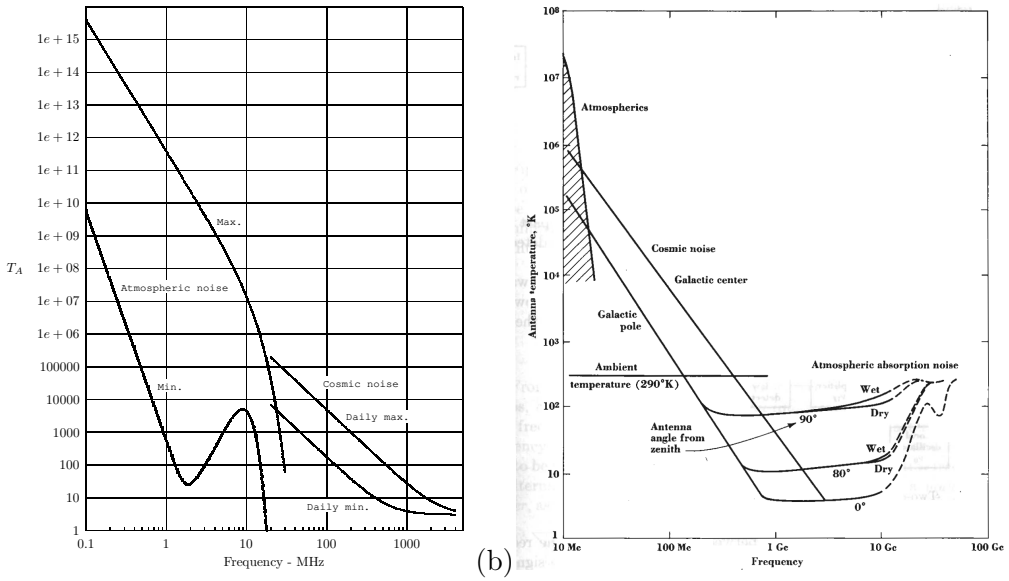


Figure 2.11: Antenna noise temperature as a function of radio frequency, with emphasis on (a) HF and VHF (from *Franke*, 2006) and (b) microwaves (from *Kraus,* Radio astronomy, 1966)

sky within the antenna's field-of-view[9]. Measurements show that antenna temperature varies with frequency as $T \propto f^{-2}$ across most of the RF band, and drops from ~ 3000 K at VHF to ~ 10 K or less at microwave frequencies (see Figure 2.11). Perhaps the most surprising aspect of this result is why the antenna temperature is independent of antenna size or directivity. Why do smaller antennas with smaller effective areas collect the same amount of noise power from the sky as large antennas and arrays?

The reason can be explained as follows: Let SA denote the available noise power from a single external source providing a flux S upon an antenna with an effective area A. If similar sources were distributed uniformly across the celestial sphere, then the number of similar contributions to P_n would be proportional to the beam solid angle Ω_o. Therefore, given that $\Omega_o = \frac{4\pi}{D}$ and $A = \frac{\lambda^2}{4\pi}D$,

$$P_n \propto SA\Omega_o = S(\frac{\lambda^2}{4\pi}D)(\frac{4\pi}{D}) = \lambda^2 S$$

for any antenna. Clearly, this result shows that low-directivity antennas with wider beams "see" a larger number of sources with less sensitivity and/but the overall result is the same for all antennas independent of their directivities. This result also suggests $T \propto \lambda^2 \propto f^{-2}$, which is consistent with the measurements mentioned above. Of course any directional dependence of the parameter S used in above explanation will cause a dependence of T on antenna beam width and pointing direction.

[9]Which is well approximated usually, except when a strong source, e.g., Cassiopeia A, a compact supernova remnant, the Sun, or some man-made interference source, is within the field-of-view...

Example: An antenna impedance is given as $R_a = R_1 + R_2$, where R_1 is the radiation resistance and R_2 represents the Ohmic losses in the antenna. Two components of Z_a provide independent noise outputs, with mean-square noise phasors $\langle|V_1|^2\rangle = 8R_1KT_s$ and $\langle|V_2|^2\rangle = 8R_2KT_0$, respectively, where T_s is the equivalent antenna sky temperature due to external radiation, and T_0 the physical temperature of the antenna. Express the available noise power of the antenna over a bandwidth B in terms of T_s, T_0, and *antenna radiation efficiency* defined as $\eta \equiv R_1/R_a$.

Solution: The open-circuit voltage phasor of the antenna is clearly $V_1 + V_2$ in a unit bandwidth. Hence the available noise power over bandwidth B is

$$\begin{aligned} P_n &= \frac{\langle|V_1+V_2|^2\rangle}{8R_a}B = \frac{\langle|V_1|^2\rangle + \langle|V_2|^2\rangle}{8R_a}B \\ &= \frac{8R_1KT_s + 8R_2KT_0}{8R_a}B = KT_sB\frac{R_1}{R_a} + KT_0B\frac{R_2}{R_a} \\ &= KT_sB\eta + KT_0B(1-\eta). \end{aligned}$$

2.5 Space- and ground-wave links

Figure 2.12 shows a pair of antennas at heights h_t and h_r above a ground plane ($z = 0$) forming a *space-wave* link spanning a *ground distance* x. Assume that the antenna at $z = h_t$ transmits an average power P_t while the antenna at $z = z_r$ delivers an average power P_r to a matched load. We want to find the relationship between P_r and P_t describing the *power budget* of the link. The significance of power budget equations lies in the fact that the received power P_r of a link is in practice always in competition with noise power P_n at the receiving antenna output, and it is in general desirable to have a P_r which is at least not negligible compared to P_n.

We will first determine P_r ignoring the effect of the ground plane and then modify the expression in an appropriate manner. Denoting the slant distance between the two antennas as r as shown in Figure 2.12 we can write

$$P_r = \underbrace{\frac{P_t}{4\pi r^2}G_t}_{\text{Poynting}} \underbrace{\frac{\lambda^2}{4\pi}G_r}_{A} = P_t\frac{\lambda^2 G_tG_r}{(4\pi r)^2},$$

where G_t and G_r denote the transmission and reception antenna gains in the link direction (along the direct path of length r). We will next modify this result — which is known as Friis' transmission formula — by multiplying it with a factor $|F|^2$ that takes into account the "multi-path" between the two antennas.

Let us denote the open circuit voltage of the receiving antenna as

$$V_o = V_1 + V_2,$$

where V_1 is the voltage due to the field arriving directly from the transmitting antenna and V_2 due to ground-reflected contributions which we ignored above. We can then

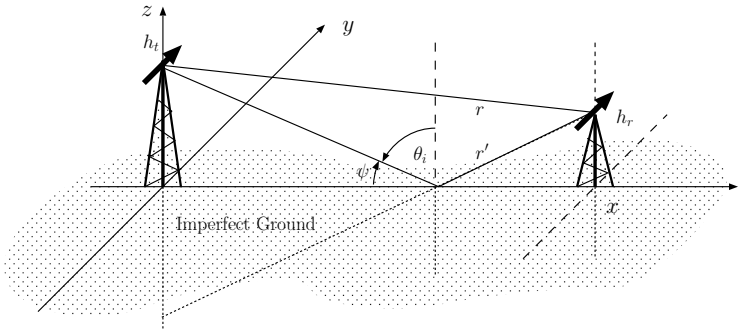


Figure 2.12: A space wave link above an imperfect ground plane. Note that for $x \gg h_t$, h_r, the incidence angle θ_i is nearly 90^o.

express the available power of the receiving antenna as

$$P_r = \frac{|V_o|^2}{8R_{rad}} = \frac{|V_1 + V_2|^2}{8R_{rad}} = \frac{|V_1|^2}{8R_{rad}}|1 + \frac{V_2}{V_1}|^2 = P_t \frac{\lambda^2 G_t G_r}{(4\pi r)^2}|F|^2,$$

where

$$F \equiv 1 + \frac{V_2}{V_1}$$

is known as *space factor*. It is clear from Figure 2.12 that if $x \gg h_t$, h_r, then

$$r = \sqrt{x^2 + (h_t - h_r)^2} \approx x + \frac{(h_t - h_r)^2}{2x}$$

while "the reflected path length" r' marked in the same diagram is

$$r' = \sqrt{x^2 + (h_t + h_r)^2} \approx x + \frac{(h_t + h_r)^2}{2x}.$$

Furthermore, for $x \gg h_t$, h_r the angles of departure and arrival of the two paths (with respect to transmitting and receiving antennas) are nearly the same and we are justified to ignore the gain and effective area differences between the two paths and model V_2 as

$$V_2 = V_1 R e^{-jk(r'-r)}$$

where R denotes the reflection coefficient of the ground (see Chapter 1) at angle θ_i. The exponential term of course accounts for the phase difference of direct and indirect path signals due to path difference

$$r' - r \approx (x + \frac{(h_t + h_r)^2}{2x}) - (x + \frac{(h_t - h_r)^2}{2x}) = \frac{(h_t + h_r)^2 - (h_t + h_r)^2}{2x} = \frac{2h_t h_r}{x}.$$

Hence we obtain

$$F = 1 + \frac{V_2}{V_1} = 1 + R e^{-j\frac{4\pi h_t h_r}{\lambda x}}.$$

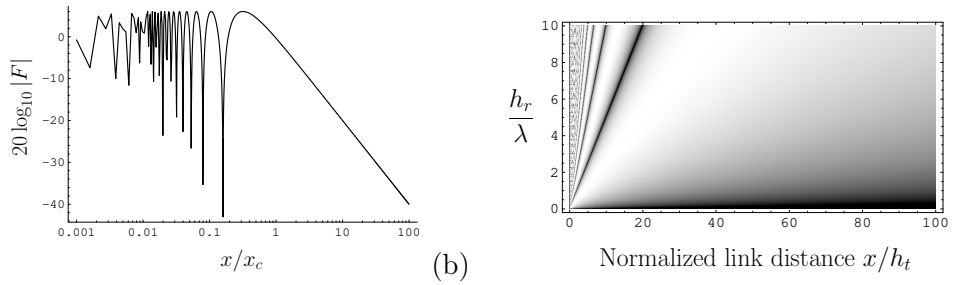


Figure 2.13: (a) Space factor $|F|^2$ in dB versus $\frac{x}{x_c}$ and (b) density plot of $|F|^2$ versus $\frac{x}{h_t}$ and $\frac{h_r}{\lambda}$. Black indicates a zero space factor. The diagonal white region of maximum space factor 4 (or 6 dB) separates the "interference zone" to its left ($x < x_c$) from the far-field ($x > x_c$) to its right.

Let us examine this result assuming that reflection coefficient $R = -1$ (which is true for horizontal polarized fields above a perfect ground). In that case

$$F = 1 - e^{-j\frac{4\pi h_t h_r}{\lambda x}} \approx j\frac{4\pi h_t h_r}{\lambda x} \quad \rightarrow \quad |F| \approx \frac{4\pi h_t h_r}{\lambda x} \text{ for } x \gg x_c \equiv \frac{4\pi h_t h_r}{\lambda},$$

which, it turns out, is pertinent for all space wave links because (as shown in Chapter 1) the reflection coefficient $R \approx -1$ for both horizontal and vertical polarized fields above an imperfect ground as $\theta_i \rightarrow 90^o$. Note that $|F|$ oscillates between extreme values of 0 and 2 (meaning destructive and constructive interference, respectively) for relatively small ground separations $x < x_c$ and then beyond x_c it decreases smoothly as $\propto x^{-1}$ with increasing x (see Figure 2.13a). For $\lambda = 3$ m (corresponding to FM frequencies of ~ 100 MHz), for instance, $x_c \approx 4$ km with $h_t = h_r = 30$ m, and x_c is less than 2 km with $h_t = 300$ m and h_r of a few meters (a typical case of an FM transmission tower serving a portable receiver). The point is, at least in the FM band, x_c is usually a very small distance and in most circumstances the far-field approximation $|F| \propto h_t h_r x^{-1}$ applies (see Figure 2.13b). Given that h_r values of more than a few meters cannot be realistically expected in FM, broadcast stations are forced to maintain their transmitting antennas at substantial heights h_t in order to reach a broader area (usually limited by the curvature of the Earth).

The space wave link model developed above fails in h_t and/or $h_r \rightarrow 0$ limit. In that limit $F \rightarrow 0$ while in fact *ground wave links* using surface mounted antennas function with non-zero transmission gains $\frac{P_r}{P_t}$ at least within the AM band. This discrepancy can be explained as follows:

Applying a single reflection coefficient to a single reflected ray is an approximation. The exact procedure to determine F is to first find out the *plane wave spectrum* of the transmitted wave field and then apply angle-varying reflection coefficients to each plane wave component of the spectrum (covering all angles of incidence rather than just $\theta_i \approx 90^o$) before superposing the reflected waves. When the procedure — the details are given in the Appendix — is carried out for vertical polarized monopole transmitters it is

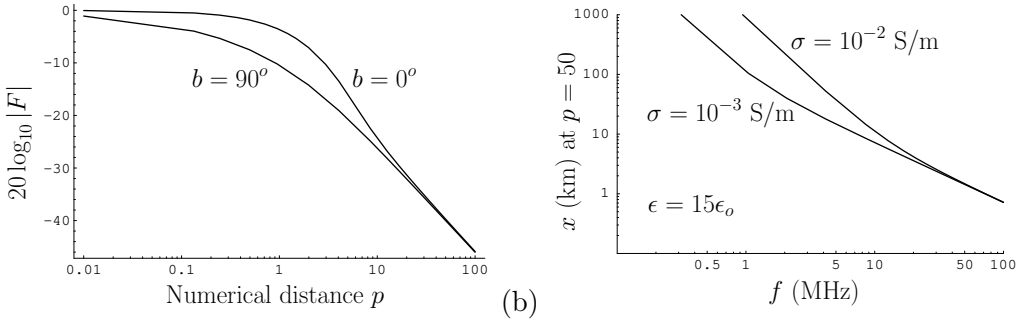


Figure 2.14: (a) Surface wave attenuation factor $|A_s|$ in dB versus numerical distance $p = \frac{\pi r}{\lambda|\epsilon_{rg} - j\frac{\sigma_g}{\omega\epsilon_o}|}$ for $b = \tan^{-1}\frac{\omega\epsilon_o}{\sigma_g} = 0^o$ and 90^o, and (b) variation of distance x at $p = 50$ as a function of frequency f for typical ground parameters.

found that[10] in $h_t \to 0$ limit the total vertical field just above a ground with a refractive index n_g is

$$E_z = -j\eta_o I_o kh \frac{e^{-jkx}}{4\pi x} A_s \quad \text{for a short vertical monopole of length } h$$

where

$$A_s = 1 - j\sqrt{\pi p}e^{-jb/2}e^{-pe^{-jb}}\text{erfc}(j\sqrt{p}e^{-jb/2}),$$

with

$$\text{erfc}(\xi) \equiv \frac{2}{\sqrt{\pi}}\int_{\xi}^{\infty} du\ e^{-u^2}, \quad p \equiv \frac{kx}{2|n_g|^2}, \quad \text{and} \quad b \equiv \angle jn_g^2.$$

In these formulae $n_g^2 = \epsilon_{rg} - j\frac{\sigma_g}{\epsilon_o\omega}$, $p \propto x$ is known as *numerical distance*, and $|A_s|$ is known as "surface wave" *attenuation factor*.

For $A_s = 1$ — which is true for $x = 0$ — the field expression above is just the field of a short vertical monopole above a perfect conductor (see homework problem). However, with finite σ, the attenuation factor $|A_s| < 1$ for non-zero x and varies with $p \propto x$ as shown in Figure 2.14a. Note that for $p = 50$ attenuation is about -40 dB. Distance x corresponding to numerical distance $p = 50$ is given by

$$x = \frac{50\lambda}{\pi}|\frac{\epsilon_g}{\epsilon_o} - j\frac{\sigma_g}{\omega\epsilon_o}|,$$

which is plotted versus frequency $f = \frac{\omega}{2\pi}$ in Figure 2.14b assuming typical ground parameters. Clearly, Figure 2.14b indicates that imperfect conductivity of the ground does not preclude an effective use of monopoles in the AM band (or at frequencies lower than 1 MHz) but causes a severe departure from the ideal case at higher frequencies.

Other complicating factors that impact the operation of space wave links include rough surface scattering (when the ground plane is irregular), so called "beam divergence" due

[10]See Chapter 16 in "Electromagnetic waves and radiating systems" by *Jordan and Balmain* or Section 6.5 in "Antennas and Radiowave Propagation" by *Collin*.

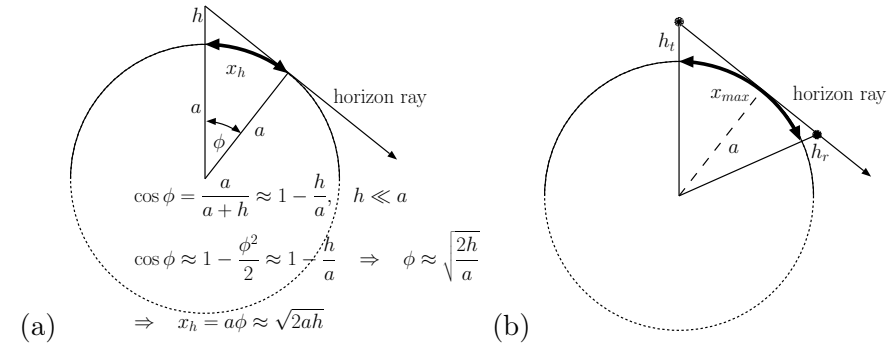


Figure 2.15: (a) Derivation of horizon distance formula $x_h \approx \sqrt{2ah}$, and (b) the geometry of a space wave link with $x = x_{max} = \sqrt{2ah_t} + \sqrt{2ah_r}$. Note that as $x \to x_{max}$ space factor F for the link requires corrections to account for curvature effects.

to the spherical shape of the ground surface, mid-path obstacles, tropospheric refraction, etc[11]. The reach of a space wave link is of course ultimately limited by the curvature of the Earth.

As shown in Figure 2.15a the *horizon distance* x_h for an antenna elevated to some height h above the spherical Earth is approximately $x_h \approx \sqrt{2ah}$ where a is the Earth radius (6371.2 km). Also as illustrated in Figure 2.15b a pair of antennas at heights h_t and h_r above the surface can form a link only if their ground separation x is less than $x_{max} \approx \sqrt{2ah_t} + \sqrt{2ah_r}$, or less than

$$x_{max} \approx \sqrt{2a'h_t} + \sqrt{2a'h_r} \qquad \text{with} \quad a' = \frac{4}{3}a,$$

taking tropospheric refractions into account. For instance, $x_{max} \approx 83$ km if we assume $h_t = h_r = 100$ m.

2.6 Antenna arrays and imaging

We have mainly discussed so far small-sized antennas with relatively broad beam patterns and low directivities. In this section we will examine how such antennas can be combined to design high directivity systems with narrow beam patterns. Such systems are in general referred to as antenna arrays. Antenna arrays can be used to image the angular structure of far-away radio sources as discussed at the end of the section.

2.6.1 1-D antenna arrays

Consider an *array* of identical antennas with identical effective lengths $\mathbf{l}_e(\hat{r})$ and positions $\mathbf{r}_n = (nD, 0, 0) = \hat{x}Dn$ along the x-axis, as shown in Figure 2.16a. Also assume that the

[11] Details of space wave link design including all conceivable complications are covered in "Propagation of short radio waves" by *Kerr*.

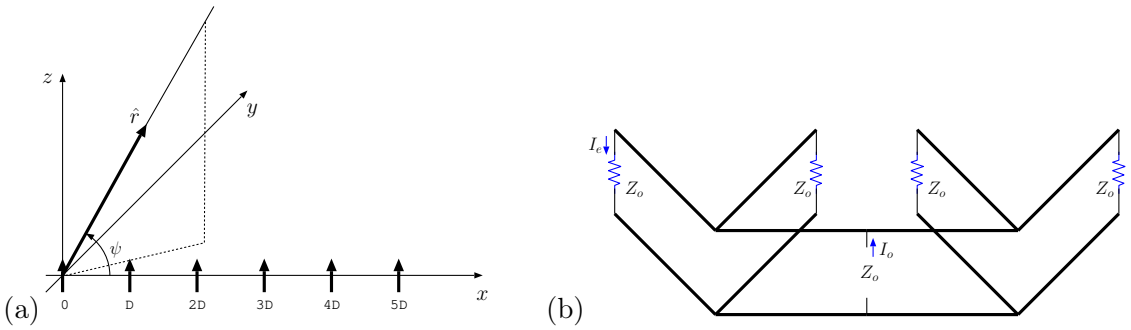


Figure 2.16: (a) An N-element 1-D antenna array with an element-to-element spacing D along the x-axis. The beam pattern of the array is most conveniently expressed in terms of the cosine of angle ψ measured from the array axis towards the observing direction $\hat{r} = \hat{x}\sin\theta\cos\phi + \hat{y}\sin\theta\sin\phi + \hat{z}\cos\theta$ — the *direction cosine* is $\cos\psi = \hat{x}\cdot\hat{r} = \sin\theta\cos\phi$ in terms of zenith and azimuth angles θ and ϕ. For 1-D arrays along y- or z-axes, use alternate direction cosines given by $\hat{y}\cdot\hat{r} = \sin\theta\sin\phi$ or $\hat{z}\cdot\hat{r} = \cos\theta$, respectively. (b) A *corporate ladder* network for a 4-element antenna array consisting of quarter-wave transformer segments with characteristic impedances matched to antenna impedances Z_o. It is easy to show that $|I_e| = |I_o|/2$ using power conservation.

antenna array is exposed to a plane TEM-wave

$$\mathbf{E}(\mathbf{r}) = \mathbf{E}_o e^{j\mathbf{k}\cdot\mathbf{r}}$$

arriving from an arbitrary direction $\hat{r} = \hat{k} \equiv \mathbf{k}/k$. Each *element* of the array will then respond to $\mathbf{E}(\mathbf{r})$ with an open-circuit voltage

$$V_n = \mathbf{l}_e(\hat{r})\cdot\mathbf{E}(\mathbf{r}_n) = \mathbf{l}_e(\hat{r})\cdot\mathbf{E}_o e^{j\mathbf{k}\cdot\mathbf{r}_n} = \mathbf{E}_o\cdot\mathbf{l}_e(\hat{r})e^{jk\hat{r}\cdot\hat{x}nD},$$

where the scalar $\mathbf{E}_o\cdot\mathbf{l}_e(\hat{r})$ corresponds to voltage V_0 of element $n=0$ located at $\mathbf{r}=0$; the other V_n's are just phase shifted versions of V_0 depending on direction of arrival $\hat{r}=\hat{k}$ and element locations $\hat{x}Dn$.

Let us next envision a linear, time-invariant network connected to the elements above — see Figure 2.16b for an example network — to produce a weighted superposition

$$V = \sum_{n=0}^{N-1} W_n V_n = \mathbf{E}_o\cdot\mathbf{l}_e(\hat{r})\sum_{n=0}^{N-1} W_n e^{jk\hat{r}\cdot\hat{x}nD}$$

at some terminal, and, reciprocally, generate element currents $I_n = W_n I_o$ when a current I_o is applied to the same terminal. The elements and their connection network just described constitute an *array antenna*, for which the output voltage $\sum_{n=0}^{N-1} W_n V_n$ (during reception) can be re-expressed as

$$V = \mathbf{E}_o\cdot\mathbf{f}(\hat{r}),$$

in terms of an *effective length*

$$\mathbf{f}(\hat{r}) \equiv \mathbf{l}_e(\hat{r}) \sum_{n=0}^{N-1} W_n e^{jnkD\cos\psi}.$$

Above, $\cos\psi \equiv \hat{x}\cdot\hat{r}$ is a *direction cosine* as pictured in Figure 2.16a. Also, the transmitted field of the array, when driven with an input current I_o, can be expressed as

$$\mathbf{E}(\mathbf{r}) = j\eta_o I_o k\mathbf{f}(\hat{r})\frac{e^{-jkr}}{4\pi r} = \sum_{n=0}^{N-1} j\eta_o I_o W_n k\mathbf{l}_e(\hat{r})e^{jnkD\cos\psi}\frac{e^{-jkr}}{4\pi r},$$

which is clearly a linear superposition of the radiation fields[12] of individual array elements with driving point currents $I_n = I_o W_n$.

The effective lengths $\mathbf{f}(\hat{r})$ of array antennas with element lengths $\mathbf{l}_e(\hat{r})$ are commonly written as

$$\mathbf{f}(\hat{r}) = \mathbf{l}_e(\hat{r})F(\psi),$$

where the scalar function

$$F(\psi) \equiv \sum_{n=0}^{N-1} W_n e^{jnkD\cos\psi}$$

is said to be the *array factor*. Note that the array factor is a spatial Fourier transform (discrete) of the current distribution $W_n = I_n/I_o$ across the array elements, just as

$$\mathbf{l}_e(\hat{r}) = \hat{r}\times\hat{r}\times\int_L d\mathbf{l}\frac{I_e(\mathbf{r}')}{I_e}e^{jk\hat{r}\cdot\mathbf{r}'}$$

describes the Fourier transform of the current distributions within array elements.

The most common[13] current distribution for 1-D arrays is $W_n = We^{j\alpha n}$, where α is some real constant known as *phase increment*, and a constant W implies uniform current amplitudes across the array elements. Such antennas are known as *uniform phased-arrays* and, in general, their effective length

$$\mathbf{f}(\hat{r}) \propto \mathbf{l}_e(\hat{r}) \sum_{n=0}^{N-1} e^{jn(kD\cos\psi+\alpha)}.$$

The corresponding beam-pattern (for transmission and reception) is of the form

$$|\mathbf{f}(\hat{r})|^2 = |\mathbf{l}_e(\hat{r})|^2|F(\psi)|^2,$$

[12] Note the additional phase offset $nkD\cos\psi$ in the field contributions from elements offset from the origin by a distance nD. The offset term is only accurate for distances $r \gg ND$ as discussed in Section 6.4.

[13] The design of array *feed networks* to implement arbitrary current distributions W_n is an engineering challenge. In principle array antennas with arbitrary W_n's can be operated as follows: in transmission each element is driven by its own source with an input current $W_n I_o$; in reception the individual antenna outputs V_n are recorded (often in digital form) and combined mathematically (in software) to produce $V = \sum W_n V_n$.

with an array factor[14]

$$F(\psi) = \sum_{n=0}^{N-1} e^{jn(kD\cos\psi+\alpha)}$$

and an *array pattern* $|F(\psi)|^2$. Clearly, a choice of phase increment

$$\alpha = -kD\cos\psi_o$$

insures a peak response of $F(\psi) = N$ in direction $\psi = \psi_o$ — e.g., if the peak of $|F(\psi)|^2$ is desired in $\psi = 30^\circ$ direction, then the required

$$\alpha = -kD\cos 30^\circ = -\frac{1}{2}kD = -\pi D/\lambda.$$

Using the summation formula

$$\sum_{n=0}^{N-1} w^n = \frac{1-w^N}{1-w}$$

with $w = e^{j(kD\cos\psi+\alpha)}$, we find that

$$|F(\psi)| = \left|\frac{1-e^{jN(kD\cos\psi+\alpha)}}{1-e^{j(kD\cos\psi+\alpha)}}\right| = \frac{|\sin(\frac{N}{2}(kD\cos\psi+\alpha))|}{|\sin(\frac{1}{2}(kD\cos\psi+\alpha))|}$$

in general. L'Hospital's rule can now be used to show that the array pattern $|F(\psi)|^2$ maximizes for angles ψ satisfying the constraints

$$kD\cos\psi + \alpha = 0, \pm 2\pi, \pm 4\pi, \cdots \quad \text{and} \quad |\cos\psi| \le 1.$$

Depending on the numerical value for kD, only a finite number of $|F(\psi)|^2$ peaks will be obtained in the physically meaningful range of $0 \le \psi \le 180^\circ$ for angle ψ.

Example: Figures 2.17a, b, and c show the array patterns $|F(\psi)|^2$ of 1-D uniform phased-arrays with (a) $D = \frac{\lambda}{4}$, $\alpha = -\frac{\pi}{2}$ rad, $N = 8$; (b) $D = \frac{\lambda}{2}$, $\alpha = 0$, $N = 8$; $D = \frac{\lambda}{2}$, $\alpha = -\frac{\pi}{4}$ rad, $N = 16$. The plotted patterns are

$$|F(\psi)|^2_{a,b,c} = \frac{|\sin(\frac{8}{2}(\frac{\pi}{2}\cos\psi - \frac{\pi}{2}))|^2}{|\sin(\frac{1}{2}(\frac{\pi}{2}\cos\psi - \frac{\pi}{2}))|^2}, \ \frac{|\sin(\frac{8}{2}(\pi\cos\psi))|^2}{|\sin(\frac{1}{2}(\pi\cos\psi))|^2}, \ \frac{|\sin(\frac{16}{2}(\pi\cos\psi - \frac{\pi}{4}))|^2}{|\sin(\frac{1}{2}(\pi\cos\psi - \frac{\pi}{4}))|^2},$$

respectively.

The beam pattern $|F(\psi)|^2_a$ shown in Figure 2.17a is called "endfire" because it shows a peak response along the array axis (i.e, $\psi = 0$ direction). In reception the array is mainly sensitive to waves propagating in -x-direction towards the antenna, whereas in transmission the strongest outgoing Poynting flux is in x-direction. You can visualize the 3-D beam pattern of the array by revolving Figure 2.17a in your mind round the horizontal axis (see also Figure 2.6b).

[14]Here we assume $W = 1$ for convenience, which will leave the magnitude of $\mathbf{f}(\hat{r})$ uncalibrated unless $W = 1$ is true. An uncalibrated magnitude for $\mathbf{f}(\hat{r})$ is of no concern in gain and effective area calculations shown in Section 6.3.

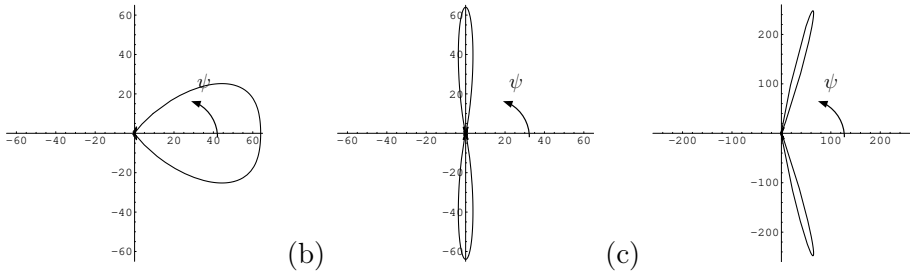


Figure 2.17: Polar diagrams of $|F(\psi)|^2$ versus ψ for 1-D uniform phased-arrays with elements along the x-axis ($\psi = 0$ direction in the plots): (a) an 8-element endfire array (see Figure 2.6b for a 3-D plot of the same beam pattern), (b) an 8-element broadside array, and (c) a 16-element array with the beam steered away from the broadside direction.

Pattern $|F(\psi)|_b^2$ shown in Figure 2.17b is called "broadside" because it is directed perpendicular to the array axis.

Finally, pattern $|F(\psi)|_c^2$ shown in Figure 2.17c has been steered away from broadside direction as a consequence of a non-zero α. Also the pattern is narrower (smaller solid angle) than the pattern $|F(\psi)|_b^2$ because array (c) is physically longer than array (b).

Although all the results above were obtained for 1-D arrays with x-axis elements, the formulae developed can also be used for arrays with y- or z-axes elements, provided that direction cosine $\cos\psi$ in the equations is re-interpreted as follows: Since ψ stands for the angle from the array axis to the observing direction

$$\hat{r} = \sin\theta\cos\phi\hat{x} + \sin\theta\sin\phi\hat{y} + \cos\theta\hat{z},$$

it follows that, in general,

$$\cos\psi = \begin{cases} \hat{x}\cdot\hat{r} = \cos\psi_x \equiv \sin\theta\cos\phi & \text{for } x-\text{axis arrays,} \\ \hat{y}\cdot\hat{r} = \cos\psi_y \equiv \sin\theta\sin\phi & \text{for } y-\text{axis arrays,} \\ \hat{z}\cdot\hat{r} = \cos\theta & \text{for } z-\text{axis arrays;} \end{cases}$$

the direction cosine*s* $\cos\psi_x$, $\cos\psi_y$, and $\cos\theta$ satisfy an identity

$$\cos^2\psi_x + \cos^2\psi_y + \cos^2\theta = 1;$$

furthermore, just as the zenith angle θ increases in the direction of unit vector

$$\hat{\theta} = \frac{(xz, yz, z^2 - r^2)}{r\sqrt{x^2+y^2}},$$

angles ψ_x and ψ_y increase in directions

$$\hat{\psi}_x \equiv \frac{(x^2 - r^2, yx, zx)}{r\sqrt{y^2+z^2}} \quad \text{and} \quad \hat{\psi}_y \equiv \frac{(xy, y^2 - r^2, zy)}{r\sqrt{z^2+x^2}},$$

respectively (as shown in a homework problem).

Exercise 1: (i) Assuming small $\cos^2\psi_{x,y}$ obtain an approximate formula for θ in radians in terms of direction cosines $\cos^2\psi_{x,y}$. (ii) If $\psi_y = 90^o$, show that $\theta \approx |\cos\psi_x|$ for small $\cos\psi_x$.

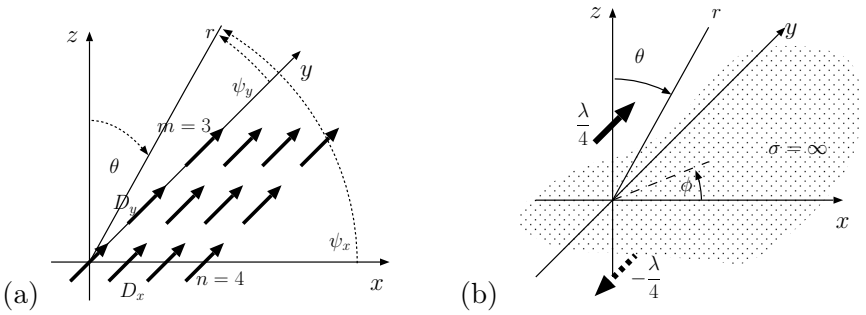


Figure 2.18: (a) A planar 2-D array with D_x and D_y element spacings along x- and y-axes, respectively—note that the 2-D array can be viewed as a 1-D y-axis array ($M = 3$ elements) of 1-D x-axis arrays ($N = 4$ elements). (b) The sketch of an antenna element located $\frac{\lambda}{4}$ above a reflecting ground plane ($z = 0$) and its "image" below the ground. The image is envisioned as the source of a reflected field that cancels the tangential field on $z = 0$ plane. This requires the image to be polarized as shown in the diagram (opposite to the element), giving rise to a 180^o phase difference between the element and image input currents.

2.6.2 Multi-D arrays

The elements of an antenna array need not be constrained to a single axis. Clearly, 2- or 3-D arrays covering a plane or a volume can be constructed. Figure 2.18a, for instance, shows a planar 2-D array with D_x and D_y element spacings along x- and y-axes, respectively. Assuming a progressive phasing of the array elements with α_x and α_y increments along x- and y-directions, respectively, the effective length of the array can be written as

$$\mathbf{f}(\hat{r}) = \mathbf{l}_e(\hat{r})F_x(\psi_x)F_y(\psi_y)$$

with

$$|F_{x,y}(\psi_{x,y})| = \frac{|\sin(\frac{N,M}{2}(kD_{x,y}\cos\psi_{x,y} + \alpha_{x,y}))|}{|\sin(\frac{1}{2}(kD_{x,y}\cos\psi_{x,y} + \alpha_{x,y}))|}.$$

This approach, known as *pattern multiplication*, is justified because the described 2-D array is really a 1-D y-axis array of "1-D x-axis arrays".

For an unphased array with $\alpha_x = \alpha_y = 0$, the array pattern $|F_x(\psi_x)|^2|F_y(\psi_y)|^2$ maximizes in $\theta = 0^o$ and 180^o directions. A plot of such a pattern with $N = M = 4$ and $D_{x,y} = \frac{\lambda}{2}$ is shown in Figure 2.19. The pattern in Figure 2.19 can be steered by using non-zero values of α_x and/or α_y.

If a planar array is constructed above a reflecting ground plane (e.g., $z = 0$), then the ground causes a change in the beam pattern of the array with respect to free space. Reflections from ground cause the radiation to vanish in $z < 0$ region and add to the original field in $z > 0$ hemisphere. Since a reflection can be viewed as radiation from an *image array* below the ground plane, the overall field in $z > 0$ region can be computed

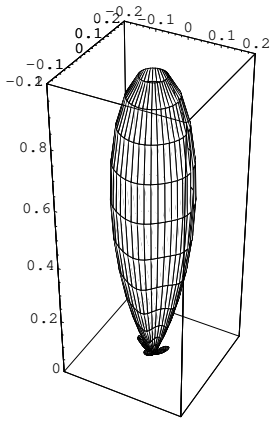


Figure 2.19: A self-normalized plot of array pattern $|F_x(\psi_x)|^2|F_y(\psi_y)|^2$ for a 4×4 unphased array with $\frac{\lambda}{2}$ element spacings shown over only the upper hemisphere, $z > 0$. Such an array placed above a reflecting plane would have a single beam like the one shown here, but only over the upper hemisphere (as discussed in the text). Such one-sided beams are known as "pencil beams" and are desirable for high resolution remote sensing applications.

using pattern multiplication once again. This results in

$$\mathbf{f}(\hat{r}) = \mathbf{l}_e(\hat{r})F_x(\psi_x)F_y(\psi_y)F_z(\theta)$$

for $z > 0$, where $F_z(\theta)$ is the array factor of a 1-D z-axis array formed by an array element and its image below the ground (see Figure 2.18b and the caption). Since the element and its image are 180^o out of phase, $\alpha_z = -\pi$ rad, and with $D_z = \frac{\lambda}{2}$, corresponding to an array $\frac{\lambda}{4}$ above the ground,

$$|F_z(\theta)| = \frac{|\sin(\frac{2}{2}(k\frac{\lambda}{2}\cos\theta - \pi))|}{|\sin(\frac{1}{2}(k\frac{\lambda}{2}\cos\theta - \pi))|} = \frac{|\sin(\pi(\cos\theta - 1))|}{|\sin(\frac{\pi}{2}(\cos\theta - 1))|} \approx 2 \text{ for small } \theta.$$

Therefore, the field intensity of a 2-D array is effectively doubled in the upper hemisphere near $\theta = 0$ and lower hemisphere lobes vanish. Overall, the ground plane doubles the directivity of the array.

Exercise 1: Show that the array factor $F(\psi_x, \psi_y)$ of an arbitrary 2-D array with element locations (nD_x, mD_y) and an arbitrary current distribution $W_{n,m}$ can be expressed as

$$F(\psi_x, \psi_y) = \sum_{n=0}^{N-1}\sum_{m=0}^{M-1} W_{n,m}e^{j(nkD_x\cos\psi_x + mkD_y\cos\psi_y)},$$

assuming a $N \times M$ grid for the array elements and grid spacings of D_x and D_y in x and y directions, respectively.

Exercise 2: Plot the 2-D array pattern $|F(\psi_x, \psi_y)|^2$ expression from Exercise 1 (in dB) for $N = M = 64$, $D_x = D_y = 6\lambda$, and $W_{n,m} = e^{j\pi(n+m)/4}$, as a function of direction cosines $\theta_x \equiv \cos\psi_x$ and $\theta_y \equiv \cos\psi_y$. Hint: use a *Mathematica* or MATLAB based 2-D FFT routine to compute $F(\psi_x, \psi_y)$ over a suitably defined grid in θ_x and θ_y with sufficient angular resolution. Be careful about the sign convention of the FFT routine so that the pointing direction of the antenna beam is determined correctly.

Exercise 3: What would be the physical area of the array examined in Exercise 2 for a frequency of $f = 50$ MHz?

2.6.3 Gain, directivity, and beam solid-angle for arrays

The gain and directivity of arrays can be calculated starting with the usual gain formula

$$G = \frac{4\pi|\mathbf{f}|^2}{\int d\Omega|\mathbf{f}|^2}.$$

With a complicated θ, ϕ dependence for $|\mathbf{f}|^2$, the solid angle integral $\int d\Omega|\mathbf{f}|^2$ is usually difficult to evaluate. An alternative approach is to first *estimate* the beam solid angle

$$\Omega_o \equiv \int d\Omega|\mathbf{f}|^2/|\mathbf{f}|^2_{max},$$

possibly using a graphical approach, and then use the relation

$$G = D|\mathbf{f}|^2/|\mathbf{f}|^2_{max} \quad \text{with} \quad D = \frac{4\pi}{\Omega_o}.$$

As pointed out earlier, the beam solid angle Ω_o relates the antenna beam shape to its directivity D. For an isotropic radiator with $D = 1$, the solid angle is $\Omega_o = 4\pi$, which is also the surface area of a unit-radius sphere, i.e., the shape of the beam of an isotropic radiator. In general, for an arbitrary antenna, Ω_o corresponds to an effective area on the surface of a unit-radius sphere which is illuminated by the antenna beam.

For instance, the beam of an unphased ($\alpha_x = \alpha_y = 0$) 2-D array with

$$|F_x(\psi_x)||F_y(\psi_y)| = \frac{|\sin(\frac{N}{2}kD_x\cos\psi_x)|}{|\sin(\frac{1}{2}kD_x\cos\psi_x)|}\frac{|\sin(\frac{N}{2}kD_y\cos\psi_y)|}{|\sin(\frac{1}{2}kD_y\cos\psi_y)|}$$

illuminates its unit sphere mostly between the "first nulls" of its array factor, i.e., between angles $\pm\theta$ such that $\frac{N}{2}kD_{x,y}\sin\theta = \pi\frac{ND_{x,y}}{\lambda}\sin\theta = \pi$ in both $\phi = 0$ and 90^o planes. The corresponding solid angle Ω_o of beam pattern $\propto |F_x(\psi_x)|^2|F_y(\psi_y)|^2$ is slightly less than a product of the *beam width between first nulls*, i.e.,

$$\text{BWFN} = 2\sin^{-1}(\frac{\lambda}{ND_{x,y}})$$

defined for the same two planes. A more accurate estimate for Ω_o turns out to be the product of *half-power beam widths*[15]

$$\text{HPBW} \approx \sin^{-1}(\frac{\lambda}{ND_{x,y}}).$$

[15]Exact definition of the array HPBW is as follows: HPBW $= 2\theta$, where $\sqrt{2}\sin(\pi\frac{ND_{x,y}}{\lambda}\sin\theta) = N\sin(\pi\frac{D_{x,y}}{\lambda}\sin\theta)$. However, HPBW $\approx \frac{1}{2}$BWFN works well for arrays with $ND_{x,y} \gg \lambda$.

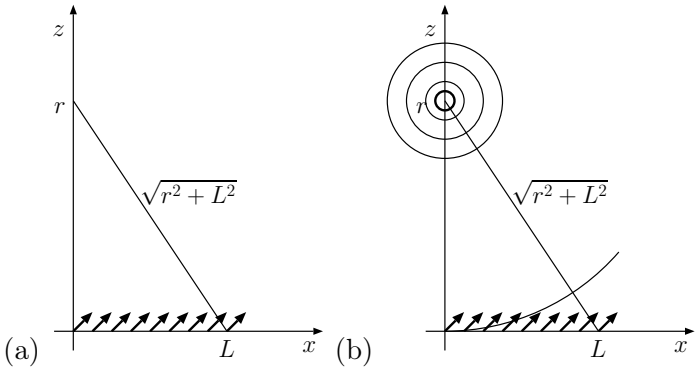


Figure 2.20: (a) A 1-D array of a physical dimension L is depicted for which the usual far-field formulae derived so far are valid only if transmission distance $r > \frac{2L^2}{\lambda} \equiv R_{ray}$ (see text), and (b) the geometry of a spherical wave emitted by a source at distance r towards the same array.

Thus, for $ND_{x,y} \gg \lambda$,

$$\Omega_o \sim \sin^{-1}(\frac{\lambda}{ND_x})\sin^{-1}(\frac{\lambda}{MD_y}) \approx \frac{\lambda^2}{(ND_x)(MD_y)} = \frac{\lambda^2}{A_p},$$

where $A_p \equiv (ND_x)(MD_y)$ is the physical area of the 2-D array. The term aperture area is also used commonly to refer to A_p.

For large 2-D arrays the maximum possible directivity (which is achieved with no phasing) is therefore[16]

$$D = \frac{4\pi}{\Omega_o} \sim \frac{4\pi}{\lambda^2}A_p.$$

For *phased* 2-D arrays, on the other hand,

$$D \sim \frac{4\pi}{\lambda^2}A_p\epsilon_a,$$

where $\epsilon_a \leq 1$ is known as *aperture efficiency* and decreases with increasing beam steering angle θ.

Overall, 2-D antenna arrays with large physical areas $A_p \gg \lambda^2$ will have large directivities but exact determination of D and $G(\theta, \phi)$ will require the determination of aperture efficiency ϵ_a.

2.6.4 Rayleigh distance and plane-wave approximation

The array formulae derived in this section are strictly valid only if the distance r, where the radiation field $\mathbf{E}(\mathbf{r})$ is being calculated, is sufficiently large compared to the physical size L of the array. A possible geometry involving r and L is illustrated in Figure 2.20a,

[16]This result turns out to be exact for arrays with $\frac{\lambda}{2}$-dipole elements with $D_x = D_y = \frac{\lambda}{2}$ spacings. Such arrays are called *filled* arrays.

(a)  (b)  (c) 

Figure 2.21: (a) Augustin Jean Fresnel (1788-1827), (b) John William Strutt Lord Rayleigh (1842-1919), and (c) Joseph von Fraunhofer (1787-1826).

which we will use to examine by what extent r should exceed L in order for our array formulae to be considered accurate enough.

The standard far field formula for $\mathbf{E}(\mathbf{r})$ derived earlier effectively neglects the phase difference due to paths r and $\sqrt{r^2+L^2}$ shown in Figure 2.20a. If the neglected phase difference $k(\sqrt{r^2+L^2}-r)$ is small in radians, e.g., if

$$k(\sqrt{r^2+L^2}-r) < \frac{\pi}{2}$$

the corresponding approximation will have a minor impact on $\mathbf{E}(\mathbf{r})$ calculations. For $r \gg L$ we can re-write the same condition as

$$kr(\sqrt{1+(\frac{L}{r})^2}-1) \approx \frac{kL^2}{2r} < \frac{\pi}{2},$$

or, equivalently, as

$$r > \frac{kL^2}{\pi} = \frac{2L^2}{\lambda} \equiv R_{ray},$$

where

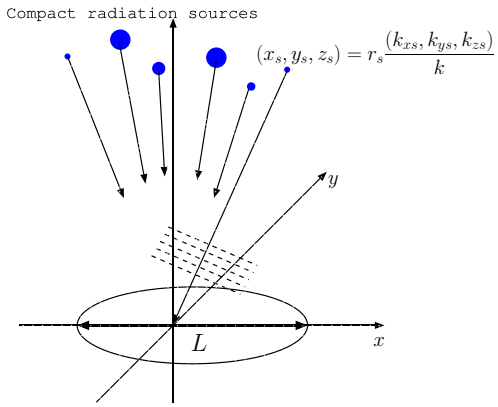
$$R_{ray} \equiv \frac{2L^2}{\lambda}$$

is known as *Rayleigh distance*.

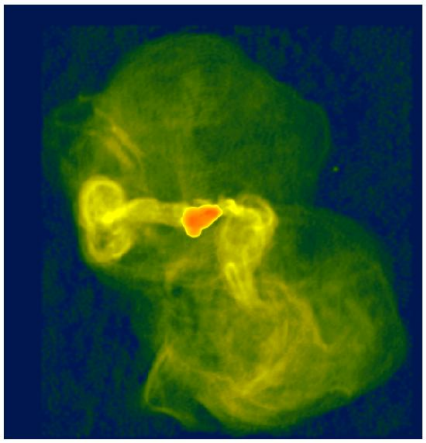
Consequently the array formula for $\mathbf{E}(\mathbf{r}) \propto \mathbf{f}(\hat{r})$ should be accurate for $r > R_{ray}$ in the array *far-field*, also known as the *Fraunhofer region*. By contrast, in the *Fresnel region,* $r < R_{ray}$, the radiation field of the array is not accurately described $\mathbf{E}(\mathbf{r}) \propto \mathbf{f}(\hat{r})$. Clearly, $r \sim R_{ray}$ corresponds a transition region.

The Rayleigh distance of an array of size L has another useful interpretation which can be described with the help of Figure 2.20b. The figure shows a point source located at a distance r emitting a spherical wave at some wavelength λ. Although the incident field on the array has spherical wavefront, the corresponding curvature will be small if $r \gg L$ (i.e., the wavefront will appear straight across the scale L), or , equivalently, the difference of phases detected by individual array elements can be neglected because

$$k(\sqrt{r^2+L^2}-r) < \frac{\pi}{2}.$$



(a)



(b)

Figure 2.22: (a) Spherical radiation arriving from compact sources at large distances $r_s > \frac{2L^2}{\lambda} = R_{ray}$ can be modeled as plane-waves across a measurement platform or an *observatory* of a size L. (b) An image of a radio galaxy (M87, the central galaxy of the Virgo Cluster, constructed with VLA radio telescope, see http://www.seds.org/messier/more/m087_nrao.html) obtained with radio imaging techniques.

This, leads once again to $r > R_{ray}$, now as a condition for *plane-wave* approximation of the incident field. That is, radiation from point sources onto a measurement platform can be regarded as plane waves when the sources are located within the Fraunhofer region as suggested in Figure 2.22a.

2.6.5 Radio imaging

Consider a large number of identical and identically polarized antennas at positions (x, y) within a measuring platform as in Figure 2.22a responding to a superposition of plane wave fields at some specified wave frequency ω (arriving from compact sources located within the Fraunhofer region of the platform). The open-circuit voltage of each antenna can be expressed as

$$V(x,y) = \sum_s \mathbf{l}(\hat{k}_s) \cdot \mathbf{E}_s e^{j(k_{xs}x + k_{ys}y)} = \sum_s v_s e^{j(k_{xs}x + k_{ys}y)},$$

where $\mathbf{l} = \mathbf{f}$ is the effective antenna length, and

$$v_s \equiv \mathbf{l}(\hat{k}_s) \cdot \mathbf{E}_s$$

is a voltage phasor due to source at $r_s\mathbf{k}_s/k$ detected by the central antenna at the origin. Notice that

$$V(0,0) = \sum_s v_s,$$

and, therefore, it is not possible to obtain estimates of individual source strengths $\propto |v_s|^2$, nor their directions of arrival $\hat{k}_s$, using a single antenna element. However, if many elements are employed, we can find suitable "imaging procedures" to picture the $|v_s|^2$ distributions across the sky, like the one shown in Figure 2.22b.

We will first consider a 1-D imaging system like one depicted in Figure 2.23a, with individual antenna outputs

$$V_n \equiv V(nD, 0) = \sum_s v_s e^{jk_{xs}nD}.$$

Multiplying both sides of the V_n data by e^{jk_xDn} and summing the products over all n, we obtain

$$\begin{aligned} v(k_x) &= \sum_{n=0}^{N-1} V_n e^{jk_xDn} = \sum_{n=0}^{N-1} \{\sum_s v_s e^{jk_{xs}nD}\} e^{jk_xDn} \\ &= \sum_s v_s \sum_{n=0}^{N-1} e^{j(k_x+k_{xs})Dn} = \sum_s v_s \frac{1 - e^{jN(k_x+k_{xs})D}}{1 - e^{j(k_x+k_{xs})D}}, \end{aligned}$$

where $v(k_x)$ is clearly the discrete Fourier transform (DFT) of V_n sequence, which, according to the last expression on the right, is a linear superposition of shifted versions of an antenna array factor

$$F(k_x) \equiv \frac{1 - e^{jNk_xD}}{1 - e^{jk_xD}}.$$

Given that

$$|F(k_x + k_{xs})| = |\frac{\sin(\frac{N}{2}(k_x + k_{xs})D)}{\sin(\frac{1}{2}(k_x + k_{xs})D)}|$$

is sharply peaked about $k_x = -k_{xs}$ when N is large, we have, in the same limit,

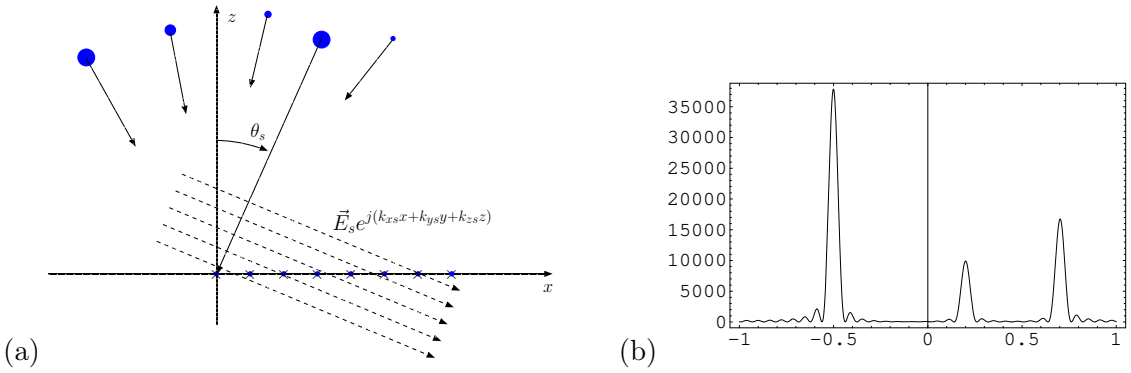
$$|v(k_x)|^2 \approx \sum_s |v_s|^2 |F(k_{xs} + k_x)|^2,$$

which is the *Fourier spectrum* of the antenna V_n data — here we are referring to $|\text{DFT}|^2$ as *spectrum* using a common practice in radio signal processing. A plot of the spectrum provides an inverted 1-D image of the source distribution as shown in Figure 2.23b.

The spectral image in 2.23b implies the existence of three distinct sources in the sky, corresponding to, say, $k_{x1} \approx 0.5k$, $k_{x2} \approx -0.2k$, and $k_{x3} \approx -0.7k$, with k_{x1} source being the strongest and k_{x2} the weakest (notice the sign inversions from spectral peaks to k_{xs} estimates). Spectral amplitude for each source goes as $N^2|v_s|^2$, and the angular resolution of the image is the same as the beam width of an N-element 1-D antenna array with D spacings, that is

$$\text{HPBW} = \frac{\lambda}{ND} = \frac{\lambda}{L} \,\text{rad}$$

for small zenith angles θ. Note that the antenna spacing D in the imaging system should be small enough to ensure a single peak per source across the field of view of the system



(a) (b)

Figure 2.23: (a) 1-D interferometer imaging system with N antennas at $x = nD$, $n \in [0, N-1]$, and (b) a possible source image $|v(k_x)|^2$ plotted versus k_x/k of a scene with three far field sources. The image illustrates the case for $N = 32$ and $D = \frac{\lambda}{2}$.

(which in turn is determined by the element pattern), e.g., $D = \frac{\lambda}{2}$, if an all-sky image is being formed.

A 2-D radio image of the sky (see Figure 2.22b) can be formed likewise by plotting versus k_x and k_y the squared magnitude of 2-D DFT of the antenna voltage data $V(x, y)$ collected over a rectangular grid. The angular resolution of a 2-D image would be the solid angle (or HPBW's) of a 2-D antenna array with equivalent dimensions.

The imaging procedures described above are nothing more than a *software based* — instead of hard-wired — implementation of antenna arrays discussed in the previous section (why and in what sense?) and are based on an implicit assumption of monochromatic and deterministic sources. However, there exists alternate imaging strategies applicable to scenes consisting of stochastic signal sources with fluctuating v_s values. Any image of a *stochastic scene* obtained with a squared DFT would be just *one realization* of an *ensemble* of possible source images, with an *ensemble average* (or *expected value*)

$$\langle |v(k_x)|^2 \rangle \approx \sum_s \langle |v_s|^2 \rangle |F(k_x + k_{xs})|^2.$$

While an estimate of the ensemble average image $\langle |v(k_x)|^2 \rangle$ can be obtained by averaging[17] many independent realizations of $|v(k_x)|^2$ in time, it may be more convenient, or even better, to use one of the alternate methods described next.

Observe that

$$\begin{aligned} |v(k_x)|^2 &= |\sum_n V_n e^{jk_x Dn}|^2 = (\sum V_n e^{jk_x Dn})^* (\sum_m V_m e^{jk_x Dm}) \\ &= \sum_n \sum_m V_n^* V_m e^{jk_x D(m-n)} \end{aligned}$$

[17] i.e., using a finite *exposure time*, using a well known photography term.

and, therefore, the expected value

$$\langle |v(k_x)|^2 \rangle = \sum_n \sum_m \langle V_n^* V_m \rangle e^{jk_x D(m-n)}.$$

Above, the expected value $\langle V_n^* V_m \rangle$ represents the *cross-correlation* of individual antenna outputs. But the output model

$$V_n = \sum_s v_s e^{jk_{xs} Dn}$$

with random phased v_s components implies[18] that the cross-correlation

$$\begin{aligned} \langle V_n^* V_m \rangle &= \sum_s \sum_{s'} \underbrace{\langle v_s^* v_{s'} \rangle} e^{-jnk_{xs} D} e^{jmk_{xs'} D} \\ &\qquad\qquad \langle |v_s|^2 \rangle \text{ if } s' = s, \text{ 0 otherwise.} \\ &= \sum_s \langle |v_s|^2 \rangle e^{jk_{xs} D(m-n)} \end{aligned}$$

is a function of only the antenna-to-antenna spacings $D(m-n)$ rather than being a function of two variables, *Dn and Dm*, separately. Therefore, we can adopt the notation

$$R_{m-n} \equiv \langle V_n^* V_m \rangle$$

for the cross-correlations, and since all R_{m-n} are equal to R_p, with $p \equiv m-n$, we can re-write

$$\langle |v(k_x)|^2 \rangle = \sum_n \sum_m \langle V_n^* V_m \rangle e^{jk_x D(m-n)} = \sum_n \sum_m R_{m-n} e^{jk_x D(m-n)}$$

as

$$\langle |v(k_x)|^2 \rangle = \sum_{p=-N}^{N} (N - |p|) R_p e^{jk_x Dp},$$

where the triangular weighting factor $N - |p|$ can be understood with the help of Figure 2.24a.

The last equation above is effectively the image formula which can be used with the measured values of the cross-correlations

$$R_p = \langle V_n^* V_{n+p} \rangle = \sum_s \langle |v_s|^2 \rangle e^{jk_{xs} Dp}.$$

Clearly, to measure $R_p = \langle V_n^* V_{n+p} \rangle$ for any given p we only need a single occurrence of antenna spacing Dp within the array, which, in turn, allows the use of *non-uniformly spaced arrays* like the one shown in Figure 2.24b. In fact, the spacing shown in Figure

[18] Here the assumption is, phases $\phi \equiv \angle v_s$ and $\phi' \equiv \angle v_{s'}$ are uniform distributed independent random variables. In detail, $\langle v_s^* v_{s'} \rangle = |v_s||v_{s'}| \langle e^{-j\phi} e^{j\phi'} \rangle = |v_s|^2$ if $s' = s$. Otherwise $\langle v_s^* v_{s'} \rangle = 0$ because $\langle e^{-j\phi} e^{j\phi'} \rangle = \langle e^{-j\phi} \rangle \langle e^{j\phi'} \rangle = 0$ given that ϕ and ϕ' are independent and uniform distributed phases.

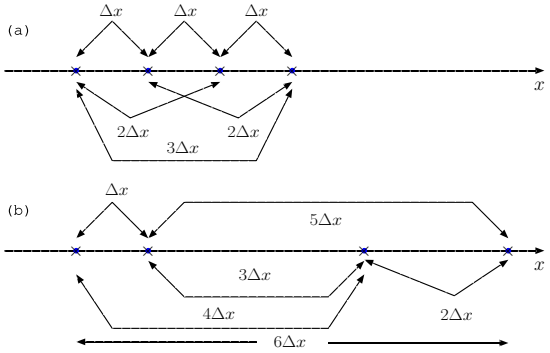


Figure 2.24: (a) A 4-element system with equal spacings will produce 4 estimates of R_0, 3 estimates of R_1, 2 estimates of R_2, and a single estimate of R_3). (b) The alternate spacing of 4 elements shown here provides all estimates of R_0 through R_6 with no gaps, and, therefore, higher resolution images than the system in (a).

2.24b is preferable to the spacing in Figure 2.24a, since an image resolution of a 7-element array can be obtained using 2.24b. Also, the triangular weighting function $N - |p|$ in the image formula above can be replaced by alternative window functions (e.g., rectangle, raised cosine, etc.) if so desired, with familiar impacts on the image quality (sidelobe vs main lobe tradeoffs, etc.).

For imaging with *missing* R_p data, we first define a *brightness distribution*

$$B(k_x) \equiv \sum_s 2\pi\delta(k_x + k_{xs})\langle|v_s|^2\rangle$$

such that the antenna cross-correlations can be expressed in the following form:

$$R_p = \int \frac{dk_x}{2\pi} B(k_x) e^{-jk_x Dp}.$$

Note that the brightness distribution is, in effect, the "infinite-resolution" source image[19] that we would like to see. To obtain an estimate of $B(k_x)$ — of some *finite* resolution, of course — some suitable *initial guess* for $B(k_x)$ can be declared and then adjusted iteratively until a *global misfit*

$$\chi^2 \equiv \sum_p \sigma_p^{-2} |R_p - \int \frac{dk_x}{2\pi} B(k_x) e^{-jk_x Dp}|^2$$

between the available R_p *data* and R_p *models* $\int \frac{dk_x}{2\pi} B(k_x) e^{-jk_x Dp}$ (written in practice as some Riemann sum of the integral over a suitable grid in k_x) is minimized. The weight

[19] Infinite resolution but distorted by the element pattern of the imaging system since $v_s \equiv \mathbf{l} \cdot \mathbf{E}_s$ where $\mathbf{l}$ is the effective length of the array element. On the other hand, the brightness distribution is a more *fundamental* way of describing a radio scene than the point-source model we started with — it is proper to think of the point-source model as a particular type of brightness distribution.

coefficients σ_p^{-2}'s are generally taken as the variance of the cross-correlation magnitudes $|R_p|$.

Likewise, in the 2-D case, an image $B(k_x, k_y)$ can be estimated in the *weighted least-squares* sense just describe by using the 2-D cross-correlation model[20]

$$R_{p,q} = \int \frac{dk_x}{2\pi} \int \frac{dk_y}{2\pi} B(k_x, k_y) e^{-jD(k_x p + k_y q)}$$

for antenna pair spacings pD and qD along x and y, respectively. In least-squares image estimation (in both 1- and 2-D) it is common to employ *regularization* methods to enhance the image resolution.

2.6.6 Aperture synthesis and SAR imaging of stationary targets

High resolution images of stationary sources can also be formed by moving a single antenna across a suitably sized measurement platform. Assuming that the antenna position varies as

$$x = ut,$$

the DFT of a sequence of voltage outputs

$$V_n = \sum_s v_s e^{jk_{xs}nuT}$$

taken at sampling intervals T yields

$$v(\omega) \equiv \sum_{n=0}^{N-1} V_n e^{j\omega nT} = \sum_s v_s \frac{1 - e^{jN(\omega + k_{xs}u)T}}{1 - e^{j(\omega + k_{xs}u)T}} \equiv \sum_s v_s F(\omega + k_{xs}u),$$

where u is the antenna velocity. Clearly, for a sufficiently large N, the Fourier spectrum

$$|v(\omega)|^2 \approx \sum_s |v_s|^2 |F(\omega + k_{xs}u)|^2$$

can be interpreted as a 1-D source image, with each source characterized by a peak at a frequency given as $\omega = -k_{xs}u$ from which k_{xs} and source direction can be inferred. Since $\omega = -k_{xs}u$ represents a *Doppler frequency shift* of the incident field due to the antenna motion u, it is common to refer to $|v(\omega)|^2$ as *Doppler* spectrum (or, periodogram, in more proper terms). The angular resolution of the image for small θ can be shown to be $\frac{\lambda}{L}$, where $L \equiv uNT$, the width of the antenna aperture synthesized over the observation time NT. The analysis presented here is valid for sources at distances $r_s > R_{ray} = \frac{2L^2}{\lambda}$ as usual, which clearly places a limit on maximum usable observation time NT.

The imaging procedure just described, where only discrete-time samples of antenna voltages are needed, can be utilized in *active* imaging applications where the antenna is also used to transmit short pulses of monochromatic waves at T intervals. In such

[20] Known as *van Cittert-Zernike theorem.* See Thompson, A. R., J. M. Moran, G. W. Swenson, *Interferometry and synthesis in radio astronomy*, Wiley-Interscience, 1986.

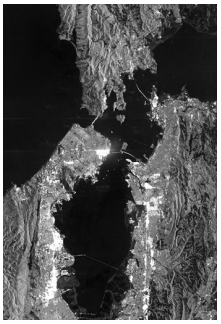
(a)  (b) 

Figure 2.25: (a) A satellite based SAR image of San Fransisco Bay area, and (b) Christian Andreas Doppler (1803-1853)
(http://www.jpl.nasa.gov/radar/sircxsar/)

applications — known as *synthetic aperture radar* (SAR) imaging — the imaged signals are the echoes of the transmitted pulses from reflecting targets situated in the antenna far-field. Using a *pulse length* δt much shorter than the *inter-pulse period* (IPP) T, it becomes possible to obtain a sequence of 1-D images representing the angular distributions of reflecting targets situated at different distances r to the radar. The resolution of such images in *radar range* r works out to be $\delta r = \frac{c\delta t}{2}$ (as shown in the next chapter) while the angular resolution remains $\frac{\lambda}{L} = \frac{\lambda}{uNT}$ as established above. The image pertinent for any radar range r is formed from the DFT of a V_n sequence formed by sampling the antenna output after a delay of $t = \frac{2r}{c}$ after each pulse transmission. 2-D images like the one shown in Figure 2.25a are then formed by stacking the Doppler spectra of distinct V_n sequences representing different ranges δr apart.

The term *range-Doppler analysis* is used to refer to the SAR imaging procedure outlined above. Note that in SAR imaging returns from a target at zenith θ_s peaks at a frequency $\omega = -2k_{xs}u = -2k\sin\theta_s u$ instead of $\omega = -k_{xs}u$ because of the motion of the transmitter with respect to the target. In other words, Doppler effect occurs twice in SAR, first from the transmitter to the target, and then from target to the receiver (which is same as the transmitter). Radar applications with moving targets will be discussed in the next chapter.

3 Ionospheric propagation and vertical soundings

In this chapter we start examining plane-wave solutions of Maxwell's equations in ionospheric plasmas. The term *plasma* refers to partially or fully ionized gases such as the Earth's upper atmosphere and interplanetary medium. The partially ionized portion of the upper atmosphere above ∼50 km height is known as the *ionosphere* and facilitates intercontinental radio communications by causing the reflection and ducting of radiowaves at HF and lower frequencies. Also the same medium impacts radio signals exchanged with communication satellites. When the full complexity of ionospheric plasma is taken into account the medium appears as *dispersive*, *inhomogeneous*, *anisotropic*, as well as *lossy*. However, the extent of these propagation effects depends on radiowave frequency and often only one or two of them are dominant in a given setting.

In this chapter we disregard the anisotropy effects caused by the Earth's DC magnetic field, and focus our attention primarily on vertical-incidence propagation in a horizontally stratified ionosphere. We begin with dispersion effects under a cold plasma approximation, and then examine propagation and reflections in an inhomogeneous ionosphere, describe an important vertical sounding technique — the ionosonde — used in ionospheric research, and, finally, conclude the chapter with a discussion of ionospheric absorption caused by electron-neutral collisions.

3.1 TEM plane waves in the upper-atmosphere

We know from Chapter 1 that the plane-wave form of Ampere's law can be written as

$$-j\mathbf{k}\times\mathbf{H}=\mathbf{J}+j\omega\mathbf{D}=\sigma\mathbf{E}+j\omega\epsilon_o\mathbf{E}+j\omega\epsilon_o\chi\mathbf{E}=j\omega\epsilon_o n^2\mathbf{E}, \tag{3.1}$$

where

$$n=\sqrt{1+\chi+\frac{\sigma}{j\omega\epsilon_o}} \tag{3.2}$$

denotes the *refractive index* of a medium with $\mu=\mu_o$ in terms of *susceptibility* χ due to bound charge carriers, and *conductivity* σ due to free charge carriers.

In a partially ionized gas χ in (3.2) is mainly due to neutral atoms and molecules in the gas (see Appendix 2) whereas σ is mainly dominated by the presence of free-electrons which are light and, therefore, respond strongly to the applied electric fields. In the atmosphere the neutral density decreases exponentially with increasing height so that at upper heights, where the number density of electrons is relatively large, χ can be neglected in the equations above. Conversely, at lower altitudes, and, in particular,

below 50 km, the free-electron density is negligibly small in the atmosphere so that the σ term above can be safely ignored. However, χ is always small, even in this region, to the extent that above equation (in the absence of σ) can be expressed as $n = 1 + \frac{1}{2}\chi$ all the way to the ground, where, at $z = 0$ level, $\frac{1}{2}\chi$ is on the order of a few hundred parts in a million. Leaving a discussion of the impact of the height derivative of this small quantity to Chapter 3, we will now focus on the refractive index model

$$n = \sqrt{1 + \frac{\sigma}{j\omega\epsilon_o}} \tag{3.3}$$

pertinent for the ionosphere (when the effect of the DC magnetic field permeating the ionosphere can be neglected).

The scalar conductivity σ required in (3.3) can be modeled in a number of different ways with varying degrees of accuracy. We will next derive a simplified model for σ using the *cold plasma* approximation. In this approximation all the free electrons in the medium are envisioned to be stationary in the absence of an applied AC electric field (thus the model overlooks the fact that free electrons in an actual plasma will be moving around with random thermal velocities in a frame of reference where the plasma as a whole can be regarded stationary).

When their random motions are ignored the dynamics of electrons are governed by Newton's 2nd Law, namely

$$\frac{d}{dt}(m\mathbf{v}) = \mathbf{F}$$

where m is electron mass, $\mathbf{v}$ the electron velocity, and $\mathbf{F}$ the total force on the electron. In non-relativistic limit $|\mathbf{v}| \ll c$, and for $\mathbf{F} = -e\mathbf{E}$, this equation is reduced to

$$m\frac{d\mathbf{v}}{dt} = -e\mathbf{E}. \tag{3.4}$$

Although the plasma conductivity σ can be easily deduced from the phasor form of (3.4) — as we will ultimately do — it is useful to view (3.4) as a special case ($\nu = \omega_o = 0$) of a 2nd-order differential equation

$$m\frac{d^2\mathbf{r}}{dt^2} + m\nu\frac{d\mathbf{r}}{dt} + m\omega_o^2\mathbf{r} = -e\mathbf{E} \tag{3.5}$$

for a position vector $\mathbf{r}$ of an electron with a velocity $\mathbf{v} = \frac{d\mathbf{r}}{dt}$. This damped harmonic-oscillator equation describes the motion of an electron subject to additional friction and spring-like restoring forces corresponding to the negative of the second and third terms on the left, respectively. Its solution $\mathbf{r}$ provides a model for the displacement of a bound electron from its atomic nucleus, the basis of the so-called *classical electron theory of matter*[1] [e.g., *Bekefi and Barret*, 1984] which provides a susceptibility model χ for

[1] Also known as *Lorentz* or *Lorentz-Drude* model — basically atoms and molecules are envisioned as dipole oscillators (or collections of them) described by (3.5), where ω_o and ν are chosen to correspond to absorption/emission line frequencies and bandwidths of relevant atoms/molecules. The model was constructed before a detailed understanding of atomic structure became available. The term *Lorentzian* originates from this work and describes the shape of the imaginary part of χ.

dielectric materials as a function of ν and ω_o parameterizing the additional forces. We will not make use of that approach to obtain our working model for atmospheric χ, but still follow it to a point where the distinctions between free- and bound-electron effects on radiowaves become clear. We will next obtain a phasor solution of (3.5) and use the result with $\nu = \omega_o = 0$ to obtain the free-electron response.

In phasor form (3.5) is modified as[2]

$$m(j\omega)^2\mathbf{r} + m\nu j\omega\mathbf{r} + m\omega_o^2\mathbf{r} = -e\mathbf{E}, \tag{3.6}$$

from which it follows that

$$\mathbf{r} = \frac{-e}{m(\omega_o^2 + j\nu\omega - \omega^2)}\mathbf{E}. \tag{3.7}$$

Thus, for free-electrons, the displacement

$$\mathbf{r} = \frac{e}{m\omega^2}\mathbf{E}, \tag{3.8}$$

velocity

$$\mathbf{v} = j\omega\mathbf{r} = \frac{j\omega e}{m\omega^2}\mathbf{E} = \frac{je}{m\omega}\mathbf{E},$$

and multiplying both sides by $-eN$, where N is number of free electrons per unit volume, we obtain

$$\mathbf{J} \equiv -Ne\mathbf{v} = -\frac{jNe^2}{m\omega}\mathbf{E} = \sigma\mathbf{E}, \tag{3.9}$$

where

$$\sigma = \frac{Ne^2}{j\omega m}. \tag{3.10}$$

is the AC conductivity in cold plasma approximation. Note that we have effectively ignored the current contributions of positive ions in writing the current density as $\mathbf{J} = -eN\mathbf{v}$ in terms of electron motions only, which is very well justified in practice since $\mathbf{v} \propto m^{-1}$ and the ions are thousands of times more massive than the electrons.

A parallel calculation for the bound electrons yields an oscillation velocity

$$\mathbf{v}_b = j\omega\mathbf{r} = j\omega\frac{-e}{m(\omega_o^2 + j\nu\omega - \omega^2)}\mathbf{E}, \tag{3.11}$$

and a polarization current

$$\mathbf{J}_p \equiv j\omega\mathbf{P} = -eN_b\mathbf{v}_b = j\omega\epsilon_o\frac{N_be^2}{m\epsilon_o(\omega_o^2 + j\nu\omega - \omega^2)}\mathbf{E} = j\omega\epsilon_o\chi\mathbf{E},$$

where

$$\chi = \frac{N_be^2}{m\epsilon_o(\omega_o^2 + j\nu\omega - \omega^2)}.$$

[2]Notice that according to (3.6), in sinusoidal steady-state, the electric force $-e\mathbf{E}$ is balanced by the restoring force $-m\omega_o^2\mathbf{r}$ in the case of bound electrons *as opposed to* the inertial force $m\omega^2\mathbf{r}$ in the case of free electrons. Since $\omega \ll \omega_o$ in the RF band, the free electron response $\mathbf{r}$ far exceeds the bound electron response. Also the free electron response is in the opposite direction with respect to bound electron response.

In cases with multiple bound electrons with different resonance and damping parameters, the classical susceptibility model is extended as a sum

$$\chi = \frac{N_b e^2}{m\epsilon_o} \sum_i \frac{f_i}{(\omega_{oi}^2 + j\nu_i\omega - \omega^2)},$$

where f_i are weighting coefficients related to the relative strengths of different resonances. Parameters ν_{oi}, ω_{oi}, and f_i of this extended model are constrained by quantum physics of atomic and molecular structures.

In terms of the σ and χ models above, we can re-write the Ampere's law (3.1) as

$$-j\mathbf{k}\times\mathbf{H} = \overbrace{\frac{Ne^2}{j\omega m}}^{\sigma}\mathbf{E} + j\omega\epsilon_o\mathbf{E} + j\omega\epsilon_o \overbrace{\sum_i \frac{N_b f_i e^2}{m\epsilon_o(\omega_{oi}^2 + j\nu_i\omega - \omega^2)}}^{\chi}\mathbf{E}. \tag{3.12}$$

Note that the conduction current and polarization current terms in (3.12) approach the $\omega \to \infty$ limit in a similar manner. However, for radiowave frequencies $\omega < \omega_{oi}$ — typically ω_{oi} are found in infrared or at higher frequencies — the frequency dependence of the same terms are in opposite sense and their impacts on wave propagation are also opposite — while polarization effect due to bound electrons tends to increase the refractive index, conduction due to free-electrons lowers it, as it can be seen directly from

$$n = \sqrt{1 + \chi + \frac{\sigma}{j\omega\epsilon_o}} = \sqrt{1 + \chi - \frac{Ne^2/m\epsilon_o}{\omega^2}}. \tag{3.13}$$

In any case, at ionospheric heights we can neglect χ, and re-write (3.13) as

$$n = \sqrt{1 - \frac{\omega_p^2}{\omega^2}}, \tag{3.14}$$

where

$$\omega_p \equiv \sqrt{\frac{Ne^2}{m\epsilon_o}} \approx 2\pi \times \sqrt{80.6N} \tag{3.15}$$

is known as *plasma frequency*. In the formula $2\pi \times \sqrt{80.6N}$, electron density N should be entered in MKS units in order to obtain ω_p in radians per seconds. Equivalently, $\sqrt{80.6N}$ gives the plasma frequency $f_p \equiv \frac{\omega_p}{2\pi}$ in Hz units.

Using the refractive index expression above, wavenumber $k = \frac{\omega}{c}n$ and intrinsic impedance $\eta = \eta_o/n$ for a plasma are found to be

$$k = \frac{\omega}{c}\sqrt{1 - \frac{\omega_p^2}{\omega^2}} \quad \text{and} \quad \eta = \frac{\eta_o}{\sqrt{1 - \frac{\omega_p^2}{\omega^2}}},$$

respectively. Both of these wave parameters are purely imaginary if the wave frequency $\omega < \omega_p$. Wave propagation is therefore possible if and only if $\omega > \omega_p$.

Radiowaves with frequencies $\omega < \omega_p$ become *evanescent* or *cut-off* in a plasma, in analogy with the TE and TM mode waves in metallic waveguides for $\omega < \omega_c$. Also, in analogy with waveguides, there exists a distinction between phase and group velocities of propagation — the phase velocity

$$v_p = \frac{\omega}{k} = \frac{c}{n} = \frac{c}{\sqrt{1 - \frac{\omega_p^2}{\omega^2}}} \tag{3.16}$$

is real valued only for $\omega \geq \omega_p$, and exceeds c in that frequency range. Since $v_p > c$, the phase velocity *cannot* be the velocity involved in energy propagation according to *special relativity*. Indeed, it is the *group velocity* to be defined in the next section which is associated with energy/information transfer and is limited by c, again just like in waveguide theory.

What happens in a plasma when the wave frequency $\omega = \omega_p$? At $\omega = \omega_p$, the right hand side of Ampere's law (for ionospheric heights where polarization current is neglected)

$$-j\mathbf{k} \times \mathbf{H} = \sigma\mathbf{E} + j\omega\epsilon_o\mathbf{E}$$

vanishes because of the fact that $\sigma = -j\omega_p^2\epsilon_o/\omega$. Thus we get $\mathbf{k} \times \mathbf{H} = 0$ even though $\mathbf{E} \neq 0$. Since $\mathbf{k} \times \mathbf{H} = 0$ and $\mathbf{k} \cdot \mathbf{H} = 0$ together imply — assuming a non-zero $\mathbf{k}$ — that $\mathbf{H} = 0$, Faraday's law

$$-j\mathbf{k} \times \mathbf{E} = -j\omega\mu_o\mathbf{H}$$

then implies that $\mathbf{k} \times \mathbf{E} = 0$ at $\omega = \omega_p$. But $\mathbf{k} \times \mathbf{E} = 0$ does not necessarily mean $\mathbf{E} = 0$, because *any* $\mathbf{E}$ parallel to $\mathbf{k}$ will also result in $\mathbf{k} \times \mathbf{E} = 0$.

In summary, then, if $\omega = \omega_p$, then $\mathbf{H} = 0$, and any non-zero $\mathbf{E}$ must be in the direction of (non-zero) $\mathbf{k}$, implying, via Gauss' law

$$-j\mathbf{k} \cdot \mathbf{E} = \frac{\rho}{\epsilon_o},$$

the possibility of *longitudinal* space-charge waves[3] with non-zero ρ and $\mathbf{E}$. In a longitudinal wave a non-zero space-charge wave supports a non-zero electric field wave, both time varying at the frequency $\omega = \omega_p$. Thus ω_p is the characteristic frequency of space-charge waves in a plasma which could exist with an arbitrary $\mathbf{k}$. When we relax the cold plasma approximation and take into account the finite temperature T_e of the "electron gas", the nature of longitudinal space-charge waves changes to permit a *range of frequencies*

$$\omega^2 = \omega_p^2 + k^2 3C_e^2$$

where

$$C_e \equiv \sqrt{\frac{KT}{m}}$$

is known as the thermal speed of the electrons. The derivation of this modified *dispersion relation* for space-charge waves in "hot plasmas" requires the inclusion of the pressure

[3]Their quantized chunks with energy $\hbar\omega_p$ are known as *plasmons*.

gradient force in the equation for electron motion. The electron space-charge waves discussed above account for part of the spectrum of plasma density variations δN which scatter radiowaves crossing the ionosphere as noted in Chapter 1 (and discussed in more detail in Chapter 5 on ionospheric radars).

Exercise 1: Show that the dispersion relation of plane TEM waves in a collisionless plasma with the refractive index

$$n = \sqrt{1 - \frac{\omega_p^2}{\omega^2}}$$

can be expressed in a form

$$\omega^2 = \omega_p^2 + k^2 c^2$$

similar to the dispersion relation for space-charge waves.

3.2 Dispersion and group velocity

Consider now wave propagation in a *dispersive medium* in some direction x, where the refractive index n and the phase velocity $v_p = \frac{c}{n}$ are dependent on the wave frequency ω (as in a plasma). In such a medium — where the *dispersion relation* $\omega = \omega(k)$ is non-linear in k as shown in Figure 3.1 — no waveform $E(x,t)$ other than a pure co-sinusoidal such as

$$\frac{Pe^{j(\omega t - kx)} + P^* e^{-j(\omega t - kx)}}{2} = \frac{1}{2} P e^{-jk(x - \frac{\omega}{k} t)} + c.c.$$

can satisfy the *distortionless propagation criterion*

$$E(x,t) = E(x - vt)$$

involving some propagation velocity v. Yet, even in dispersive media, sufficiently *narrowband* waveforms such as

$$E(x,t) \equiv \frac{1}{2} \int_{k_o - \Delta k}^{k_o + \Delta k} \frac{dk}{2\pi} P(k) e^{j(\omega(k)t - kx)} + c.c. \approx \mathrm{Re}\{f(x - v_g t) e^{j(\omega_o t - k_o x)}\} \qquad (3.17)$$

— over a time-scale depending on bandwidth Δk — in which case v_g is said to be the *group velocity* of the medium and the waveform. We associate the information content of such waveforms with a carrier frequency $\omega_o \equiv \omega(k_o)$ with its traveling *envelope* function $f(x - v_g t)$ and it is the group velocity of the envelope which is constrained by relativity to an upper limit of c. Note that the envelope function can be complex valued and describe amplitude as well as phase modulations.

The narrowband waveform (3.17) is clearly a superposition of co-sinusoidal plane-waves with wavenumbers k in the vicinity of some k_o and wave frequencies $\omega(k)$ in the vicinity of $\omega_o = \omega(k_o)$. Taylor expansion of the dispersion relation $\omega(k)$ about $k = k_o$ yields

$$\omega(k) = \omega_o + \omega'(k_o)(k - k_o) + \frac{1}{2}\omega''(k_o)(k - k_o)^2 + \cdots, \qquad (3.18)$$

where the primes indicate derivatives with respect to k. Substituting the first two terms of (3.18) in (3.17) we obtain

$$E(x,t) \approx \frac{1}{2} e^{j(\omega_o t - k_o x)} \int_{k_o-\Delta k}^{k_o+\Delta k} \frac{dk}{2\pi} P(k) e^{-j(k-k_o)(x-\omega'(k_o)t)} + c.c., \tag{3.19}$$

which is of the form

$$E(x,t) \approx \mathrm{Re}\{e^{j(\omega_o t - k_o x)} f(x - v_g t)\}$$

with

$$v_g \equiv \omega'(k_o) = \frac{d\omega}{dk}|_{\omega=\omega_o}, \tag{3.20}$$

and

$$f(x) \equiv \int_{k_o-\Delta k}^{k_o+\Delta k} \frac{dk}{2\pi} P(k) e^{-j(k-k_0)x} = \int_{-\Delta k}^{\Delta k} \frac{dk}{2\pi} F(k) e^{-jkx}, \tag{3.21}$$

clearly a baseband function of position x with a spatial bandwidth Δk. The approximation above is justified if and only if

$$\frac{1}{2}|\omega''(k_o)|(k-k_o)^2 t \leq \frac{1}{2}|\omega''(k_o)|\Delta k^2 t \ll 1,$$

or, equivalently, if

$$t \ll \frac{1}{\frac{1}{2}|\omega''(k_o)|\Delta k^2} \equiv t_{max}.$$

Relatively small Δk (i.e., waveforms with narrow bandwidths) and/or small $|\omega''(k_o)|$ (i.e., weak dispersion, nearly straight line dispersion curve) imply large t_{max}, and, hence, the validity of group velocity concept to describe distortion-free envelope propagation over long propagation times t and distances $\sim v_g t$. Note that the above results also show that the carrier *phase* $\omega_o t - k_o x = -k_o(x - \frac{\omega_o}{k_o} t)$ of $E(x,t)$ travels with the phase velocity

$$v_p = \frac{\omega_o}{k_o}.$$

Example 1: In some propagation medium $\omega^2 = Ak$ is the dispersion relation over some frequency range where A is some constant. Determine the group velocity v_g.

Solution: Differentiating the dispersion relation with respect to k we find

$$2\omega \frac{d\omega}{dk} = A \quad \Rightarrow \quad v_g = \frac{d\omega}{dk} = \frac{A}{2\omega}.$$

Notice that group velocity decreases as frequency ω increases. Also note that $v_g \to \infty$ as $\omega \to 0$, which means that the given dispersion relation cannot be accurate in the low frequency limit. We deduce that from the fact that v_g should not exceed c for any physically valid dispersion relation in view of special relativity constraint.

Example 2: In some medium the phase velocity varies with wavenumber k as $v_p = \frac{k}{Q}$ where Q is some constant. Determine the group velocity v_g and express it in terms of wavelength $\lambda = \frac{2\pi}{k}$.

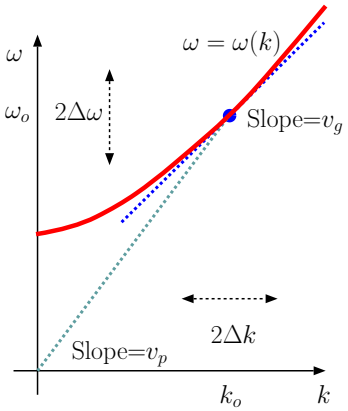


Figure 3.1: A dispersion curve $\omega = \omega(k)$ and a graphical interpretation of v_g and v_p for a signal with carrier parameters ω_o and k_o and bandwidths $\Delta\omega$ and Δk in time and space, respectively, related via $\frac{\Delta\omega}{\Delta k} \approx v_g \equiv \omega'(k_o)$.

Solution: Since the phase velocity $v_p = \frac{k}{Q} = \frac{\omega}{k}$, the dispersion relation is

$$k^2 = Q\omega.$$

Differentiating the dispersion relation with respect to k we find

$$2k = Q\frac{d\omega}{dk} \quad \Rightarrow \quad v_g = \frac{d\omega}{dk} = \frac{2k}{Q} = \frac{4\pi}{\lambda Q}.$$

Note that $v_g = 2v_p$ and it blows up as $\lambda \to 0$. Hence the dispersion relation cannot be valid (because of relativity) for very short λ.

Example 3: In a plasma with the dispersion relation

$$c^2k^2 = \omega^2 - \omega_p^2$$

a message is to be transmitted with a carrier frequency $\omega_o = \sqrt{2}\omega_p$, where $\omega_p = 2\pi f_p$, $f_p = 10$ MHz, and c is the speed of light in free space. If the message bandwidth $\Delta\omega = \frac{\omega_p}{100}$ what is the longest distance x that can be traversed before the message signal begins to distort?

Solution: Differentiating the dispersion relation twice with respect to k we find

$$2c^2k = 2\omega\omega' \text{ and } c^2 = \omega'^2 + \omega\omega''.$$

Therefore,

$$v_g = \omega' = \frac{c^2}{\omega/k} = c\sqrt{1 - \frac{\omega_p^2}{\omega^2}} = c\sqrt{1 - \frac{\omega_p^2}{2\omega_p^2}} = \frac{c}{\sqrt{2}}$$

and

$$\omega'' = \frac{c^2 - v_g^2}{\omega} = \frac{c^2 - c^2/2}{\sqrt{2}\omega_p} = \frac{c^2}{2\sqrt{2}\omega_p}.$$

Also, a bandwidth of $\Delta\omega = \frac{\omega_p}{100}$ implies $\Delta k = \frac{\Delta\omega}{v_g} = \frac{\omega_p}{100 v_g} = \frac{\sqrt{2}\omega_p}{100c}$. Thus,

$$t_{max} = \frac{2}{|\omega''|\Delta k^2} = \frac{2}{\frac{c^2}{2\sqrt{2}\omega_p}\frac{2\omega_p^2}{10^4 c^2}} = \frac{\sqrt{2}22\times 10^4}{\omega_p}$$

and

$$x_{max} = v_g t_{max} = \frac{\sqrt{2}22\times 10^4}{\omega_p}\frac{c}{\sqrt{2}} = \frac{3\times 10^8 \text{m/s}\ 2\times 10^4}{2\pi 10^7\ 1/\text{s}} = 10^5\text{m} = 100\,\text{km}.$$

Thus, for $f_p = 10$ MHz, a narrowband signal with 100 kHz bandwidth and 14 MHz carrier frequency will be distorted before it reaches a distance of about 100 km.

Exercise 1: The dispersion relation of space-charge waves in a plasma is given as

$$\omega^2 = \omega_p^2 + k^2 3C_e^2.$$

where

$$C_e \equiv \sqrt{\frac{KT}{m}}$$

is electron thermal speed. What is the group velocity v_g of space charge waves and what is v_g in $T \to 0$ limit.

Exercise 2: Find the group velocity of *matter waves* governed by the Schrodinger equation

$$j\hbar\frac{\partial\psi}{\partial t} = \frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} - V(x)\psi,$$

for $V(x) = V_o$, where V_o is some constant? Hint: first obtain a dispersion relation $\omega = \omega(k)$ that describes monochromatic matter waves of the form $\psi = e^{j(\omega t - kx)}$.

3.3 Ionospheric dispersion and propagation

We now return to the discussion of plane wave propagation in ionospheric plasmas initiated in Section 1.

Recall that ionospheric refractive index is

$$n = \sqrt{1 - \frac{\omega_p^2}{\omega^2}}$$

neglecting electron collisions and further complications to be considered later. Thus, under the present assumptions, ionospheric phase velocity is

$$v_p = \frac{\omega}{k} = \frac{c}{n} = \frac{c}{\sqrt{1 - \frac{\omega_p^2}{\omega^2}}}$$

and ionospheric dispersion relation can be implicitly written as

$$k = \frac{\omega}{c}n$$

(which is on fact valid in general in *any* isotropic medium with a refractive index n).

Differentiating both sides of the dispersion relation with respect to ω we find

$$\frac{dk}{d\omega} = \frac{1}{c}\frac{d}{d\omega}(\omega n),$$

implying that

$$v_g = \frac{d\omega}{dk} = \frac{1}{dk/d\omega} = \frac{c}{\frac{d}{d\omega}(\omega n)}.$$

It is convenient to define a *group refractive index*

$$n_g \equiv \frac{d}{d\omega}(\omega n)$$

and write next

$$v_g = \frac{c}{n_g} \text{ in analogy with } v_p = \frac{c}{n}.$$

If a medium is non-dispersive n must be independent of ω, in which case

$$n_g \equiv \frac{d}{d\omega}(\omega n) = n\frac{d}{d\omega}(\omega) = n$$

and

$$v_g = \frac{c}{n_g} = \frac{c}{n} = v_p.$$

Otherwise, i.e., if n is a function of ω, then $n_g \neq n$ and $v_g \neq v_p$.

In a plasma, which is a dispersive medium, it turns out that $n_g = \frac{1}{n}$ as shown below:

$$\begin{aligned} n_g &\equiv \frac{d}{d\omega}(\omega n) = \frac{d}{d\omega}(\omega\sqrt{1-\frac{\omega_p^2}{\omega^2}}) \\ &= \frac{d}{d\omega}\sqrt{\omega^2-\omega_p^2} = \frac{\frac{1}{2}}{\sqrt{\omega^2-\omega_p^2}}2\omega = \frac{1}{\sqrt{1-\frac{\omega_p^2}{\omega^2}}} = \frac{1}{n}. \end{aligned}$$

Hence, in a plasma

$$v_g = \frac{c}{n_g} = cn = c\sqrt{1-\frac{\omega_p^2}{\omega^2}} \text{ and } v_p = \frac{c}{n} = \frac{c}{\sqrt{1-\frac{\omega_p^2}{\omega^2}}}.$$

Thus, as $\omega \to \omega_p$ and $v \to \infty$, the group velocity $v_g \to 0$, as shown in Figure 3.2.

Consider now a vertical incidence plane wave $\propto e^{-jk_o z}$, $k_o \equiv \frac{\omega}{c}$, propagating upwards in a homogeneous ionosphere situated above the $z = 0$ plane. The wavenumber k for $z > 0$ will be

$$k = \frac{\omega}{c}n = \frac{\omega}{c}\sqrt{1-\frac{\omega_p^2}{\omega^2}}$$

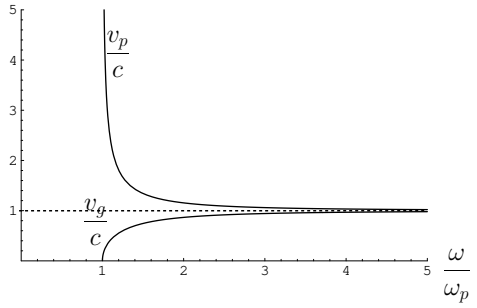


Figure 3.2: Plots of $\frac{v_p}{c}$ and $\frac{v_g}{c}$ vs $\frac{\omega}{\omega_p}$ in a plasma with plasma frequency ω_p. For $\omega < \omega_p$ waves in the medium are evanescent and v_p and v_g do not exist as meaningful parameters. Note that $v_g \leq 1$ for $\omega \geq \omega_p$.

and the field phasor will be

$$\mathbf{E}_o e^{-j\frac{\omega}{c}\sqrt{1-\frac{\omega_p^2}{\omega^2}}z}, \;\; \mathbf{E}_o \cdot \hat{z} = 0.$$

For the case $\omega > \omega_p$ the wavenumber k is real and the field propagates with a phase velocity $v_p > c$ as shown in Figure 3.2. Since k is purely real there is no wave attenuation and the Poynting flux

$$S = \frac{|\mathbf{E}_o|^2}{2\eta} = \frac{|\mathbf{E}_o|^2 n}{2\eta_o} = \frac{|\mathbf{E}_o|^2}{2\eta_o}\sqrt{1-\frac{\omega_p^2}{\omega^2}}$$

is independent of z (since under the assumption of a homogeneous plasma $\omega_p^2 \propto N$ is constant). A narrowband signal constructed as a superposition of waves with frequencies $\omega > \omega_p$ will then propagate with a group velocity $v_g < c$ as shown in Figure3.2.

For the case $\omega < \omega_p$, however, $n^2 < 0$, and hence wavenumber

$$k = \frac{\omega}{c}n = \frac{\omega}{c}\sqrt{1-\frac{\omega_p^2}{\omega^2}} = \pm j\frac{\omega}{c}\alpha, \quad \alpha \equiv \sqrt{\frac{\omega_p^2}{\omega^2}-1} > 0$$

is purely imaginary. This leads to a field phasor

$$\mathbf{E}_o e^{-j(\pm j\frac{\omega}{c}\alpha)z} = \mathbf{E}_o e^{\mp\frac{\omega}{c}\alpha z}$$

which describes a non-propagating but exponentially varying field amplitude. Assuming that the field source is a plane wave arriving from below (i.e., from $z < 0$ region) only a decaying solution

$$\mathbf{E}_o e^{-\frac{\omega}{c}\alpha z}$$

is physical; such a wave field is said to be *evanescent*. The accompanying $\mathbf{H}$ field

$$\mathbf{H} = \frac{\hat{k}\times\mathbf{E}}{\eta} = \frac{n}{\eta_o}\hat{z}\times\mathbf{E} = \frac{-j\alpha}{\eta_o}\hat{z}\times\mathbf{E}_o e^{-\frac{\omega}{c}\alpha z}$$

has also an evanescent character and as a consequence time-average Poynting vector

$$\mathbf{S} = \frac{1}{2}\mathrm{Re}\{\mathbf{E}\times\mathbf{H}^*\} = \frac{1}{2}\mathrm{Re}\{\mathbf{E}_o e^{-\frac{\omega}{c}\alpha z}\times(\frac{j\alpha}{\eta_o}\hat{z}\times\mathbf{E}_o^* e^{-\frac{\omega}{c}\alpha z})\} = 0$$

because $\mathbf{E}\times\mathbf{H}^*$ is purely imaginary.

In summary, plasma medium does not accept energy from an incident plane wave with a frequency $\omega < \omega_p$ and confines the influence of the incident wave within an evanescent field that decays *exponentially* away from the interface. As confirmed in Exercise 1 below this is a "total internal reflection" situation with a unity magnitude reflection coefficient R. When a plane wave of frequency ω encounters a plasma with a plasma frequency $\omega_p > \omega$ it suffers a total internal reflection, turns around, and propagates back towards it source.

Exercise 1: Consider a normal incidence reflection occurring at $z = 0$ surface. The medium below the interface is some dielectric with a non-dispersive refractive index n_i and above the interface is a homogeneous plasma with a refractive index $n_t = \sqrt{1-\frac{\omega_p^2}{\omega^2}}$. Show that for $\omega < \omega_p$, the magnitude of the normal incidence reflection coefficient $R = \frac{n_t - n_i}{n_t + n_i}$ equals 1.

3.4 Group delay, group path, and ionospheric TEC

Consider propagation in z direction in some dispersive medium such that a signal envelope moves up a short distance dz during a short time interval $d\tau$. Since the signal envelope propagates with the group velocity

$$v_g = \frac{c}{n_g}$$

it follows that

$$\frac{dz}{d\tau} = v_g = \frac{c}{n_g}$$

or, equivalently,

$$d\tau = \frac{n_g}{c}dz.$$

If group refractive index n_g is constant, i.e., independent of z, we can write

$$\tau = \frac{n_g}{c}z$$

and interpret the time interval τ as a *group delay*, meaning specifically the propagation time of signal envelope across a distance z. However, if n_g is z dependent, i.e., $n_g = n_g(z)$, then the group delay across a distance z will have to be written as

$$\tau = \int_{z_o}^{z_o+z} \frac{n_g(z')}{c}\,dz'$$

where z_o is the location of the signal source (or an equivalent reference location). We also define a *group path*

$$S_g \equiv c\tau = \int_{z_o}^{z_o+z} n_g(z')\,dz',$$

which is the distance travelled at the speed of light c in free space during a time interval equal to the group delay τ.

In a collisionless plasma

$$n(z) = \sqrt{1 - \frac{f_p^2(z)}{f^2}}$$

and

$$n_g(z) = \frac{1}{n(z)} = \frac{1}{\sqrt{1 - \frac{f_p^2(z)}{f^2}}}.$$

Thus, *assuming* $f \gg f_p(z)$, the group delay in a plasma can be expressed as

$$\tau = \int_{z_o}^{z_o+z} \frac{1}{c\sqrt{1 - \frac{f_p^2(z')}{f^2}}}\, dz' \approx \frac{1}{c}\int_{z_o}^{z_o+z} (1 + \frac{1}{2}\frac{f_p^2(z')}{f^2})\, dz',$$

where we used the fact that $(1+a)^p \approx 1+pa$, which is a valid approximation[4] for $|a| \ll 1$. Hence, at sufficiently high frequencies the group delay within a plasma is given as

$$\tau \approx \frac{1}{c}[z + \frac{1}{2f^2}\int_{z_o}^{z_o+z} dz' f_p^2(z')],$$

or, using $f_p^2 \approx 80.6N$, as

$$\tau \approx \frac{1}{c}[z + \frac{40.3}{f^2}\int_{z_o}^{z_o+z} dz' N(z')].$$

The corresponding group path can be written as

$$S_g = c\tau \approx z + \frac{40.3}{f^2}\mathrm{TEC},$$

where

$$\mathrm{TEC} \equiv \int_{z_o}^{z_o+z} dz' N(z')$$

is the *total electron content* of the physical path between z_o and $z_o + z$.

Note that the group path S_g is obtained by adding to the actual propagation distance z a "correction factor" proportional to the TEC and the *group delay* can be obtained from group path by dividing it by c. Because $v_g \leq c$ is true universally, it follows that the group path $S_g \geq z$ in any type of propagation medium under any circumstance — i.e., group path taken by a signal is never shorter than the physical length of the signal path. In GPS applications (see below) the group path is also referred to as *pseudorange*.

[4] First two terms of Taylor series expansion of $(1 + a)^p$ about $a = 0$.

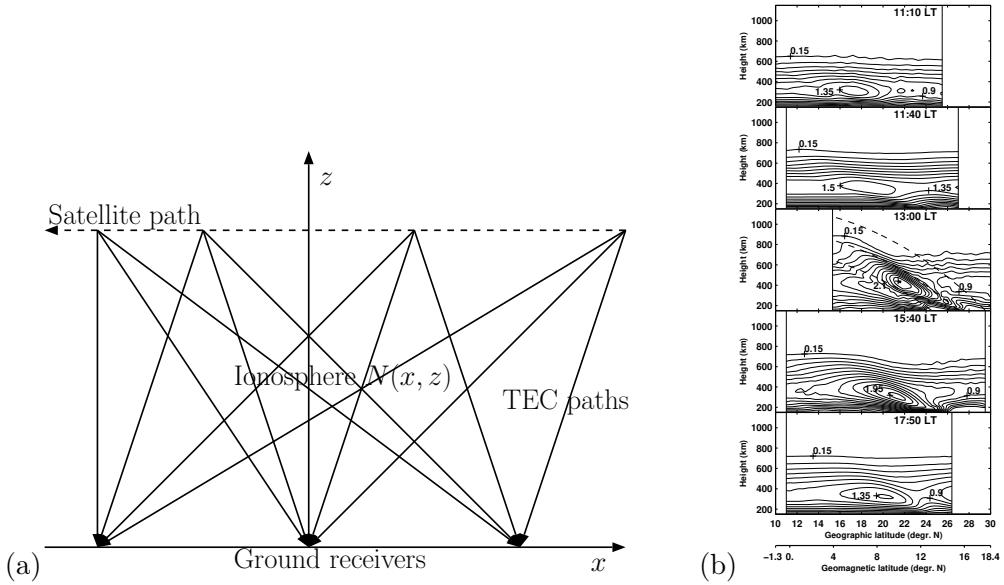


Figure 3.3: (a) Ionospheric tomography geometry and (b) sample $N(x, z)$ maps obtained by tomographic inversion of October 7, 1994 (contours are in units of $10^{12}\,\mathrm{m}^{-3}$) TEC data collected with 6 ground based receivers located between Shanghai and Manila.

Ionospheric TEC is the cause of errors in GPS[5] distance and position measurements when ionospheric effects are neglected. However, with multi-frequency GPS operations the ionospheric error can be eliminated:

Consider GPS pulse transmissions at two carrier frequencies[6] f_1 and f_2 over the same trans-ionospheric path (satellite-to-ground or satellite-to-satellite) of an unknown length z. The corresponding group path expressions

$$S_{g1} \approx z + \frac{40.3}{f_1^2}\mathrm{TEC}$$

and

$$S_{g2} \approx z + \frac{40.3}{f_2^2}\mathrm{TEC}$$

[5]A satellite based "global positioning system" which works based on the *triangulation principle*. GPS receivers receive transmissions from 3 or more GPS satellites and deduce their own locations based on the satellite locations and distances to each of the transmitting satellites. Distances are computed from the propagation time delays. Although 3 satellites are sufficient if the receiver clock is correctly synced with the satellite clock, the effect of small clock errors in the receiver can be eliminated by using signals received from 4 satellites.

[6]The GPS system uses $f_1 = 1575.42 =$ MHz (called L1) and $f_2 = 1227.60$ MHz (called L2) in the L-band. Note that $f_1 = 154 f_o$ and $f_2 = 120 f_o$, where $f_o = 10.23$ MHz is the chips/sec rate for the PRN codes applied to L1 and L2 signals. L1 is open for public use, but an encryption applied to L2 restricts its use to military applications and/or to so-called to codeless techniques.

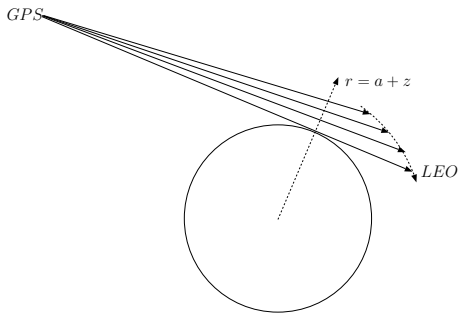


Figure 3.4: Geometry of occultation measurements of electron density profile $N(z)$ using TEC data between a GPS and LEO (low-Earth orbit) satellites. The inversion of TEC data to obtain $N(z)$ requires the assumption of spherical stratification assumption for $N(r)$, $r > a$.

constitute two equations in two unknowns, z and TEC. Solving for TEC, we find

$$\mathrm{TEC} = \frac{S_{g2} - S_{g1}}{\frac{40.3}{f_2^2} - \frac{40.3}{f_1^2}} = c\frac{\tau_2 - \tau_1}{\frac{40.3}{f_2^2} - \frac{40.3}{f_1^2}},$$

where $\tau_2 - \tau_1$ is detection delay of the pulse with carrier f_2 with respect to the other pulse (assuming that the pulses are emitted simultaneously). Once the TEC is known,

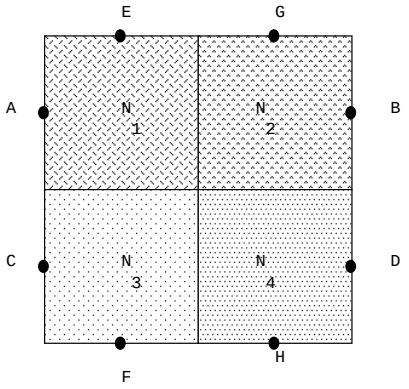
$$z \approx S_{g1} - \frac{40.3}{f_1^2}\mathrm{TEC},$$

for instance, assuming that $S_{g1} = c\tau_1$ can be deduced from the reception time and the content of the message carried by f_1 (which will include the pulse emission time and location). The approximations implied in above equations are extremely good if f_1 and f_2 are much larger that the maximum f_p of the ionosphere.

Also, multiple TEC data collected over multiple and criss-crossing ionospheric paths can be used for "imaging" the ionospheric density distribution $N(x, y, z)$ using "tomographic" inversion techniques as illustrated in Figure 3.3. Tomographic inversion of the GPS TEC data collected on a low-orbit satellites carrying suitable GPS receivers can also be used to determine the ionospheric density profile $N(z)$ using a so-called "occultation" method illustrated in Figure 3.4. Just before the occultation of the receiving satellite behind the globe (with respect to a transmitting GPS satellite), the propagation paths scan the ionosphere as shown in the figure. Assuming spherical stratification of the ionospheric density, $N(z)$ can be obtained by minimizing the misfit between measured TEC data and the expected TEC values based on guesses of $N(z)$. The main shortcoming of this method is that spherical stratification is not always a valid assumption (particularly near sunrise and sunset).

Exercise 1: Consider a collisionless (and contrived;-) plasma distribution depicted in the following diagram where the electron densities N_1, N_2, N_3, and N_4 are unknown and the

dimensions of each square shaped region is 10 km by 10 km.



1. Given that $\text{TEC}_{AB} = 5$, $\text{TEC}_{CD} = 2$, $\text{TEC}_{EF} = 3$, and $\text{TEC}_{EH} = 3\sqrt{1.25}$ in TEC units ($10^{16}\,\text{m}^{-2}$), determine the electron electron densities N_1, N_2, N_3, and N_4.
2. What is TEC_{AD}?
3. What is the group path S_{AB} if the radio frequency $f = 50$ MHz?
4. What is the group path S_{AD} if $f = 10$ MHz?

3.5 Phase delay, differential Doppler, and differential TEC

In a *homogeneous* propagation medium with a constant refractive index n, a TEM wave propagating in z-direction can be represented in terms of a phasor

$$\mathbf{E} = \mathbf{E}_o e^{-j\Phi(z)},$$

where

$$\Phi(z) = \frac{\omega}{c} nz$$

is the *phase delay* function of the wave field. In an *inhomogeneous* medium, where $n = n(z)$, the phase delay of a similar wave field propagating over a distance z can be expressed as

$$\Phi = \frac{\omega}{c} \int_{z_o}^{z_o+z} dz' n(z').$$

If the inhomogeneities in the medium are weak and refraction effects (see Chapter 3) can be neglected, we can then represent the wave field phasor in the medium as

$$\mathbf{E} = \mathbf{E}_o(z) e^{-j\Phi(z)}$$

where $\mathbf{E}_o(z)$ would be a slowly varying field amplitude that we will discuss at a later stage.

For example, in the ionosphere and for $f \gg f_p$, we have

$$n(z) = \sqrt{1 - \frac{f_p^2(z)}{f^2}} \approx 1 - \frac{1}{2}\frac{f_p^2(z)}{f^2} = 1 - \frac{40.3N(z)}{f^2},$$

and hence

$$\Phi = \frac{\omega}{c}\int_0^z dz'(1 - \frac{40.3N(z')}{f^2}) = \frac{\omega}{c}z - \frac{80.6\pi}{cf}\text{TEC}.$$

Since the signal phase Φ is intrinsically 2π-ambiguous, this phase delay formula cannot be used for estimating the TEC's from measured phase delays Φ. However, the time-rate-of-change of Φ — an effective Doppler frequency shift from the carrier frequency f of a wave transmission — can be modeled as

$$\omega_d \equiv \frac{d\Phi}{dt} = \frac{\omega}{c}\frac{dz}{dt} - \frac{80.6\pi}{cf}\frac{d}{dt}\text{TEC} = \frac{2\pi f}{c}\frac{dz}{dt} - \frac{80.6\pi}{cf}\frac{d}{dt}\text{TEC},$$

where $\frac{dz}{dt}$ is the line-of-sight velocity of a moving transmitter with respect to a receiver and $\frac{d}{dt}$TEC is a *differential TEC* which is caused as a result of the change of the integration path. This equation can be inverted for the *differential* TEC. Hence a collection of ω_d data acquired during satellite passes can be used for tomographic estimation of ionospheric electron density $N(x, y, z)$ when augmented by additional data — e.g., an absolute TEC over one of the propagation paths or ionograms.

Another method for obtaining the differential TEC involves using dual carrier frequencies $f_1 = q_1 f$ and $f_2 = q_2 f$, and define a *differential Doppler* parameter by

$$\Delta\omega_d \equiv \frac{\omega_{d2}}{q_2} - \frac{\omega_{d1}}{q_1} = \frac{80.6\pi}{cf}(\frac{1}{q_1^2} - \frac{1}{q_2^2})\frac{d}{dt}\text{TEC}.$$

Using this formula for estimating the differential TEC does not require having the knowledge of the transmitter velocity $\frac{dz}{dt}$. U. S. Navy Navigational Satellites (NNSS) and CICADA satellites of the former Soviet Union carry dual frequency beacons operating at $f_1 = 150$ MHz and $f_2 = 400$ MHz which are used for differential TEC measurements. Same type of beacons are also being installed in new generation research satellites such as the ROCSAT series of the Republic of China (recently renamed as FORMOSAT series).

In closing this section, we define a *phase path* parameter

$$S \equiv \int_{z_o}^{z_o+z} dz' n(z')$$

in analogy with the group path (or pseudorange) S_g. While the ratio $\frac{S_g}{c}$ signifies a group delay, the product $\frac{\omega}{c}S$ signifies a phase delay. The term *optical path* is also used to refer to the phase path S.

Exercise 1: Calculate the phase path from A to B and C to D in Exercise 1 of the previous section.

3.6 WKB and full-wave solutions of propagation in inhomogeneous media

Consider a y-polarized TEM plane wave

$$\mathbf{E}(z,t) = Re\{\hat{y}E(z)e^{j\omega t}\} \tag{3.22}$$

propagating in z-direction in a homogeneous medium with a refractive index n. In that case the wave phasor can be expressed as

$$E(z) = E_o e^{-j\Phi(z)}, \tag{3.23}$$

where

$$\Phi(z) = \frac{\omega}{c} nz \tag{3.24}$$

is a propagation phase delay, and (3.22-3.24) represents an exact solution of Maxwell's equations.

An *approximate* solution of Maxwell's phasor equations describing the propagation of wave (3.22) in an *inhomogeneous* medium with $n = n(z)$ can be expressed as

$$E(z) = E_o(z)e^{-j\Phi(z)} \tag{3.25}$$

in terms of a *phase delay*

$$\Phi(z) = \frac{\omega}{c}\int_0^z n(z')dz' \tag{3.26}$$

and an *amplitude function*

$$E_o(z) = \frac{E_o}{\sqrt{n(z)}}, \tag{3.27}$$

provided that $n(z)$ varies *slowly* with z over height scales of many wavelengths $\lambda = \frac{2\pi}{k}$, where $k = \frac{\omega}{c}n$. The amplitude variation (3.27) follows from *energy conservation* that requires (assuming no absorption or partial reflections)

$$\frac{|E(z)|^2}{2\eta(z)} = \underbrace{\frac{|E_o|^2}{2\eta_o}}_{\text{Poynting flux at } z=0};$$

hence

$$|E(z)|^2 = |E_o|^2\frac{\eta(z)}{\eta_o} = \frac{|E_o|^2}{n(z)} \;\Rightarrow\; E_o(z) = \frac{E_o}{\sqrt{n(z)}}$$

where

$$\eta(z) = \frac{\eta_o}{n(z)}$$

is the intrinsic impedance at height z in terms of local refractive index $n(z)$. We can combine (3.25-3.27) as

$$E(z) = \frac{E_o}{\sqrt{n(z)}}e^{\mp j\frac{\omega}{c}\int_0^z n(z')dz'} \tag{3.28}$$

to represent the upgoing and downgoing waves in the medium with $\mp$ signs, respectively. This approximate wave solution of Maxwell's equations is known as "WKB solution" for historical reasons and will be discussed in detail below.

First, we note that the WKB solution (3.28) gracefully relaxes to the proper form of an ordinary plane TEM wave if the medium is homogeneous and n is independent of z. Otherwise, that is if $n = n(z)$ is inhomogeneous, then the solution predicts that the electric field wave amplitude is larger (smaller) in regions the where refractive index is smaller (larger) and also that the phase varies in proportion to the *phase path* $\int_0^z n(z')dz'$.

The approximate nature of the WKB solution should be evident in that it blows up in the $n(z) \to 0$ limit. This "non-physical" behavior is related to the violation of our assumption that $n(z)$ varies slowly with z in comparison with $\lambda = \frac{2\pi}{k}$, where $k = \frac{\omega}{c}n$, in $n(z) \to 0$ limit where $\lambda \to \infty$. To understand this limitation of the WKB solution better and to find out how it can be improved let's go back to Maxwell's equations with the assumption that

$$\mathbf{E}(z,t) = Re\{\mathbf{E}(z)e^{j\omega t}\}. \tag{3.29}$$

For a medium where $\mu = \mu_o$ and $\epsilon = \epsilon_o$, Maxwell's curl equations

$$\nabla \times \mathbf{E} = -j\omega\mu_o\mathbf{H} \text{ and } \nabla \times \mathbf{H} = \sigma\mathbf{E} + j\omega\epsilon_o\mathbf{E} \equiv j\omega\epsilon_o n^2(z)\mathbf{E}$$

and the Gauss's law can be combined[7] to obtain a vector wave equation

$$\nabla^2\mathbf{E} + \frac{\omega^2}{c^2}n^2(z)\mathbf{E} = 0. \tag{3.30}$$

where

$$n(z) = \sqrt{1 + \frac{\sigma(z)}{j\omega\epsilon_o}} = \sqrt{1 - \frac{\omega_p^2(z)}{\omega^2}}.$$

With $\mathbf{E}(z) = \hat{y}E(z)$ the vector wave equation reduces to the scalar wave equation

$$\frac{d^2E(z)}{dz^2} + \frac{\omega^2}{c^2}n^2(z)E(z) = 0. \tag{3.31}$$

An exact solution of this equation would furnish an exact wave-field solution of Maxwell's equations in the stratified medium. However, in the following sections on *WKB approximation* and the *full-wave* approach, we will see that the exact solution is well matched by the WKB solution (3.28) in regions where $n(z)$ is slowly varying.

[7] Combining the curl equations

$$\nabla \times \mathbf{E} = -j\omega\mu_o\mathbf{H} \text{ and } \nabla \times \mathbf{H} = \sigma\mathbf{E} + j\omega\epsilon_o\mathbf{E} \equiv j\omega\epsilon_o n^2(z)\mathbf{E}$$

we first obtain

$$\nabla \times \nabla \times \mathbf{E} = -j\omega\mu_o\nabla \times \mathbf{H} = -j\omega\mu_o(j\omega\epsilon_o n^2(z)\mathbf{E}) = \frac{\omega^2}{c^2}n^2(z)\mathbf{E}.$$

The vector wave equation (3.30) follows from this and

$$\nabla \times \nabla \times \mathbf{E} = \nabla(\nabla \cdot \mathbf{E}) - \nabla^2\mathbf{E} \text{ as well as } \nabla \cdot \mathbf{E} = 0.$$

3.6.1 WKB approximation

Equation (3.31) has the same form as the time-independent Schrodinger equation from quantum mechanics. Approximate solutions of (3.31) with slowly varying $n^2(z)$ can be obtained by using the WKB method[8] developed to solve the Schrodinger equation. In WKB method it is first assumed that the differential equation (3.31) has a solution of the form

$$E(z) = E_o e^{-j\Phi(z)}, \tag{3.32}$$

where $\Phi(z)$ is an unknown function to be determined. Substitution of (3.32) in (3.31) leads to an ODE

$$(\frac{d\Phi}{dz})^2 + j\frac{d^2\Phi}{dz^2} = \frac{\omega^2}{c^2}n^2(z). \tag{3.33}$$

The second step of the method is to assume that the first term in (3.33) on the left dominates over the second one — because Φ varies almost linearly with z — so that crudely speaking

$$\frac{d\Phi}{dz} \approx \pm\frac{\omega}{c}n(z).$$

Thus

$$\frac{d^2\Phi}{dz^2} \approx \pm\frac{\omega}{c}\frac{dn}{dz},$$

and a less crude formula for $d\Phi/dz$ can be obtained by using this within (3.33) to iterate as

$$(\frac{d\Phi}{dz})^2 \pm j\frac{\omega}{c}\frac{dn}{dz} = \frac{\omega^2}{c^2}n^2(z) \quad \Rightarrow \quad \frac{d\Phi}{dz} \approx \pm\sqrt{\frac{\omega^2}{c^2}n^2(z) \mp j\frac{\omega}{c}\frac{dn}{dz}}$$

which constitutes the third step of the WKB method. Finally simplifying the result as

$$\frac{d\Phi}{dz} \approx \pm\frac{\omega}{c}n(z) - \frac{j}{2}\frac{d}{dz}\ln n = \pm\frac{\omega}{c}n(z) + j\frac{d}{dz}\ln n^{-1/2}$$

(which is well justified by our prior assumptions in step 2), and integrating over z, we find that

$$\Phi(z) \approx \pm\frac{\omega}{c}\int_0^z dz' n(z') + j\ln n^{-1/2}$$

assuming that $\Phi(0) = 0$ and $n(0) = 1$. Therefore, the WKB solution of (3.31) is found to be

$$E(z) = E_o e^{-j\Phi(z)} = E_o e^{\mp j\frac{\omega}{c}\int_0^z dz' n(z')} e^{\ln n^{-1/2}} = \frac{E_o}{\sqrt{n(z)}} e^{\pm j\frac{\omega}{c}\int_0^z dz' n(z')}, \tag{3.34}$$

identical with (3.28) obtained earlier by heuristic arguments.

The importance of the WKB approach outlined above is that its result provides us with a validation (a posteriori justification) of our usage of energy conservation to obtain

[8] ... established by Wentzel, Kramers, and Brillouin.

(3.28). More importantly, we now know that the condition for the validity of (3.28) should be

$$(\frac{d\Phi}{dz})^2 \gg |\frac{d^2\Phi}{dz^2}|,$$

or, equivalently,

$$\frac{\omega^2}{c^2}n^2(z) \gg |\frac{\omega}{c}\frac{dn}{dz}|,$$

invoked in step 2 of the WKB procedure. Clearly, this condition cannot be satisfied near a reflection height $z = z_r$, where, by definition, $n^2(z_r) = 0$. The full-wave solution which will be obtained in the next section will show us the shortcomings of the WKB solution in the vicinity of $z = z_r$.

Note that if an upgoing wave field were reflected from a height $z = z_r$, then the wave field for $z < z_r$ would have to be the sum of the incident and reflected wave fields, which is given, in the WKB approximation, as

$$\begin{aligned} E(z) &= \frac{E_o}{\sqrt{n(z)}}(e^{-j\frac{\omega}{c}\int_0^z dz' n(z')} + e^{j\phi_r}e^{j\frac{\omega}{c}\int_0^z dz' n(z')}) \\ &= \frac{2E_o e^{j\phi_r/2}}{\sqrt{n(z)}}\cos(\frac{\omega}{c}\int_0^z dz' n(z') + \frac{\phi_r}{2}), \end{aligned} \tag{3.35}$$

where $e^{j\phi_r}$ represents, effectively, the reflection coefficient at $z = z_r$ having a unity magnitude (expected by energy conservation arguments) — this model assumes ω to be a positive frequency variable so that $e^{j\phi_r}$ term is indeed the amplitude of the reflected downgoing WKB solution. This total *standing wave field* model shares the shortcomings of the WKB solution for the upgoing wave near $z = z_r$. However, with a proper choice of ϕ_r obtained (see next section) from the full-wave solution, (3.35) works well for $z < z_r$ by a few free-space wavelengths.

3.6.2 Full wave solution

Full wave method is none other than finding an exact solution of Maxwell's equations near the reflection height $z = z_r$, which amounts to finding an exact solution of the scalar wave equation (3.31) for a specific $n^2(z)$ model applicable for z near z_r. We will use the model

$$n^2(z) = \frac{z_r - z}{L}, \tag{3.36}$$

where $L > 0$ is some constant scale length such that $n^2 > 0$ for $z < z_r$ and $n^2 < 0$ for $z > z_r$. This model should always be valid sufficiently close to $z = z_r$, assuming that $n^2(z)$ is differentiable at $z - z_r$. Substituting (3.36) in (3.31), that is, in , namely,

$$\frac{d^2E(z)}{dz^2} + k_o^2 n^2(z)E(z) = 0,$$

where

$$k_o \equiv \frac{\omega}{c}$$

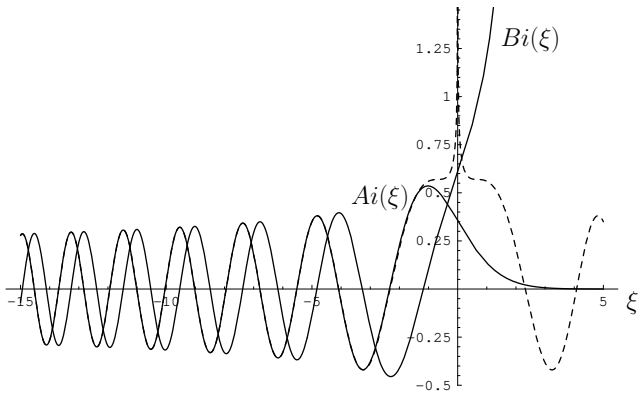


Figure 3.5: Airy functions $Ai(\xi)$ and $Bi(\xi)$ vs ξ. Note that $Bi(\xi)$ is an unbounded function; $Ai(\xi)$, on the other hand, is bounded and $Ai(\xi) \to 0$ as $\xi \to \infty$. The dashed curve is $\frac{|\xi|^{-1/4}}{\sqrt{\pi}} \cos(\frac{2}{3}|\xi|^{3/2} - \frac{\pi}{4})$ which well approximates $Ai(\xi)$ for $\xi < -1$.

is the free-space wavenumber, and defining a non-dimensional height parameter

$$\xi \equiv (\frac{k_o^2}{L})^{1/3}(z - z_r), \tag{3.37}$$

(3.31) can be re-expressed as

$$\frac{d^2E(\xi)}{d\xi^2} = \xi E(\xi). \tag{3.38}$$

The general solution of (3.38), known as *Stoke's equation*, can be written as

$$E(\xi) = K_1 Ai(\xi) + K_2 Bi(\xi), \tag{3.39}$$

a weighted superposition of tabulated Airy functions[9] $Ai(\xi)$ and $Bi(\xi)$ plotted in Figure 3.5.

Plots of $Ai(\xi)$ and $Bi(\xi)$ in Figure 3.5 show that $Bi(\xi)$ is unbounded while $Ai(\xi) \to 0$ as $\xi \to \infty$. Since unbounded solutions cannot be justified for $E(z)$, we take $K_2 = 0$ in (3.39) and obtain

$$E(z) = K_1 Ai((k_o^2/L)^{1/3}(z - z_r)). \tag{3.40}$$

We see from Figure 5 that unlike the WKB approximation the full-wave solution (3.40) is bounded at $z = z_r$, but otherwise it shows the evanescent and oscillatory features of

[9]Note that Stoke's equation requires, for negative ξ, the curvature $\frac{d^2E}{d\xi^2}$ of $E(\xi)$ to have the opposite algebraic sign as $E(\xi)$. This forces$E(\xi)$ to have an oscillatory character for $\xi < 0$ exhibited by $Ai(\xi)$ and $Bi(\xi)$. For $\xi > 0$ the curvature $\frac{d^2E}{d\xi^2}$ and solution E must have the same algebraic signs — this forbids E to have an oscillatory character and forces it either to decay or blow up like $Ai(\xi)$ and $Bi(\xi)$ for $\xi > 0$. Finally, at $\xi = 0$ the solution must go through an inflection point as the functions $Ai(\xi)$ and $Bi(\xi)$ do.

WKB solution above and below z_r. For $\xi < -1$, $Ai(\xi)$ is known to be well approximated by

$$Ai(\xi) \approx \frac{\cos(\frac{2}{3}|\xi|^{3/2} - \frac{\pi}{4})}{\sqrt{\pi}|\xi|^{1/4}}; \tag{3.41}$$

this is confirmed graphically by comparing the plot of (3.41) included in Figure 3.5 (dashed curve) with the full $Ai(\xi)$ curve.

Using (3.41) in (3.40) and matching the result with (3.35), namely

$$E(z) = \frac{2E_o e^{j\phi_r/2}}{\sqrt{n(z)}} \cos(k_o \int_0^z dz' n(z') + \frac{\phi_r}{2})$$

obtained from the superposition of WKB solutions, it can be shown (see Exercise 1 below) that

$$\phi_r = \frac{\pi}{2} - 2k_o \int_0^{z_r} dz' n(z')$$

under the assumption of a positive ω (or else negate it if ω has a negative value). Hence,

$$R \equiv e^{j(\frac{\pi}{2} - 2k_o \int_0^{z_r} dz' n(z'))}$$

can be used as a reflection coefficient at the ground ($z = 0$) to convert an upgoing field amplitude E_o into a reflected field amplitude RE_o from $z = z_r$, the total reflection height.

Later in the chapter we will make use of this result for R in formulating the total reflections of vertically propagating radiowave *pulses* (i.e., modulated carriers) in the ionosphere.

Exercise 1:

1. In MATLAB issue the commands **z=[-25:0.1:5]** and **plot(z,real(airy(z)))** to obtain a plot of Airy function $Ai(\xi)$ for $-25 < \xi < 5$. Compare with **plot(z,abs(z).^(-0.25).*cos(2*abs(z).^(3/2)/3-pi/4))** and also issue a **help airy** command to find out further information about the Airy function.

2. Given that $n^2(z) = \frac{z_r - z}{L}$ and $\xi = (\frac{k_o^2}{L})^{1/3}(z - z_r)$, verify the proportionality (for $\xi < -1$)

$$Ai(\xi) \sim \frac{|\xi|^{-1/4}}{\sqrt{\pi}} \cos(\frac{2}{3}|\xi|^{3/2} - \frac{\pi}{4}) \propto \frac{1}{\sqrt{n(z)}} \cos(\frac{2}{3}n^3(z)Lk_o - \frac{\pi}{4}).$$

3. Show that the WKB solution below the reflection height

$$E(z) \propto \frac{1}{\sqrt{n(z)}} \cos(k_o \int_0^z n(z')dz' + \frac{\phi_r}{2}) = \frac{1}{\sqrt{n(z)}} \cos(\frac{2}{3}(n^3(z) - n^3(0))Lk_o - \frac{\phi_r}{2})$$

after evaluating the phase path integral $\int_0^z n(z')dz'$ with the given $n(z)$.

4. Matching the phases in results (2) and (3) above show that

$$\phi_r = \frac{\pi}{2} - \frac{4}{3}n^3(0)Lk_o = \frac{\pi}{2} - 2k_o \int_0^{z_r} n(z')dz'.$$

3.6.3 Coupled wave equations and partial reflections in plane stratified media

The above treatment of vertical propagation in a plane stratified medium — that is $n = n(z)$ — focused on the *total reflection* of propagating waves from altitudes where $n^2 \to 0$ and the WKB solution breaks down. The possibility of "partial reflections" from regions where the WKB solution works well was not considered.

We will next re-examine WKB solutions of propagation in plane stratified media and obtain a method to estimate the amplitudes of weak partial reflected waves. Our derivation starts by establishing a set of *coupled wave equations* from Maxwell's equations:

Consider the phasors

$$\mathbf{E}(z) = \hat{y}E_y(z) \text{ and } \mathbf{H}(z) = \hat{x}H_x(z)$$

with

$$E_y(z) \equiv E_u(z) + E_d(z) \text{ and } H_x(z) \equiv -\frac{n(z)}{\eta_o}(E_u(z) - E_d(z))$$

expressed in terms for functions $E_u(z)$ and $E_d(z)$ to be determined.

Substituting the phasors within Maxwell's curl equations we find

$$\nabla \times \mathbf{E} = -j\omega\mu_o\mathbf{H} \;\Rightarrow\; \frac{dE_y}{dz} = j\omega\mu_o H_x(z)$$

and

$$\nabla \times \mathbf{H} = j\omega\epsilon_o n^2(z)\mathbf{E} \;\Rightarrow\; \frac{dH_x}{dz} = j\omega\epsilon_o n^2(z)E_y(z).$$

In terms of $E_u(z)$ and $E_d(z)$ the same scalar equations can be written as

$$\begin{aligned} \frac{d}{dz}(E_u + E_d) &= -jk_o n(z)(E_u - E_d) \\ \frac{d}{dz}(E_u - E_d) + \frac{1}{n}\frac{dn}{dz}(E_u - E_d) &= -jk_o n(z)(E_u + E_d) \end{aligned}$$

where $k_o \equiv \omega/c$ is the free-space wavenumber once again. Dividing the sum and difference of these equations by 2 we find that

$$\begin{aligned} \frac{dE_u}{dz} + jk_o nE_u + \frac{1}{2n}\frac{dn}{dz}E_u &= \frac{1}{2n}\frac{dn}{dz}E_d \\ \frac{dE_d}{dz} - jk_o nE_d + \frac{1}{2n}\frac{dn}{dz}E_d &= \frac{1}{2n}\frac{dn}{dz}E_u. \end{aligned} \tag{3.42}$$

Clearly, for $\frac{dn}{dz} = 0$, the solution of the equations just obtained are the up- and down-propagating wave phasors $E_u = e^{-jk_o nz}$ and $E_d = e^{jk_o nz}$, respectively.

We next consider the more interesting case $\frac{dn}{dz} \neq 0$.

In the above set of equations the unknowns E_u and E_d are *coupled* because of the terms on the right. If these coupling terms were *arbitrarily* ignored (i.e., if the right-hand sides were treated as zero), the equations would integrate *exactly* to

$$E_u(z) = \frac{A_u}{\sqrt{n(z)}}e^{-jk_o\int_0^z n(z')dz'} \text{ and } E_d(z) = \frac{A_d}{\sqrt{n(z)}}e^{jk_o\int_0^z n(z')dz'}, \tag{3.43}$$

respectively, which are of course the up- and down-propagating WKB solutions with arbitrary constant amplitudes A_u and A_d. Thus we find out that the neglect of the coupling terms in (3.42) and using the WKB method for (3.31) amount to the same approximation.

Next, in an attempt to solve (3.42) exactly, we substitute (3.43) in (3.42) while treating A_u and A_d as unknown functions of z (instead of constant coefficients) — as a consequence (3.42) simplifies as

$$\begin{aligned} \frac{dA_u}{dz} &= \frac{1}{2n}\frac{dn}{dz}A_d e^{j2k_o\int_0^z n(z')dz'} \\ \frac{dA_d}{dz} &= \frac{1}{2n}\frac{dn}{dz}A_u e^{-j2k_o\int_0^z n(z')dz'}. \end{aligned} \tag{3.44}$$

This set of (exact) equations — for A_u and A_d in (3.43) solutions of the original propagation problem — can be solved by using the following iterative method: (a) Assume first a constant A_u and integrate the second equation to obtain an approximate solution for $A_d(z)$, (b) substitute the approximate $A_d(z)$ in the first equation and integrate it to obtain an approximation for $A_u(z)$, (c) integrate next the second equation using $A_u(z)$ just obtained to obtain an improved $A_d(z)$ solution, and so and so forth. Frequently, we may be content with the WKB solution for the up-going $E_u(z)$ and want to find out the first order approximation for partially reflected $E_d(z)$ — in that case we just implement step (a) above and then stop (as illustrated next).

3.6.4 Ionospheric partial reflections

We will next integrate the second equation in (3.44) assuming a constant A_u and

$$n^2(z) = 1 - \frac{\omega_p^2(z)}{\omega^2} = 1 - \frac{e^2}{m\epsilon_o\omega^2}N(z)$$

in order to calculate the partial reflections of an up-propagating wave through an ionosphere with an electron density profile $N(z)$. We will also assume a sufficiently high radiowave frequency ω such that $n \approx 1$ can be used everywhere in (3.44) except in the computation of $\frac{dn}{dz}$. Thus, we find

$$2n\frac{dn}{dz} = -\frac{e^2}{m\epsilon_o\omega^2}\frac{dN}{dz} \;\Rightarrow\; \frac{dn}{dz} \approx -\frac{e^2}{2m\epsilon_o\omega^2}\frac{dN}{dz} = -\frac{e^2}{2mc^2\epsilon_o k_o^2}\frac{dN}{dz}$$

and, consequently, the second equation in (3.44) is reduced to

$$\frac{dA_d}{dz} \approx -\frac{e^2}{4mc^2\epsilon_o k_o^2}\frac{dN}{dz}A_u e^{-j2k_o z}. \tag{3.45}$$

Integrating (3.45) by parts from 0 to ∞ and using the boundary conditions $A_d(\infty) = N(\infty) = N(0) = 0$ we obtain

$$R_p \equiv \frac{A_d(0)}{A_u} \approx j\frac{e^2}{2mc^2\epsilon_o k_o}\int_0^\infty N(z)e^{-j2k_o z}dz$$

or, equivalently,

$$R_p \approx j\lambda_o r_e \int_0^\infty N(z) e^{-j2k_o z} dz,$$

where $\lambda_o = \frac{2\pi}{k_o}$ is free-space wavelength and

$$r_e \equiv \frac{e^2}{4\pi\epsilon_o mc^2} \approx 2.8 \times 10^{-15}\,\mathrm{m}$$

is a fundamental parameter known as *classical electron radius*.

Clearly, the above result indicates that *partial reflection coefficient* R_p of the entire ionosphere above the ground is proportional to the spatial Fourier transform

$$\tilde{N}(k) \equiv \int_{-\infty}^\infty N(z) e^{jkz} dz$$

of the ionospheric electron density $N(z)$ evaluated at $k = -2k_o$. Notice that in this interpretation function $N(z)$ is required to satisfy $N(z) = 0$ for $z \leq 0$ as consistent with the ionospheric electron density.

In practice, ionospheric density $N(z)$ is composed of a smoothly varying and (in principle) predictable component $N_b(z) \equiv \langle N(z) \rangle$ and a small-amplitude unpredictable part $\delta N(z) \equiv N(z) - N_b(z)$ with a zero-mean. Since the characteristic scale of $N_b(z)$ is on the order of 100's of km, the contribution of $N_b(z)$ to $\tilde{N}(k)$ at $k = -2k_o$ is negligibly small at large ω and $k_o = \frac{\omega}{c}$ compatible with our working assumption that $n \approx 1$. For that reason $N_b(z)$ can be neglected altogether in the expression for R_p to obtain

$$R_p \approx j\lambda_o r_e \int_0^\infty \delta N(z) e^{-j2k_o z} dz.$$

Even though typically $|\delta N(z)| \ll N_b(z)$ in the ionosphere, R_p above can attain substantial (meaning detectable by practical means) magnitudes when $\delta N(z)$ contains a component that varies as $e^{j2k_o z} + c.c.$ Partial reflections caused by density variations of a wavenumber $2k_o$ or wavelength $\frac{\lambda_o}{2}$ add up constructively (as opposed to destructively) to dominate the overall partial reflection process. Wavelength $\frac{\lambda_o}{2}$ important in this process is known as *Bragg wavelength* in honor of the work of William H. Bragg on scattering of X-rays from crystal lattices (where the scattering condition is the lattice spacing being an integral multiple of $\frac{\lambda_o}{2}$). More on Bragg scatter process in Chapter 5 where we study ionospheric backscatter radars. Also see Appendix VII for a brief discussion of partial reflections using pulsed (rather than monochromatic) transmissions.

3.7 Vertical incidence ionospheric sounding

Consider next *pulsed* radiowaves of the form

$$f(t)\cos(\omega_o t) = \frac{1}{2}\int \frac{d\omega}{2\pi} F(\omega - \omega_o) e^{j\omega t} + c.c. \quad \leftrightarrow \quad \frac{1}{2}F(\omega - \omega_o) + \frac{1}{2}F(\omega + \omega_o) \qquad (3.46)$$

which are transmitted up from $z = 0$ into the ionosphere where the pulse envelope $f(t) \leftrightarrow F(\omega)$ is real valued and "narrowband" with respect to carrier frequency ω_o — see HW to understand why the inverse Fourier transform can be written in this form using *c.c.* notation.

We will assume here a sufficiently small carrier frequency ω_o so that vertically propagating monochromatic waves of frequency ω_o go through a *total reflection* at a height $z_r(\omega_o)$ where the plasma frequency equals ω_o. In that case — as we will show next — a sufficiently narrowband pulsed transmission $f(t)\cos(\omega_o t)$ will also "reflect from" $z = z_r(\omega_o)$ and return back to $z = 0$ after having travelled a total distance of $2z_r(\omega_o)$ at the ionospheric group speed $c/n_g(\omega_o)$.

We model the total reflected pulse as

$$p(t) = \frac{1}{2}\int \frac{d\omega}{2\pi} F(\omega - \omega_o) e^{j\phi_r(\omega)} e^{j\omega t} + c.c., \tag{3.47}$$

where

$$\phi_r(\omega) = \frac{\pi}{2} - 2\frac{\omega}{c}\int_0^{z_r(\omega)} dz' n(z') \tag{3.48}$$

is the *total reflection phase* $e^{j\phi_r(\omega)}$ derived earlier in the chapter and the *total reflection height* $z_r(\omega)$ satisfies the constraint $n^2(z_r(\omega)) = 0$ for all ω over the $2\Delta\omega$ bandwidth specified by $F(\omega - \omega_o)$. Note that the reflection coefficient $e^{j\phi_r(\omega)}$ can be interpreted in terms of a two-way propagation phase delay in the exponent (in the WKB sense) and an additional 90° phase advance at the reflection point (revealed by full-wave calculation).

After Taylor series expanding $\phi_r(\omega)$ about $\omega = \omega_o$ as $\phi_r(\omega) \approx \phi_r(\omega_o) + \phi_r'(\omega_o)(\omega - \omega_o)$, (3.47) can be simplified and approximated as

$$\begin{aligned} p(t) &= \mathrm{Re}\{\int_{\omega=0}^{\infty} \frac{d\omega}{2\pi} F(\omega - \omega_o) e^{j\phi_r(\omega)} e^{j\omega t}\} \\ &\approx \mathrm{Re}\{e^{j\phi_r(\omega_o)} e^{j\omega_o t} \int \frac{d\omega}{2\pi} F(\omega - \omega_o) e^{j\phi_r'(\omega_o)(\omega-\omega_o)} e^{j(\omega-\omega_o)t}\} \\ &= -\sin(\omega_o t - 2k_o S(\omega_o))\, f(t - \frac{2S_g(\omega_o)}{c}), \end{aligned}$$

where

$$S(\omega_o) \equiv \int_0^{z_r(\omega_o)} n(z) dz$$

is the *phase path* to the reflection height $z_r(\omega_o)$ at frequency ω_o, whereas

$$S_g(\omega_o) \equiv \int_0^{z_r(\omega_o)} n_g(z) dz = \int_0^{z_r(\omega_o)} \frac{1}{n(z)} dz$$

is the corresponding *group height*. In reaching this result for $p(t)$ we use, in Taylor expansion of $\phi_r(\omega)$,

$$\begin{aligned} \frac{\partial}{\partial\omega}\int_0^{z_r(\omega)} \omega n(z) dz &= \frac{\partial z_r(\omega)}{\partial\omega} \omega n(z_r(\omega)) + \int_0^{z_r(\omega)} \frac{\partial}{\partial\omega}(\omega n(z)) dz \\ &= 0 + \int_0^{z_r(\omega)} n_g(z) dz = S_g(\omega), \end{aligned}$$

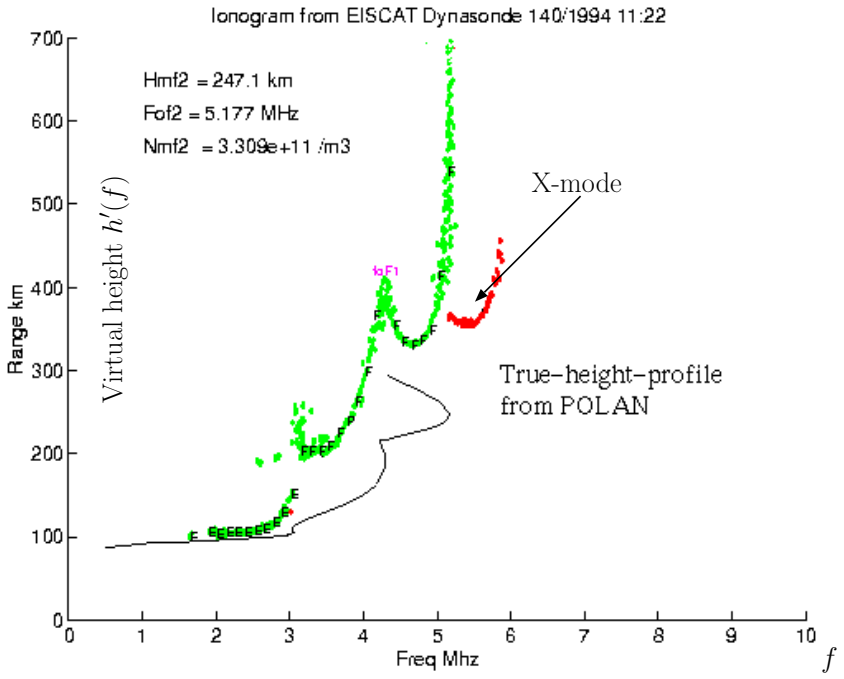


Figure 3.6: A sample ionogram measured in Norway — see http://www.eiscat.rl.ac.uk/dynasonde/true_height_analysis.html. The solid curve is the ionospheric plasma density profile $f_p(z)$ inferred from the ionogram represented by the traces identified with letters EE FF; the trace marked as "X-mode" should be ignored until Chapter 6. This ionogram indicates a critical frequency of $f_c = Fof2 = 5.177$ MHz for the ionosphere during the measurement period.

since $n(z_r(\omega)) = 0$ as already pointed out — here we used *Leibniz rule*[10] of differentiation since the derivative of an integral with variable integration limits was sought.

The result

$$p(t) = -f(t - \frac{2S_g(\omega_o)}{c}) \sin(\omega_o t - 2k_o S(\omega_o))$$

for the reflected pulse indicates a total reflection delay

$$\tau(\omega_o) = \frac{2S_g(\omega_o)}{c}$$

of the transmitted envelope $f(t)$, where $S_g(\omega_o)$ is the group path of the pulse to the reflection height $z_r(\omega_o)$ at the carrier frequency. Delay $\tau(\omega_o)$ can be easily measured

[10] $\frac{\partial}{\partial x}\int_a^b f(x,z)dz = \int_a^b \frac{\partial f}{\partial x}dz + f(x,b)\frac{\partial b}{\partial x} - f(x,a)\frac{\partial a}{\partial x}$ when integration limits a and b have x dependence.

experimentally and inverted to group path $S_g(\omega_o)$ using

$$\frac{c\tau}{2} = S_g = \int_0^{z_r} n_g(z)dz = \int_0^{z_r} \frac{dz}{n(z)} = \int_0^{z_r} \frac{dz}{\sqrt{1-\frac{f_p^2(z)}{f^2}}} \equiv h'(f)$$

at any desired carrier frequency f. The function $h'(f)$ defined above is known as ionospheric *virtual reflection height.* — the prime after height h is the usual notation to remind the reader that the height is not the real reflection height.

Since the virtual height

$$h'(f) = \int_0^{z_r} \frac{dz}{\sqrt{1-\frac{f_p^2(z)}{f^2}}}$$

is a group path, it follows that $h'(f) \geq z_r$, in which the true reflection height z_r satisfies the equation

$$f_p(z_r) = f$$

for a given $f_p(z)$ profile and a sounding frequency f.

Since the total group delay of a reflected pulse (from ground to z_r and back) satisfies

$$c\tau = 2h'(f)$$

a plot of $h'(f)$ vs f can be obtained experimentally by measuring the round trip group delays τ of a sequence of pulses transmitted into the ionosphere using a sequence of carrier frequencies f. Such plots of $h'(f)$ vs f are known as *ionograms* and simple RF systems used to implement ionogram measurements are known as *ionosondes.* An ionogram example is shown in Figure 3.6 which will be discussed after a sequence of exercise problems.

Example 1: Linear plasma density profile In some ionosphere the plasma density $N(z)$ increases linearly with z above some height z_o and consequently

$$f_p^2(z) = \begin{cases} \alpha(z-z_o), & z > z_o \\ 0, & z < z_o, \end{cases}$$

where α is some positive constant. Determine the virtual height function $h'(f_v)$ and plot it as a function of f assuming that $z_o = 50$ km and $\alpha = 0.1\,\text{km}^{-1}$.

Solution: The reflection condition $f_p(z_r) = f$ implies that

$$z_r = z_o + \frac{f^2}{\alpha}.$$

Therefore,

$$h'(f) = \int_0^{z_r} dz \frac{1}{\sqrt{1-f_p^2(z)/f^2}} = z_o + \int_{z_o}^{z_o+f^2/\alpha} dz \frac{1}{\sqrt{1-\alpha(z-z_o)/f^2}}.$$

Performing the integral, we find

$$h'(f) = z_o + \frac{2f^2}{\alpha}.$$

Figure 3.7 shows plots of $h'(f)$ and z_r (dashed curve) corresponding to $z_o = 50$ km and $\alpha = 0.1\,\text{km}^{-1}$. Notice that $h'(f) > z_r$ for all $f > 0$.

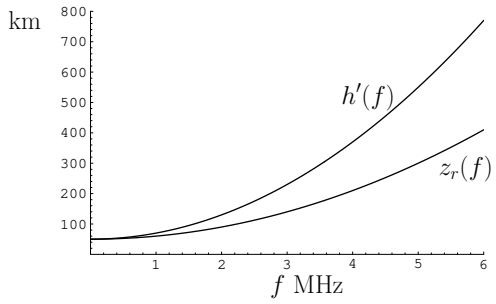


Figure 3.7: Ionogram $h'(f)$ and reflection height $z_r(f)$ for a hypothetical ionosphere described in Example 1.

Example 2: Parabolic plasma density profile Assume that

$$f_p^2(z) = \begin{cases} f_c^2\{1-(\frac{z-z_m}{s})^2\}, & z_m - s < z < z_m + s \\ 0, & \text{otherwise;} \end{cases}$$

where z_m, $s < z_m$ and f_c are positive constants corresponding to the altitude, semi-thickness, and maximum plasma frequency of an ionosphere with a parabolic shaped density profile $N(z)$. Determine $h'(f)$ and plot it for $f_c = 10$ MHz, $z_m = 200$ km, and $s = 50$ km.

Solution: Reflection condition $f_p(z_r) = f_v$ implies that

$$f_c^2\{1-(\frac{z_r - z_m}{s})^2\} = f^2 \;\Rightarrow\; \frac{z_r - z_m}{s} = \pm\sqrt{1 - f^2/f_c^2} \;\Rightarrow\; z_r = z_m - s\sqrt{1 - f^2/f_c^2},$$

where $-$ root is chosen because reflections will occur below the peak height z_m. Hence,

$$h'(f) = (z_m - s) + \int_{z_m - s}^{z_m - s\sqrt{1-f^2/f_c^2}} dz \frac{1}{\sqrt{1 - f_c^2\{1-(\frac{z-z_m}{s})^2\}/f^2}}.$$

Performing the integral it can be shown that

$$h'(f) = (z_m - s) + \frac{1}{2} s \frac{f}{f_c} \ln\left\{\frac{f_c + f}{f_c - f}\right\}.$$

Note that for $f \to f_c$, $h'(f) \to \infty$. Figure 3.8a shows plots of $h(f)$ and z_r for $f_c = 10$ MHz, $z_m = 200$ km, and $s = 50$ km. Note that $h'(f)$ is effectively ∞ for $f > 10$ MHz since pulses with carriers larger than 10 MHz are not reflected and therefore never return back to the ground. This portion of the ionogram $h'(f)$ is naturally not plotted.

Exercise 1: A piecewise continuous ionosphere Consider an ionosphere where

$$f_p(z) = \begin{cases} 0, & z < 100 \text{ km} \\ 4, & z > 100 \text{ km} \end{cases}$$

in MHz units. Sketch the ionogram curve $h'(f)$ and explain your reasoning.

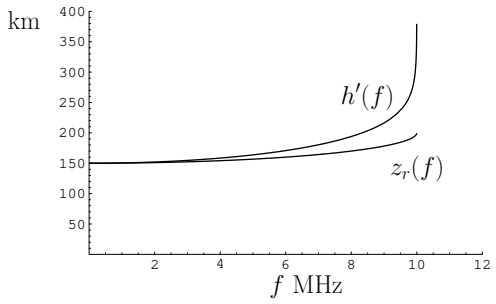


Figure 3.8: Ionogram $h'(f)$ and reflection height $z_r(f)$ for a parabolic ionosphere described in Example 2. The maximum plasma frequency of 10 MHz inferred for this ionosphere would be referred to as critical frequency f_c of the ionosphere.

Exercise 2: Another piecewise continuous ionosphere Consider an ionosphere where

$$f_p(z) = \begin{cases} 0, & z < 100 \\ 4, & 100 < z < 200 \\ 0, & 200 < z < 300 \\ 6, & z > 300 \end{cases}$$

where MHz and km units are used to refer to f_p and z. Sketch the ionogram curve $h'(f)$ and explain your reasoning.

An ionogram measured in a "single-humped ionosphere" will resemble the $h'(f)$ curve obtained in Example 2. Since the real ionosphere is usually multi-humped in its $N(z)$ and $f_p^2(z)$ distributions, real ionograms often have shapes similar to that shown in Figure 3.6. In general it is not possible to infer from an ionogram $h'(f)$ the underlying $f_p(z)$ profile uniquely because of inaccessible "valley regions" of the $f_p(z)$ variation. Although many inversion techniques have been developed over the years to obtain reasonably good estimates of $f_p(z)$ from $h'(f)$ (the solid curve in Figure 3.6 being the result of one such inversion effort), state-of-the-art ionospheric research does not depend on ionosondes as primary measurement instruments. However, they play an important support role and also provide *sufficient information* for making ionospheric sky-wave link calculations as discussed in the next chapter.

Note that the maximum plasma frequency of the ionosphere can be deduced with no ambiguity from a vertical incidence ionogram — this parameter is usually labelled as f_c and called the *critical* frequency.

3.8 Ionospheric absorption

As radiowaves traverse the ionosphere they suffer some amount of attenuation as a consequence of Ohmic losses due to electron collisions mainly with neutral atoms and molecules. In this section we will examine this loss phenomenon.

Recall that in a collisionless plasma the current response $\mathbf{J} = -eN\mathbf{v}$ of electrons to an electric field $\mathbf{E}$ is described by the differential equation

$$\frac{d\mathbf{J}}{dt} = \frac{Ne^2}{m}\mathbf{E}.$$

In a partially ionized collisional plasma every electron will suffer a collision "every-τ-often" with some neutral particle — τ is the mean free-time between collisions[11]. During the collision process the electron loses its velocity $\mathbf{v}$ and gains a new velocity $\mathbf{v}'$. But $\mathbf{v}'$ is random and zero-mean (when averaged over many collisions of a single electron, or over a single collision of many electrons). Therefore, as a result of collisions $N\frac{t}{\tau}$ electrons per unit volume lose their velocity $\mathbf{v}$ within an infinitesimal time span t, and, on the average, gain nothing from the same process. The upshot is, in the absence of $\mathbf{E}$ any initial $\mathbf{J}$ drops to $\mathbf{J} - \mathbf{J}\frac{t}{\tau} = \mathbf{J}(1 - \frac{t}{\tau}) = \mathbf{J}e^{-\frac{t}{\tau}}$ after a short t, with the implication that given an external forcing $\mathbf{E}$ and a collision frequency $\nu \equiv \frac{1}{\tau}$,

$$\frac{d\mathbf{J}}{dt} + \nu\mathbf{J} = \frac{Ne^2}{m}\mathbf{E}$$

within a collisional plasma. The phasor form of this damped first order equation,

$$j\omega\mathbf{J} + \nu\mathbf{J} = \frac{Ne^2}{m}\mathbf{E},$$

implies that in sinusoidal steady-state

$$\sigma = \frac{\mathbf{J}}{\mathbf{E}} = \frac{Ne^2/m}{(j\omega + \nu)}. \tag{3.49}$$

Notice that collisional conductivity (3.49) is no longer purely imaginary and it has an Ohmic positive-real component $\frac{Ne^2}{m}\frac{\nu}{\nu^2+\omega^2}$; its inductive component $-j\frac{Ne^2}{m}\frac{\omega}{\nu^2+\omega^2}$ reduces to collisionless conductivity for $\nu \ll \omega$. To see how these changes in σ affect ionospheric radiowaves we need to find out how the collision frequency ν varies throughout the ionosphere.

3.8.1 Ionospheric collision frequency

An ionospheric free electron that has just been deflected from a collision will fly some τ s with its new velocity, say, $\mathbf{v}$, before making another collision with another molecule with some cross-sectional area $a \equiv \pi r_a^2$. The volume swept by the molecule over interval τ due to its relative velocity $-\mathbf{v}$ towards the electron is clearly $a|\mathbf{v}|\tau$. We expect the swept volume at the instant of collision to be on the order of volume per neutral molecule, i.e., $\frac{1}{n}$, where n is the number density of the neutrals (i.e., electron moves through one

[11] Electrons will of course also collide with one another and also with positive ions, but these charged particle collisions are less frequent than electron-neutral collisions in a partially ionized plasma such as the ionosphere. Their effects can be neglected in the ionosphere up till ~ 400 km altitude where their frequency of occurrence catches up with electron-neutral collision frequencies.

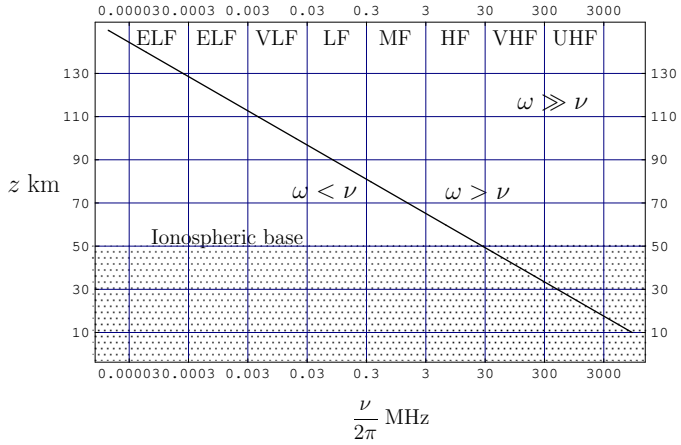


Figure 3.9: A height profile plot of $\frac{\nu}{2\pi}$ (MHz). The shaded region below 50 km contains a negligible number of free electrons, i.e., $f_p \approx 0$, and thus can be regarded as free space. For radiowave frequencies higher than 30 MHz (VHF and above) $\omega > \nu$ throughout the entire ionosphere (above the shaded region) and consequently collisions have a negligible effect (see text). $\frac{\nu}{2\pi}$curve was obtained from (3.50) for an isothermal atmosphere in hydrodynamic equilibrium with $p_o = 95000$ Pascales and $T = 235$ K (see Appendix I). At $T = 235$ km atmospheric scale height $H = \frac{KT}{Mg} \approx 6.9$ km, which is the e-fold length of ν decay with height.

volume per neutral before banging into one of them!). Thus, on the average, $a|\mathbf{v}|\tau = \frac{1}{n}$, or, equivalently, $\tau^{-1} = na|\mathbf{v}|$, the expected value of which yields the collision frequency

$$\nu \equiv na\langle|\mathbf{v}|\rangle = n\pi r_a^2\langle|\mathbf{v}|\rangle.$$

Now, since $\langle|\mathbf{v}|\rangle = \sqrt{2}C_e = \sqrt{2KT/m}$ for electrons in thermal equilibrium with neutrals[12] at a temperature T, it follows that[13]

$$\nu = n\pi r_a^2\sqrt{\frac{2KT}{m}} = 5.4 \times 10^{-16} n\sqrt{T}\,\mathrm{s}^{-1}, \tag{3.50}$$

with $r_a = 1.77 \times 10^{-10}$ m, or 1.77 angstroms, typical of an average air molecule. In the above formula for ν, the neutral density n is specified in units of m^{-3}.

Figure 3.9 shows the collision frequency profile of the ionosphere based on the model (3.50). To generate the profile we assumed an isothermal ionosphere with $T = 235$ K and

[12] $\sqrt{2KT/m}$ is specifically the *most probable electron speed* of an electron gas with a thermal speed $C_e = \sqrt{KT/m}$. Thermal speed C_e is the rms value of v_x, v_y, v_z components of electron thermal velocity $\mathbf{v} = (v_x, v_y, v_z)$, whereas most probable speed $\sqrt{2KT/m}$ corresponds to the peak of a 1-D pdf derived for the scalar quantity $|\mathbf{v}|$. Also note that $\frac{1}{2}mC_e^2 = \frac{1}{2}KT$ is average kinetic energy of electron motions for each degree of freedom.

[13] See Chapter 4 in "Introduction to ionospheric physics" by Rishbeth and Garriot. For further details of collision processes see "Ionospheres" by Schunk and Nagy.

used $p(z) = p_o e^{-\frac{z}{H}}$ from Appendix II with $p_o = 95000$ Pascales and $H = \frac{KT}{Mg}$. Notice that the horizontal axis is logarithmic and collision frequency values are indicated as $\frac{\nu}{2\pi}$ in MHz units. Since the ionosphere does not exist for $z < 50$ km, collision frequencies in the shaded portion of the diagram are inconsequential. The diagram shows that within the ionosphere $\nu < \omega$ for radiowaves in VHF or higher bands — hence, VHF and above, collisions have a negligible effect on radiowaves. For HF and MF, on the other hand at least in some portion of the ionosphere $\nu > \omega$; particularly at MF — important for AM communications and some ionospheric research radars — the collisional slab of the ionosphere where $\nu > \omega$ is substantially thick. We therefore expect important collision effects as $f = \frac{\omega}{2\pi}$ descends from VHF towards MF.

3.8.2 Ionospheric absorption in MF and HF bands

Substituting (3.49) in the refractive index formula

$$n = \sqrt{1 + \frac{\sigma}{j\omega\epsilon_o}} \equiv n' - jn'',$$

we obtain

$$\begin{aligned} n &= \sqrt{1 + \frac{1}{j\omega\epsilon_o}\frac{Ne^2/m}{(j\omega + \nu)}}. = \sqrt{1 - \frac{\omega_p^2}{\omega^2(1 - j\frac{\nu}{\omega})}} = \sqrt{1 - \frac{\omega_p^2}{\omega^2}\frac{1 + j\frac{\nu}{\omega}}{1 + (\frac{\nu}{\omega})^2}} \qquad (3.51) \\ &= \sqrt{1 - \frac{f_p^2}{f^2}\frac{1 + j\frac{\nu}{2\pi f}}{1 + (\frac{\nu}{2\pi f})^2}}. \qquad (3.52) \end{aligned}$$

Thus, for $f \gg f_p$,

$$n \approx 1 - \frac{1}{2}\frac{f_p^2}{f^2}\frac{1 + j\frac{\nu}{2\pi f}}{1 + (\frac{\nu}{2\pi f})^2} \quad \text{and} \quad n'' \approx \frac{1}{2}\frac{f_p^2}{f^2}\frac{\frac{\nu}{2\pi f}}{1 + (\frac{\nu}{2\pi f})^2},$$

and, consequently,

$$k' \equiv k_o n' = \frac{2\pi f n'}{c} = \frac{\pi f}{c}\frac{f_p^2}{f^2}\frac{\frac{\nu}{2\pi f}}{1 + (\frac{\nu}{2\pi f})^2}.$$

Hence the amplitude of a field

$$\mathbf{E}_o e^{-jkr} = \mathbf{E}_o e^{-j(k' - jk'')r} = \mathbf{E}_o e^{-k''r} e^{-jk'r},$$

attenuates at a rate of

$$\begin{aligned} A(f) = 20\log_{10}\{e^{k''(1\,\mathrm{km})}\} &= \frac{20000\pi}{c}\frac{f_p^2}{f}\frac{\frac{\nu}{2\pi f}}{1 + (\frac{\nu}{2\pi f})^2}\log_{10} e \\ &= 0.00737\frac{N}{f}\frac{\frac{\nu}{2\pi f}}{1 + (\frac{\nu}{2\pi f})^2}\,\frac{\mathrm{dB}}{\mathrm{km}}. \end{aligned}$$

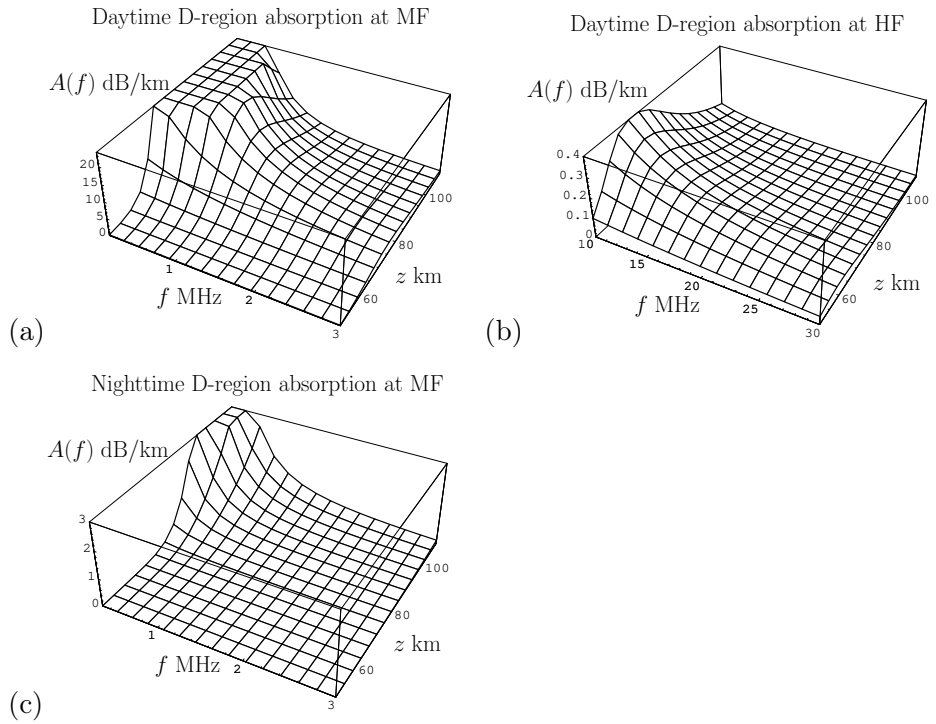


Figure 3.10: Attenuation rate $A(f)$ as a function of radio frequency f and altitude z. Plots (a) and (b) depict a daytime ionosphere specified as $N(z) = 10^{11}e^{(z-100)/10}\,\mathrm{m}^{-3}$. Plot (c) illustrates attenuation in the MF band in a nighttime ionosphere specified as $N(z) = 10^{9}e^{(z-100)/5}\,\mathrm{m}^{-3}$. As a result of reduced ionization at 1 MHz AM signals can survive the D-region absorption at night. AM reception of far-away stations at night is a consequence of the effective loss of D-region after sunset.

Figures 3.10a-c show surface plots of $A(f)$ as function of ionospheric height z and radio frequency f. The surfaces were calculated using the collision frequencies of Figure 3.9. Figures 3.10a and b, illustrating daytime MF and HF band absorption rates, respectively, were calculated using a plasma density profile $N(z) = 10^{11}e^{(z-100)/10}\,\mathrm{m}^{-3}$ representative of the daytime D-region (50-to 90 km). Clearly, 1 MHz signals in the MF band, corresponding to AM transmissions, are very heavily absorbed (see Figure 3.10a). Daytime absorption is non-negligible even at MF for $f \approx 10$ MHz (see Figure 3.10b), but certainly becomes tolerable at higher HF as 30 MHz is approached. This tendency is the reason behind for using transmission frequencies as close to MUF (maximum usable frequency, see Chapter 3) as possible in shortwave communications. Figure 3.10b showing the nighttime absorption rate MF ware calculated using a plasma density profile $N(z) = 10^{9}e^{(z-100)/5}\,\mathrm{m}^{-3}$; the figure clearly show that during the night 1 MHz AM signals can survive D-region absorption because of the effective loss of D-region after sunset. Note that $A(f) \propto \frac{N}{f}$ and thus $A(f)$ can be sufficiently small even for low values

of f such as in MF band when plasma density N is sufficiently small as during in the nighttime D-region.

A few additional comments: First, the absorption rates depicted in Figures 3.10a-c peak for each radio frequency f at the altitude where $f \approx \frac{\nu}{2\pi}$; the reason for this is discussed below. Second, Figure 3.10b clearly extrapolates into VHF and beyond with truly negligible attenuation rates; thus in our further discussion of VHF and above in Chapter 5 we will return to collisionless plasma models. Finally, in view of nighttime reduction of D-region absorption rates across the radio spectrum, nighttime reduction of MUF (due to reduced electron densities in the ionosphere) is tolerable in shortwave communications.

The absorption rates peak at an altitude where $f \approx \frac{\nu}{2\pi}$because of the $\frac{\frac{\nu}{2\pi f}}{1+(\frac{\nu}{2\pi f})^2}$ factor in the formula for $A(f)$, which, in turns is a consequence of the frequency dependence of the Ohmic component $\frac{Ne^2}{m}\frac{\nu}{\nu^2+\omega^2}$ of collisional conductivity (3.49). Notice that for a fixed plasma density N and radio frequency ω, the Ohmic conductivity peaks at a value $\frac{Ne^2}{2m}$ corresponding to the case $\nu = \omega$, and vanishes in both $\nu \to 0$ and $\nu \to \infty$ limits. The physical reasons for this dependence of Ohmic conductivity is as follows: For $\nu \to \infty$ electron collisions become so frequent that conduction via electron drifts becomes impossible—the medium then resembles an insulator like a lossless perfect dielectric with an infinite resistance. For $\nu \to 0$, on the other hand, electrons do not collide at all and therefore dissipate none of their borrowed energy from the electric field—the medium is once again non-Ohmic. *Maximum power transfer* from the field to the medium via electron dynamics therefore occurs at an intermediate collision frequency $\nu = \omega$, which is a "frequency matching" condition similar to "impedance matching" condition in electrical circuits for maximum power transfer to an external load.

3.8.3 Absorption for $f \sim f_p$

Discussions above focused on attenuation when $f \gg f_p$. This condition is generally valid for waves propagating through the ionosphere. However, the condition breaks down in vertical incidence ionospheric probing near the reflection height $z = z_r$. Near $z = z_r$ we have $f = f_p$ and therefore

$$n = \sqrt{1 - \frac{f_p^2}{f^2}\frac{1 + j\frac{\nu}{2\pi f}}{1 + (\frac{\nu}{2\pi f})^2}} \to \sqrt{1 - \frac{1 + j\frac{\nu}{2\pi f_p}}{1 + (\frac{\nu}{2\pi f_p})^2}}.$$

For the case of $\frac{\nu}{2\pi} \gg f_p$ (near the ionospheric base), this results in

$$n \approx \sqrt{1 - j\frac{2\pi f_p}{\nu}} \approx 1 - j\frac{1}{2}\frac{2\pi f_p}{\nu},$$

and hence

$$k'' = k_o n'' = \frac{2\pi f_p}{2c}\frac{2\pi f_p}{\nu} = \frac{\pi f_p}{c}\frac{f_p}{\nu/2\pi}.$$

The attenuation rate in that case is

$$A(f_p) = 20\log_{10}\{e^{k''(1\,\mathrm{km})}\} = \frac{20000\pi f_p}{c}\frac{f_p}{\nu/2\pi}\log_{10} e \approx 9.1\times 10^{-5}\frac{f_p^2}{\nu/2\pi}\,\frac{\mathrm{dB}}{\mathrm{km}}.$$

Near the ionospheric base $\frac{\nu}{2\pi} \sim 30\,\mathrm{MHz}$. For $f_p = 0.3$ MHz (upper LF) we find $0.27\,\frac{\mathrm{dB}}{\mathrm{km}}$ attenuation while for 0.03 MHz (upper ELF) only $0.0027\,\frac{\mathrm{dB}}{\mathrm{km}}$. As $f_p \to 0$ attenuation vanishes in this regime. ELF waves are reflected from the ionospheric base with virtually no attenuation (see further discussions in the next Chapter).

Strongest attenuation rates for $\frac{f_p^2}{f^2} \approx 1$ are obtained for the case $\frac{\nu}{2\pi} \approx f_p$ — as discussed in the previous section — although the case rarely applies in the ionosphere (since ν varies exponentially, $\frac{\nu}{2\pi} \approx f_p$ in a very thin height range) and is considered an anomaly in radiowave propagation.

The case $\frac{\nu}{2\pi} \ll f_p$, however, is valid in E- and F- regions and important for MF and HF sounding signals reflected from these regions. The corresponding

$$n \approx \sqrt{-j\frac{\nu}{2\pi f_p}} = \sqrt{\frac{\nu}{2\pi f_p}}\frac{1-j}{\sqrt{2}}$$

implies

$$k'' = k_o n'' = \frac{2\pi f_p}{c\sqrt{2}}\sqrt{\frac{\nu}{2\pi f_p}} = \frac{\sqrt{2}\pi}{c}\sqrt{\frac{\nu f_p}{2\pi}},$$

and, therefore,

$$A(f) = 20\log_{10}\{e^{k''(1\,\mathrm{km})}\} = \frac{20000\sqrt{2}\pi}{c}\sqrt{\frac{\nu f_p}{2\pi}}\log_{10} e \approx 1.3\times 10^{-4}\sqrt{\frac{\nu f_p}{2\pi}}\,\frac{\mathrm{dB}}{\mathrm{km}}.$$

Near 100 km altitude in the E-region $\frac{\nu}{2\pi} \approx 0.02\,\mathrm{MHz} = 20\,\mathrm{kHz}$, the attenuation rates for 1-4 MHz signals are $\sim 18 - 36\,\frac{\mathrm{dB}}{\mathrm{km}}$. Since the plasma frequency f_p changes very rapidly with height in the E-region these large attenuation rates only apply over very thin slabs compared to a kilometer and as a result attenuation in E-region reflections is usually tolerable. Higher frequency signals up to $f_c \equiv f_p(z)_{max}$ 9 MHz are reflected from F-region heights where $\frac{\nu}{2\pi}$ is many many e-folds smaller than 20 kHz; therefore the corresponding attenuation rates are quite negligible.

4 Radio links in inhomogeneous media

Solutions of Maxwell's equations in homogeneous media describe waves propagating along straight line *trajectories* defined by constant wavenormal vectors. However, in *inhomogeneous* media with position dependent refractive index n the wavenormals twist and rotate and ray trajectories deviate from straight lines curving towards the regions having the larger refractive index *and* slower phase speed. In fact — as discovered long time ago by Fermat — trajectories followed by light waves between any two points will *minimize* the travel time between the points (see Figure 4.1). In this chapter we will see how Maxwell's equations require the curved ray trajectories consistent *Fermat's principle* and learn some "ray tracing" techniques. We will then study radio links through inhomogeneous ionospheres.

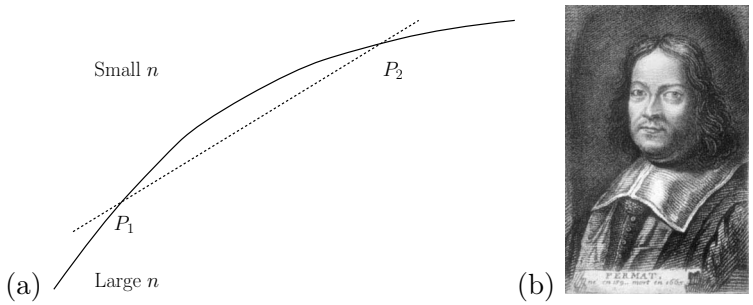


Figure 4.1: (a) Propagation velocity is large in the "small n" region on the top as compared to the "large n region" on the bottom. As a consequence ray trajectory from P_1 to P_2 follows a curved path (solid line) bending towards the region with larger n. The trajectory is the fastest path between points P_1 and P_2 in accordance with *Fermat's principle*. (b) Pierre de Fermat (1601-1665), a lawyer and part-time mathematician.

4.1 Field solutions in plane stratified media

Consider a plane stratified propagation medium with with $\mu = \mu_o$ and a z-dependent permittivity and/or conductivity so that the refractive index

$$n(z) = \sqrt{\epsilon_r(z) + \frac{\sigma(z)}{j\omega\epsilon_o}}$$

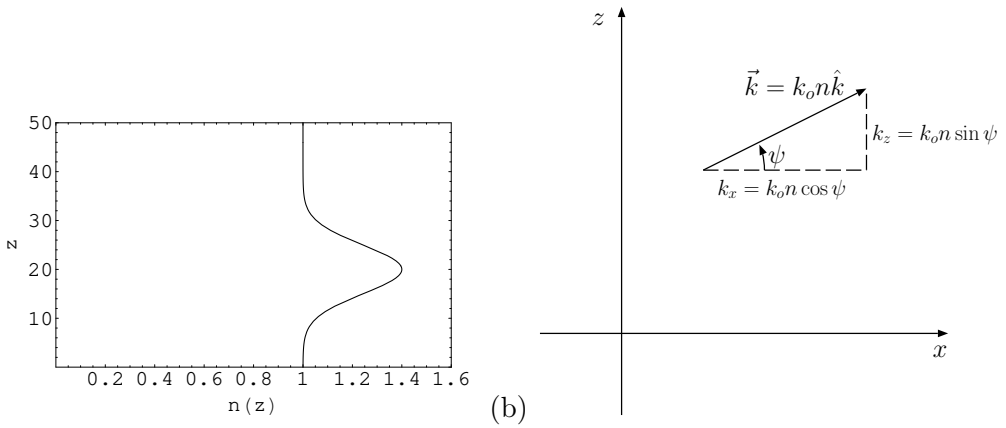


Figure 4.2: (a) An example of a refractive index profile in a plane stratified medium where the refractive index is a function of a single space variable z, and (b) components of a propagation vector $\mathbf{k} = \frac{\omega}{c} n\hat{k}$ with an elevation angle ψ.

varies with z, for instance, as illustrated in Figure 4.2a. The field phasors $\mathbf{E}(\mathbf{r})$ and $\mathbf{H}(\mathbf{r})$ in the medium will satisfy the source free Maxwell's equations

$$\nabla \cdot \mathbf{E} = 0 = \nabla \cdot \mathbf{H}, \quad \nabla \times \mathbf{E} = -j\omega\mu_o \mathbf{H}, \quad \text{and} \quad \nabla \times \mathbf{H} = \sigma\mathbf{E} + j\omega\epsilon\mathbf{E}.$$

The curl equations combine as

$$\nabla \times \nabla \times \mathbf{E} = \nabla(\nabla \cdot \mathbf{E}) - \nabla^2\mathbf{E} = \frac{\omega^2}{c^2} n^2(z)\mathbf{E},$$

and with $\nabla \cdot \mathbf{E} = 0$ lead to the vector wave equation

$$\nabla^2\mathbf{E} + \frac{\omega^2}{c^2} n^2(z)\mathbf{E} = 0.$$

For a monochromatic wave field

$$\mathbf{E}(x, z, t) = Re\{\hat{y}E(z)e^{j(\omega t - k_x x)}\}$$

with a phasor

$$\mathbf{E} = \hat{y}E(z)e^{-jk_x x}$$

and a phase delay at a rate of k_x radians per unit length in x direction, the vector wave equation reduces to the scalar wave equation

$$\frac{d^2E(z)}{dz^2} + \underbrace{\left(\frac{\omega^2}{c^2} n^2(z) - k_x^2\right)}_{\equiv k_z^2(z)} E(z) = 0. \tag{4.1}$$

The WKB and full-wave solutions of (4.1) for $k_x = 0$ were obtained in the previous chapter. Those solutions can be adapted for the $k_x \neq 0$ case by a simple replacement of

$$\frac{\omega}{c}n(z) = k_o n(z) \text{ by } k_z(z) = \pm\sqrt{\frac{\omega^2}{c^2}n^2(z) - k_x^2} = \pm\sqrt{k_o^2 n^2(z) - k_x^2}$$

as appropriate for the upgoing and downgoing waves having the opposite algebraic signs. For instance, the upward propagating WKB solution is changed as

$$E(z) = \frac{\sqrt{k_o n(0)}}{\sqrt{k_o n(z)}} E_o e^{-jk_o \int_0^z n(z')dz'} \rightarrow \frac{\sqrt{k_z(0)}}{\sqrt{k_z(z)}} E_o e^{-j\int_0^z k_z(z')dz'}$$

and it should be valid everywhere except in the vicinity of a reflection heights z_r where $k_z^2(z_r) = 0$. Also, the full-wave solution for $E(z)$ obtained with $k_z^2(z) = k_o^2(\frac{z_r - z}{L})$ will be identical with the solution obtained for the vertical incidence case — i.e., $k^2(z) = k_o^2(\frac{z_r - z}{L})$ — in Chapter 2.

Hence, the WKB field phasor for an upgoing wave, including its x dependence, is

$$\mathbf{E}(x,z) = \hat{y}E(z)e^{-jk_x x} = \hat{y}\frac{\sqrt{k_z(0)}}{\sqrt{k_z(z)}} E_o e^{-j[k_x x + \int_0^z \sqrt{k_o^2 n^2(z') - k_x^2}dz']}.$$

The geometrical interpretation of this wave solution is as follows:

At altitudes z where the solution is well justified, the refractive index $n(z)$ and hence $k_z^2(z) = k_o^2 n^2(z) - k_x^2$ vary slowly as compared to free-space wavelength $\lambda_o \equiv \frac{2\pi}{k_o}$. Hence $n(z)$ can be treated as a constant over distances over many λ_o's, which, in turn, means that the phase $\Phi(x,z)$ of the field solution varies locally like

$$k_x x + \sqrt{k_o^2 n^2 - k_x^2}z = (k_x, k_z) \cdot (x, z) = \mathbf{k} \cdot \mathbf{r},$$

where

$$k_z = \sqrt{k^2 - k_x^2}$$

and

$$k^2 \equiv k_x^2 + k_z^2 = k_o^2 n^2.$$

This means that "locally", the WKB solution describes a plane-wave with a wavenumber $k = k_o n$ just like in a homogeneous medium. However, "globally", the solution requires that we adjust the wavenumber $k = k_o n$ over distances longer than sufficiently many λ_o's by resizing $k_z = \sqrt{k^2 - k_x^2} = \sqrt{k_o^2 n^2 - k_x^2}$ while keeping k_x a constant. As k_z is re-sized while k_x remains fixed, the elevation angle ψ of vector $\mathbf{k} = (k_x, k_z)$ — see Figure 4.2b — naturally varies, describing the deviation of the wave trajectories from straight rays.

The geometrical interpretation just offered is important because it forms the basis of our description of the propagation of *all types of monochromatic disturbances* (e.g., spherical waves arriving from compact antennas, in addition to plane waves) in inhomogeneous media having *all types of refractive index variations* (e.g., spherical stratification is of practical importance in planetary atmospheres and ionospheres). The reason for this is any smooth variation of field $\mathbf{E}(\mathbf{r}, t)$ as well as medium $n(\mathbf{r})$ will appear to have

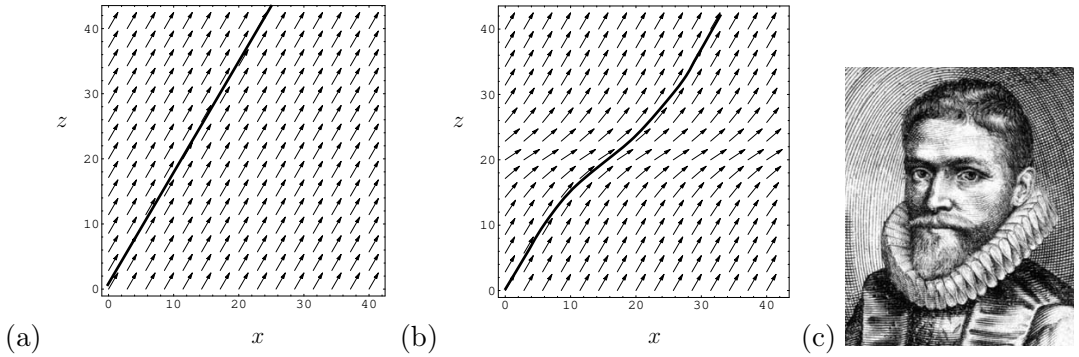


Figure 4.3: Map plots of $\hat{k} = (\cos\psi, \sin\psi)$ for (a) $n(z) = 1$, and (b) $n(z) = 1 - 0.4e^{-(z-20)^2/50}$, consistent with Snell's law $n(z)\cos\psi(z) = n_o \cos\psi_o$ — Solid lines in each $\hat{k}$-map represent ray trajectories. (c) Willebord van Roijen Snell (1580-1626).

planar geometry over sufficiently small regions of space. Accordingly we can always describe monochromatic disturbances propagating in inhomogeneous media in terms of local wavevectors $\mathbf{k}$, and, furthermore, construct from *collections* of such wavevectors (see Figure 4.3) global scale descriptions of how the disturbances spread out along suitably defined ray trajectories.

To describe the notion of ray-trajectory let us return to the plane stratified propagation problem discussed above: We note that

$$k_x = k_o n(z)\cos\psi(z)$$

in terms of the magnitude $k(z) = k_o n(z)$ and the elevation $\psi(z)$ of the wavevector $\mathbf{k}(z) = (k_x, k_z(z))$ anywhere within the wave field. Since k_x and $k_o = \omega/c$ are fixed parameters, it follows that

$$n(z)\cos\psi(z) = \text{const.}$$

This constraint on the elevation angle $\psi(z)$ of the wavevectors $\mathbf{k}(z) = (k_x, k_z(z))$ within the wave field can also be written in the familiar *Snell's law* form (recall from ECE 450)

$$n(z)\cos\psi(z) = n_o \cos\psi_o, \tag{4.2}$$

where n_o and ψ_o refer to $n(z)$ and $\psi(z)$ values found within the wave field at an arbitrary location $(x, z) = (x_o, z_o)$.

An example of how Snell's Law as stated above could be used is illustrated in Figure 4.3. The figure shows the xz-plane variations of the local wavenormals

$$\hat{k} = \frac{(k_x, k_y)}{k} = (\cos\psi, \sin\psi) = \left(\frac{n_o\cos\psi_o}{n(z)}, \sqrt{1 - \left(\frac{n_o\cos\psi_o}{n(z)}\right)^2}\right)$$

in a wave field propagating in a pair of media with (a) $n(z) = 1$, and, (b) $n(z) = 1 - 0.4e^{-(z-20)^2/50}$. Note that in each case $\hat{k} = (\cos\psi, \sin\psi)$ vectors have been computed

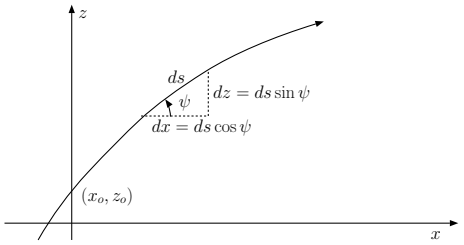


Figure 4.4: Geometry of an upgoing ray passing through a reference point (x_o, z_o). Note that the trajectory slope is $\frac{dz}{dx} = \tan\psi(z)$ and $ds = \sqrt{dx^2 + dz^2}$ is an incremental arc length along the trajectory.

using Snell's law with $z_o = 0$ and $\psi_o = 60^o$. The solid curves included in the figures are the ray trajectories which are — by definition — aligned with the wavenormals $\hat{k}$ at every point within the wave field.

4.2 Ray tracing in plane stratified media

A description of a monochromatic wave field propagating in a plane stratified medium having a refractive index $n = n(z)$ can be given in terms of a set of *ray-trajectory curves* $z(x)$ passing through $(x, z) = (x_o, z_o)$ and satisfying the constraint

$$n(z)\cos\psi(z) = n_o \cos\psi_o, \quad \text{(Snell's Law)}$$

where $n_o \equiv n(z_o)$ and $\psi_o \equiv \psi(z_o)$ denote the refractive index and elevation angle of $\mathbf{k}$ at $(x, z) = (x_o, z_o)$. In the trivial case of $n(z) = n_o$ (i.e., in a homogeneous medium) the trajectories will be straight lines and will hardly offer any benefit in the visualization of the wavefront. However, the trajectories will in general be curved when n changes with z, and many quantitative conclusions can be drawn from the trajectory curves in such cases. For instance, in isotropic propagation media where $\langle \mathbf{E} \times \mathbf{H} \rangle$ is aligned with $\mathbf{k}$, the trajectories derived from Snell's will represent the paths of energy transport in the wave field. The topology of ray-trajectories therefore indicates how the field intensity may vary in the medium in accordance with energy conservation.

A generic ray-trajectory curve is depicted in Figure 4.4 — the trajectory passes through $(x, y) = (x_o, y_o)$ and an *arc length* s measures the distance along the trajectory in such a way that

$$\frac{dx}{ds} = \cos\psi \ \text{ and } \ \frac{dz}{ds} = \sin\psi.$$

The first one of these equations, when multiplied by n, yields

$$n\frac{dx}{ds} = n\cos\psi = n_o\cos\psi_o$$

which is a constant, and therefore implies

$$\frac{d}{ds}(n\frac{dx}{ds}) = 0.$$

On the other hand,

$$\frac{d}{ds}(n\frac{dz}{ds}) = \frac{d}{ds}(n\sin\psi) = \frac{d}{dz}(n\sin\psi)\frac{dz}{ds} = \frac{d}{dz}(n\sin\psi)\sin\psi = \frac{1}{2n}\frac{d}{dz}(n^2\sin^2\psi)$$
$$= \frac{1}{2n}\frac{d}{dz}(n^2(1-\cos^2\psi)) = \frac{1}{2n}\frac{d}{dz}(n^2 - n_o^2\cos^2\psi_o)) = \frac{dn}{dz}.$$

We are also able to combine these results vectorially as

$$\frac{d}{ds}(n\frac{d\mathbf{r}}{ds}) = \nabla n$$

that can be used as the basis of ray tracing problems in non-stratified media with arbitrary variation if the refractive index function $n = n(\mathbf{r})$ such as in spherical stratified media.

Within plane stratified media where $n = n(z)$ we can make of the component equations

$$\frac{d}{ds}(n\frac{dx}{ds}) = 0 \text{ and } \frac{d}{ds}(n\frac{dz}{ds}) = \frac{dn}{dz}$$

as well as

$$n\frac{dx}{ds} = n_o\cos\psi_o \text{ and } n\frac{dz}{ds} = n\sin\psi$$

in a number of ways:

Approximate ray tracing solutions based on $\frac{d}{ds}(n\frac{dz}{ds}) = \frac{dn}{dz}$, where $x \approx s$ and $n \approx n_o$ on the left, and $dn/dz = \alpha$ on the right will be illustrated later on and very handy in tropospheric ray tracing solutions when the rays remain close to the low altitudes (i.e., small z)

The equations can also be transformed, using a change of variables defined by $s = n\tau$ and $ds = nd\tau$, into an easy-to-solve "ballistic form"

$$\frac{d^2x}{d\tau^2} = 0 \text{ and } \frac{d^2z}{d\tau^2} = n\frac{dn}{dz}$$

and

$$\frac{dx}{d\tau} = n_o\cos\psi_o \text{ and } \frac{dz}{d\tau} = n\sin\psi.$$

analogous to projectile motion equations

$$\frac{dx}{dt} = v_o\cos\psi_o \text{ and } \frac{dz}{dt} = v\sin\psi,$$

where v would be a height changing speed of a baseball with a height dependent $v_z = v\sin\psi$ and a height independent $v_x = v_o\cos\psi_o$ — even though in the ray trajectory equations neither n is speed and nor τ time, the analogy between ray tracing and projectile motion is interesting and useful as we will see.

The ballistic equations imply that

$$x(\tau) = n_o\cos\psi_o\tau + x_o$$

and, therefore, a ray trajectory

$$\mathbf{r}(\tau) = \hat{x}(x_o + n_o \cos\psi_o \tau) + \hat{z} z(\tau)$$

where $z(\tau)$ is the solution of the ODE

$$\frac{d^2 z}{d\tau^2} = n \frac{dn}{dz}$$

subject to initial conditions

$$\frac{dz}{d\tau}|_{\tau=0} = n_o \sin\psi_o \quad \text{and} \quad z(0) = z_o,$$

just like the vertical coordinate of a *ballistic object* (e.g., a baseball) with unity mass and initial upward velocity $n_o \sin\psi_o$ subject to a gravitational acceleration $g = n\frac{dn}{dz}$ (instead of $g = -9.8$ m/s^2, the normal case). Clearly, the overall ray-trajectory will exhibit some curvature so long as n has a non-zero derivative.

Example 1: Consider a plane stratified medium with a refractive index $n = n(z)$ such that

$$n^2(z) = 1 + \frac{z}{H},$$

where H is a non-zero real constant. Determine (a) Reflection height $z = z_r$ of a ray passing through $(x, z) = (0, 0)$ with an elevation angle of $\psi = 30^o$, (b) the ray trajectory function $z(x)$ of the ray described in part (a), and (c) the value of x at $z = z_r$.

Solution:

(a) With $z_o = 0$, $n_o = n(z_o) = 1$, and $\psi_o = \psi(z_o) = 30^o$, we have from Snell's law

$$n(z_r)\cos\psi(z_r) = n_o \cos\psi_o = \frac{\sqrt{3}}{2}.$$

At the reflection height $z = z_r$ the ray trajectory is locally horizontal, and, therefore, $\psi(z_r) = 0$ and $\cos\psi(z_r) = 1$. Consequently,

$$n(z_r) = \frac{\sqrt{3}}{2}$$

and hence

$$n^2(z_r) = 1 + \frac{z_r}{H} = \frac{3}{4}.$$

It follows that

$$z_r = -\frac{1}{4}H.$$

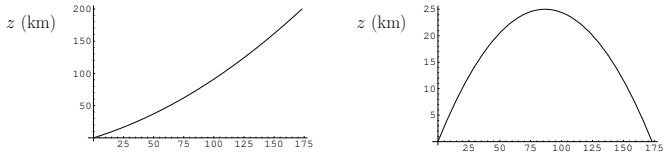
Note that $z_r > 0$ only if $H < 0$. If $H > 0$ then $z_r < 0$, which, in turn, implies either a negative x_r — see part (c) — or the absence of reflections on the ray for $x > 0$.

(b) Since

$$n^2(z) = 1 + \frac{z}{H},$$

it follows that

$$\frac{d}{dz}\frac{n^2}{2} = \frac{1}{2H}.$$



(a) (b)

Figure 4.5: Ray trajectory for $n^2(z) = 1 + \frac{z}{H}$ with $\psi(0) = 30^o$. (a) $H = 100$ km and (b) $H = -100$ km.

Thus we need to solve the differential equation

$$\frac{d^2z}{d\tau^2} = n\frac{dn}{dz} = \frac{d}{dz}\frac{n^2}{2} = \frac{1}{2H}$$

subject to initial conditions

$$\frac{dz}{d\tau}|_{\tau=0} = n_o \sin\psi_o = \sin 30^o = \frac{1}{2} \quad \text{and} \quad z(0) = 0.$$

Integrating $d^2z/d\tau^2$ once we find that

$$\frac{dz}{d\tau} = \frac{\tau}{2H} + \frac{1}{2}$$

which satisfies the initial $dz/d\tau$ at $\tau = 0$. Integrating once more we obtain

$$z(\tau) = \frac{\tau^2}{4H} + \frac{\tau}{2},$$

which can also be cast as

$$z(x) = \frac{x^2}{3H} + \frac{1}{\sqrt{3}}x$$

since

$$\frac{dx}{d\tau}|_{\tau=0} = n_o \cos\psi_o = \frac{\sqrt{3}}{2} \quad \Rightarrow \quad x(\tau) = \frac{\sqrt{3}}{2}\tau.$$

Figure 4.5 shows plots of $z(x)$ curve for $H = \pm 100$ km.

(c) To determine x at $z = z_r$, which we will call x_r, we note that

$$z'(x_r) = \frac{2x_r}{3H} + \frac{1}{\sqrt{3}} = 0,$$

which implies that

$$x_r = -H\frac{\sqrt{3}}{2}.$$

Note that

$$z_r = z(x_r) = -\frac{(-H\frac{\sqrt{3}}{2})^2}{3H} + \frac{1}{\sqrt{3}}(-H\frac{\sqrt{3}}{2}) = \frac{H}{4} - \frac{H}{2} = -\frac{H}{4},$$

consistent with part (a).

Having seen above the analogy of ray tracing to particle dynamics, let us interpret Snell's law in light of the same analogy: "invariance of k_x in a refractive index field varying strictly in the z-direction" being analogous to "invariance of horizontal momentum of a particle in a force field acting in the vertical direction". In *spherical* stratified media, where $n = n(r)$, we will have a counterpart to Snell's law known as *Bouguer's rule* which is analogous to "invariance of *angular* momentum of a particle in a central force field acting in the radial direction". More examples on plane stratified ray-tracing:

Example 2: In a plane stratified medium

$$n^2(z) = 1 - \frac{z^2}{Z^2}, \quad \text{(a refractive index duct)}$$

where Z is some constant scale length. (a) Find the reflection height z_r for a ray launched with an elevation angle $\psi_o > 0$ from height $z = 0$, and, (b) determine the trajectory equation $z(x)$ for a ray passing through $(x, z) = (0, 0)$ with an elevation angle ψ_o.

Solution:

(a) First using Snell's law,

$$n(z_r)\cos\psi(z_r) = n_o \cos\psi_o \quad \Rightarrow \quad n(z_r) = \sqrt{1 - \frac{z_r^2}{Z^2}} = \cos\psi_o,$$

which yields

$$1 - \frac{z_r^2}{Z^2} = \cos^2\psi_o \quad \Rightarrow \quad z_r = Z\sin\psi_o.$$

(b) Next, noting that

$$\frac{d}{dz}\frac{n^2}{2} = -\frac{z}{Z^2}$$

and $n_o = n(z_o) = n(0) = 1$, we establish the ODE to be solved:

$$\frac{d^2z}{d\tau^2} = \frac{d}{dz}\frac{n^2}{2} = -\frac{z}{Z^2} \quad \Rightarrow \quad \frac{d^2z}{d\tau^2} + \frac{z}{Z^2} = 0.$$

This is a familiar 2nd order ODE with a solution of the form

$$z(\tau) = A\cos(\frac{\tau}{Z}) + B\sin(\frac{\tau}{Z}),$$

and with

$$\frac{dz}{d\tau} = -\frac{A}{Z}\sin(\frac{\tau}{Z}) + \frac{B}{Z}\cos(\frac{\tau}{Z}).$$

Applying initial conditions

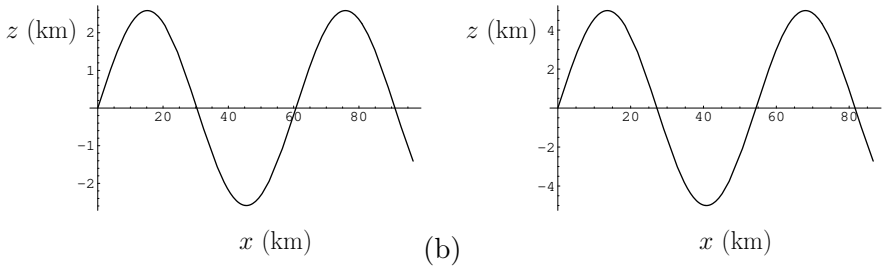
$$z(0) = 0 = A$$

and

$$\frac{dz}{d\tau}|_{\tau=0} = n_o\sin\psi_o = \frac{B}{Z} \quad \Rightarrow \quad B = Z\sin\psi_o = z_r.$$

Thus, we find

$$z(\tau) = z_r\sin(\frac{\tau}{Z}).$$



(a) (b)

Figure 4.6: Ray trajectory for $n^2(z) = 1 - \frac{z^2}{Z^2}$ with $Z = 10$ km. (a) $\psi_o = 15^o$ and (b) $\psi_o = 30^o$.

But we also have

$$\frac{dx}{d\tau}|_{\tau=0} = n_o \cos\psi_o \quad \Rightarrow \quad x = \cos\psi_o \tau \;\Rightarrow\; \tau = \frac{x}{\cos\psi_o}.$$

Therefore, the ray trajectory can also be expressed as

$$z(x) = z_r \sin(\frac{x}{Z\cos\psi_o})$$

which describes the *ducted rays* depicted in Figure 4.6.

Given *any* refractive index function $n = n(z)$, consider the Taylor series expansion of its square $n^2(z)$, about $z = z_o$, which will be of a form

$$n^2(z) = A + B(z - z_o) + C(z - z_o)^2 + \cdots$$

Using this series, we have an approximation

$$\frac{d}{dz}\frac{n^2}{2} \approx \frac{B}{2} + C(z - z_o) = (\frac{B}{2} - Cz_o) + Cz \equiv \alpha + \beta z$$

which is useful for ray tracing purposes close to $z = z_o$ as illustrated in the next example.

Example 3: For z close to z_o it is known that

$$n^2(z) = A + B(z - z_o) + C(z - z_o)^2 + \cdots \quad \text{and} \quad \frac{d}{dz}\frac{n^2}{2} \approx \alpha + \beta z,$$

where $\alpha = (\frac{B}{2} - Cz_o)$ and $\beta = C$ are constant coefficients. Determine the ray trajectory formulas $z = z(\tau)$ and $x = x(\tau)$ in terms of x_o, z_o, and ψ_o for z close to z_o.

Solution:

The pertinent differential equation is

$$\frac{d^2z}{d\tau^2} = \frac{d}{dz}\frac{n^2}{2} \approx \alpha + \beta z \quad \Rightarrow \quad \frac{d^2z}{d\tau^2} - \beta z = \alpha.$$

This is a linear 2nd order ODE with constant coefficients. Its *particular solution* is

$$z = -\frac{\alpha}{\beta}$$

and its *general solution* can be written as

$$z(\tau) = -\frac{\alpha}{\beta} + Ae^{s\tau} + Be^{-s\tau},$$

where A and B are to be determined, and

$$s \equiv \sqrt{\beta}.$$

Since

$$z'(\tau) = sAe^{s\tau} - sBe^{-s\tau},$$

it follows that

$$z(0) = z_o = -\frac{\alpha}{\beta} + A + B \quad \text{and} \quad z'(0) = n_o \sin\psi_o = sA - sB.$$

Hence, adding and subtracting the second equation to and from the product of the first equation and variable s we find that

$$sz_o + n_o \sin\psi_o = -s\frac{\alpha}{\beta} + 2sA \quad \Rightarrow \quad A = \frac{z_o + \alpha/\beta}{2} + \frac{n_o \sin\psi_o}{2s},$$

and

$$sz_o - n_o \sin\psi_o = -s\frac{\alpha}{\beta} + 2sB \quad \Rightarrow \quad B = \frac{z_o + \alpha/\beta}{2} - \frac{n_o \sin\psi_o}{2s}.$$

Thus,

$$\begin{aligned} z(\tau) &= -\frac{\alpha}{\beta} + (z_o + \frac{\alpha}{\beta})\frac{e^{s\tau} + e^{-s\tau}}{2} + n_o \sin\psi_o \frac{e^{s\tau} - e^{-s\tau}}{2s} \\ &= -\frac{\alpha}{\beta} + (z_o + \frac{\alpha}{\beta})\cosh(\sqrt{\beta}\tau) + \frac{n_o \sin\psi_o}{\sqrt{\beta}}\sinh(\sqrt{\beta}\tau). \end{aligned}$$

Also

$$\frac{dx}{d\tau} = n_o \cos\psi_o \;\Rightarrow\; x(\tau) = x_o + n_o \cos\psi_o \tau.$$

The trajectory close to $z = z_o$ is therefore described by

$$\mathbf{r}(\tau) = (x_o + A\cos\psi_o\tau, -\frac{\alpha}{\beta} + (z_o + \frac{\alpha}{\beta})\cosh(\sqrt{\beta}\tau) + \frac{A\sin\psi_o}{\sqrt{\beta}}\sinh(\sqrt{\beta}\tau)).$$

4.3 Eikonal equation and general ray tracing equations

In this section we will describe general ray tracing techniques which can be applied in inhomogeneous propagation media not constrained to be plane stratified. The techniques will be developed from Maxwell's equations in the *geometrical optics* approximation and are based on the solutions of an *eikonal equation* governing the phase of the field solutions in the frequency domain.

4.3.1 Geometrical optics and eikonal equation

Consider the Maxwell's curl equations

$$\nabla \times \mathbf{E} = -j\omega\mu_o\mathbf{H} \quad \text{and} \quad \nabla \times \mathbf{H} = j\omega\tilde{\epsilon}\mathbf{E}$$

where

$$\tilde{\epsilon} = \epsilon + \frac{\sigma}{j\omega}$$

is the effective permittivity of the medium at the wave frequency ω. Spatial variations in $\tilde{\epsilon}$ forces us to abandon the possibility of plane TEM wave solutions. However, for slowly varying $\tilde{\epsilon}$ we can try solutions of the form

$$\mathbf{E} = \mathbf{E}_o(\mathbf{r})e^{-j\Phi(\mathbf{r})} \text{ and } \mathbf{H} = \mathbf{H}_o(\mathbf{r})e^{-j\Phi(\mathbf{r})} \tag{4.3}$$

involving amplitude functions $\mathbf{E}_o(\mathbf{r})$ and $\mathbf{H}_o(\mathbf{r})$ and a phase function $\Phi(\mathbf{r})$, all to be determined by using the Maxwell's equations. If $\tilde{\epsilon} = \tilde{\epsilon}(\mathbf{r})$ varies slowly with $\mathbf{r}$, and if the frequency ω is sufficiently large, we expect that the exponential $e^{-j\Phi(\mathbf{r})}$ will vary much more rapidly with position $\mathbf{r}$ than $\mathbf{E}_o(\mathbf{r})$ and $\mathbf{H}_o(\mathbf{r})$ — this is true because as we have seen previously, local variation of $e^{-j\Phi(\mathbf{r})}$ should be like in homogeneous media where a higher ω implies a faster varying $e^{-j\mathbf{k}\cdot\mathbf{r}}$. In that case, for sufficiently large ω, the curl equations above will be well approximated by

$$-j\nabla\Phi \times \mathbf{E} = -j\omega\mu_o\mathbf{H} \quad \text{and} \quad -j\nabla\Phi \times \mathbf{H} = j\omega\tilde{\epsilon}\mathbf{E}$$

after using the identity $\nabla \times (\mathbf{A}f) = f\nabla \times \mathbf{A} + \nabla f \times \mathbf{A}$ and ignoring terms due to $\nabla \times \mathbf{E}_o$ and $\nabla \times \mathbf{H}_o$. This is the *geometrical optics* approximation of Maxwell's curl equations, justified at large $\omega = 2\pi f$ and short $\lambda = c/f$.

Let us set

$$\mathbf{k} \equiv \nabla\Phi \tag{4.4}$$

for notational convenience, and then re-write the equations above constraining the phase function Φ as

$$\mathbf{k} \times \mathbf{E} = \omega\mu_o\mathbf{H} \quad \text{and} \quad -\mathbf{k} \times \mathbf{H} = \omega\tilde{\epsilon}\mathbf{E}.$$

But notice, these equations are identical with the plane-wave form of Maxwell's curl equations obtained for a homogeneous medium, and algebraically they imply

$$\mathbf{H} = \frac{\hat{k} \times \mathbf{E}}{\eta} \text{ and } \mathbf{E} = \eta\mathbf{H} \times \hat{k}$$

with

$$\eta = \sqrt{\frac{\mu_o}{\tilde{\epsilon}}}, \ \hat{k} = \frac{\mathbf{k}}{|\mathbf{k}|}, \text{ and } \mathbf{k} \cdot \mathbf{k} = \omega^2\mu_o\tilde{\epsilon} \equiv \frac{\omega^2}{c^2}n^2$$

as before. We can make us of these relationships so long as we keep in mind that in the present case $\mathbf{k} = \nabla\Phi$ is a position dependent "wave vector".

Thus, the dispersion relation

$$\mathbf{k}\cdot\mathbf{k}-\frac{\omega^2}{c^2}n^2=0$$

above is really a short-hand for a partial differential equation

$$\nabla\Phi\cdot\nabla\Phi-\frac{\omega^2}{c^2}n^2=0\ \ \Rightarrow\ \ (\frac{\partial\Phi}{\partial x})^2+(\frac{\partial\Phi}{\partial y})^2+(\frac{\partial\Phi}{\partial z})^2-\frac{\omega^2}{c^2}n^2=0 \tag{4.5}$$

known as the *eikonal equation* — the phase function $\Phi=\Phi(x,y,z)$ is also known as the *eikonal* in the geometrical optics literature and *by definition* surfaces of $\Phi=$ const. are the wavefronts of the wave field[1].

The overall objective in geometrical optics is finding the solution of the eikonal equation for a given refractive index field $n=n(\mathbf{r})$ and boundary conditions specifying a wavefront — local values of $|\mathbf{E}_o(\mathbf{r})|$ can always be deduced using energy conservation arguments, as we have already seen in Section 3.2, once the ray-trajectories are known (see below). Formally, the solution of the eikonal equation can be cast as

$$\Phi(\mathbf{r})=\int_0^{(x,y,z)}\mathbf{k}\cdot d\mathbf{l} \tag{4.6}$$

which is a path integral where the integration path is arbitrary but the use of the proper $\mathbf{k}$ is essential. This follows from an analogy with the well-known solution methods of electrostatics: the analogy is facilitated by a comparison of

$$\mathbf{E}=-\nabla V\ \ \Leftrightarrow\ \ \nabla\times\mathbf{E}=0$$

from electrostatics with

$$\mathbf{k}\equiv\nabla\Phi\ \ \Leftrightarrow\ \ \nabla\times\mathbf{k}=0$$

from geometrical optics. Because $\mathbf{k}$ is the gradient of a *scalar* eikonal, it is necessarily a curl-free vector field, and hence its path-independent contour integral (4.6) yields back the eikonal[2], just as in electrostatics the path-independent contour integral

$$V(\mathbf{r})=-\int_0^{(x,y,z)}\mathbf{E}\cdot d\mathbf{l} \tag{4.7}$$

[1] Needless to say, the concept of eikonal is also valid in homogeneous propagation media. In such a medium, a plane TEM wave will have an eikonal $\Phi(x,y,z)=k_xx+k_yy+k_zz$, for instance, with $k_x^2+k_y^2+k_z^2\equiv k^2=\frac{\omega^2}{c^2}$, say, in vacuum, whereas the eikonal of a spherical TEM wave will be $\Phi(x,y,z)=kr$, where $r\equiv\sqrt{x^2+y^2+z^2}$. We see that the eikonal will be influenced by both the medium and the type of the TEM disturbance produced by its source.

[2] A distinction should be made between the eikonal $\int_0^{(x,y,z)}\mathbf{k}\cdot d\mathbf{l}$ and the so-called "optical action" $\int_0^{(x,y,z)}kdl$, where $k\equiv|\mathbf{k}|=\frac{\omega}{c}n$, and an arbitrary integration path is implied between 0 and (x,y,z) in both integrals. Since in general $|\mathbf{k}\cdot d\mathbf{l}|\le|\mathbf{k}||d\mathbf{l}|=kdl$, action $\int_0^{(x,y,z)}kdl\ge\int_0^{(x,y,z)}\mathbf{k}\cdot d\mathbf{l}$, with equality being achieved only when the integration path happens to be an actual "ray path" along which $\mathbf{k}$ and $d\mathbf{l}$ are entirely co-aligned. Hence, determining the path that minimizes the action integral — e.g., by calculus of variations — is another ray tracing technique. This is in fact Fermat's *principle of least time*, since optical action $\int_0^{(x,y,z)}kdl$ is proportional to the propagation time $\int_0^{(x,y,z)}\frac{dl}{v_p}=\int_0^{(x,y,z)}n\frac{dl}{c}=\omega^{-1}\int_0^{(x,y,z)}kdl$ from 0 to (x,y,z) via the integration path. See e.g., Molcho and Censor, *Am. J. Phys.*, **54**, 351 (1986); Evans and Rosenquist, *Am. J. Phys.*, **54**, 876 (1986).

of electrostatic field $\mathbf{E}(\mathbf{r})$ yields back the electrostatic potential $V(\mathbf{r})$.

In electrostatics one can obtain $\mathbf{E}(\mathbf{r})$ to be used in (4.7) by using Gauss's law

$$\nabla \cdot \mathbf{E} - \frac{\rho}{\epsilon_o} = 0$$

along with the curl-free condition $\nabla \times \mathbf{E} = 0$, unless it is easier to obtain $V(\mathbf{r})$ directly from Poisson's equation. Analogously, in geometrical optics, one can obtain $\mathbf{k}(\mathbf{r})$ to be used in (4.6) by using the dispersion relation

$$\mathbf{k} \cdot \mathbf{k} - \frac{\omega^2}{c^2} n^2 = 0$$

along with the curl-free condition $\nabla \times \mathbf{k} = 0$, unless it is easier to obtain $\Phi(\mathbf{r})$ directly by guessing the solution of the eikonal equation.

If the eikonal $\Phi(\mathbf{r})$ cannot be easily guessed, then "ray tracing" to determine $\mathbf{k} = \mathbf{k}(\mathbf{r})$ comes into play. We already know how to do that in a plane stratified medium with $n = n(z)$ by using *Snell's law* and/or solving the ray tracing equations resembling ballistic equations. Next we extend those methods for the case of an arbitrary refractive index field $n = n(\mathbf{r})$.

4.3.2 General ray tracing equations

In the general dispersion relation

$$\underbrace{\mathbf{k} \cdot \mathbf{k} - \frac{\omega^2}{c^2} n^2}_{\equiv D(\mathbf{r}, \omega, \mathbf{k})} = 0,$$

where $n = n(\mathbf{r})$, consider $\mathbf{k} = \nabla\Phi$ and position variable $\mathbf{r}$ to be independent variables linked by some "delay parameter" τ such that $\mathbf{r} = \mathbf{r}(\tau)$ and $\mathbf{k} = \mathbf{k}(\tau)$. Since the dispersion relation stipulates that

$$D(\mathbf{r}, \omega, \mathbf{k}) = 0, \tag{4.8}$$

function $D(\mathbf{r}, \omega, \mathbf{k})$ must be an invariant independent of τ, so that

$$\frac{dD}{d\tau} = \frac{\partial D}{\partial \mathbf{r}} \cdot \frac{d\mathbf{r}}{d\tau} + \frac{\partial D}{\partial \mathbf{k}} \cdot \frac{d\mathbf{k}}{d\tau} = 0,$$

in which we have used the shorthand

$$\frac{\partial D}{\partial \mathbf{r}} \cdot \frac{d\mathbf{r}}{d\tau} \equiv \frac{\partial D}{\partial x}\frac{dx}{d\tau} + \frac{\partial D}{\partial y}\frac{dy}{d\tau} + \frac{\partial D}{\partial z}\frac{dz}{d\tau} = \nabla D \cdot \frac{d\mathbf{r}}{d\tau}$$

and

$$\frac{\partial D}{\partial \mathbf{k}} \cdot \frac{d\mathbf{k}}{d\tau} \equiv \frac{\partial D}{\partial k_x}\frac{dk_x}{d\tau} + \frac{\partial D}{\partial k_y}\frac{dk_y}{d\tau} + \frac{\partial D}{\partial k_z}\frac{dk_z}{d\tau} = \nabla_k D \cdot \frac{d\mathbf{k}}{d\tau}$$

where ∇_k denotes a $\mathbf{k}$-space gradient in analogy with the usual ∇. Thus

$$\nabla D \cdot \frac{d\mathbf{r}}{d\tau} + \nabla_k D \cdot \frac{d\mathbf{k}}{d\tau} = 0,$$

which we can satisfy by selecting

$$\frac{d\mathbf{k}}{d\tau} = f(\tau)\nabla D \text{ and } \frac{d\mathbf{r}}{d\tau} = -f(\tau)\nabla_k D \tag{4.9}$$

where $f(\tau)$ is an arbitrary function for which we have a complete freedom of choice (to be utilized as we see fit later on). These equations happen to be the *general ray tracing equations* that we were after.

Notice that, since $\nabla \times \nabla D = 0$ for any scalar D, the first ray tracing equation above implies that

$$\nabla \times \frac{d\mathbf{k}}{d\tau} = 0 \quad \Rightarrow \quad \nabla \times \mathbf{k}(\tau) = 0$$

(so long as $\nabla \times \mathbf{k}(0) = 0$) as required by $\mathbf{k} = \nabla\Phi$. The choice of function $f(\tau)$ above is related to the physical interpretation of the parameter τ. It can be shown that with the choice

$$f(\tau) = \frac{1}{\partial D/\partial\omega},$$

the right-hand-side of the second ray tracing equation becomes the group velocity vector and thus the parameter τ becomes the group delay (as verified in HW). This choice is ideal for ionospheric applications but we will postpone its use for the time being in favor of a simpler choice,

$$f(\tau) = -\frac{c}{2\omega},$$

invoked below:

Since

$$D = \mathbf{k}\cdot\mathbf{k} - \frac{\omega^2}{c^2}n^2 = k_x^2 + k_y^2 + k_z^2 - \frac{\omega^2}{c^2}n^2,$$

we have

$$\nabla D = \frac{\partial D}{\partial \mathbf{r}} = -\frac{\omega^2}{c^2}\nabla n^2$$

and

$$\nabla_k D = \frac{\partial D}{\partial \mathbf{k}} = (2k_x, 2k_y, 2k_z) = 2\mathbf{k},$$

so that the ray tracing equations (4.9) can be written as

$$\frac{d\mathbf{k}}{d\tau} = -f(\tau)\frac{\omega^2}{c^2}\nabla n^2 \text{ and } \frac{d\mathbf{r}}{d\tau} = -f(\tau)2\mathbf{k}.$$

Choosing

$$f(\tau) = -\frac{c}{2\omega}, \tag{4.10}$$

we simplify these as

$$\frac{d\mathbf{k}}{d\tau} = \frac{\omega}{c}\nabla\frac{n^2}{2} \text{ and } \frac{d\mathbf{r}}{d\tau} = \frac{c}{\omega}\mathbf{k}. \tag{4.11}$$

While the parameter τ has no neat physical interpretation with choice (4.10), the resulting ray tracing equations combine to yield

$$\frac{d^2\mathbf{r}}{d\tau^2} = \nabla\frac{n^2}{2} \text{ and } \frac{d\mathbf{r}}{d\tau} = n\hat{k}, \tag{4.12}$$

since $\frac{c}{\omega}\mathbf{k} = \frac{c}{\omega}\frac{\omega}{c}n\hat{k}$, giving us the vector based generalization of the ballistic-type ray tracing equations we have used in Section 3.2 (also mentioned in a footnote in Section 3.2). They can be used in inhomogeneous media with an arbitrary $n(\mathbf{r})$ so long as the medium is isotropic[3].

We will approach our study of ray tracing in spherical stratified media in the next section using these equations. Note that the relations (4.12) imply that

$$\frac{d}{d\tau}(\mathbf{r}\times\frac{d\mathbf{r}}{d\tau}) = \frac{d\mathbf{r}}{d\tau}\times\frac{d\mathbf{r}}{d\tau} + \mathbf{r}\times\frac{d^2\mathbf{r}}{d\tau^2} = \mathbf{r}\times\frac{d^2\mathbf{r}}{d\tau^2} = 0$$

in a spherical stratified medium where $n = n(|\mathbf{r}|) = n(r)$, because in that case $\nabla n^2 \propto \hat{r}$. The conserved quantity $\mathbf{r}\times d\mathbf{r}/d\tau$ in spherical stratified problems is analogous to "angular momentum" in particle dynamics and provides us with *Bouguer's rule* used in the next section.

In spherical stratified propagation medium with $n = n(r)$ — e.g., the mean state in a planetary atmosphere — the general ray tracing equations (4.12) imply that

$$\frac{d^2\mathbf{r}}{d\tau^2} = \hat{r}\frac{\partial}{\partial r}\frac{n^2}{2} \text{ and } \frac{d\mathbf{r}}{d\tau} = n(\hat{r}\sin\psi + \hat{\theta}\cos\psi)$$

where $\psi(\tau)$ is the local elevation angle of the ray measured from constant $n(r)$ surfaces as shown in Figure 4.7.

These are the same as the equations of motion of a particle of unity mass subject to a central force such as a planetary gravitational field — the velocity $n\hat{k}$ of the particle[4] will vary with time along its trajectory because of centripetal acceleration $\hat{r}\frac{\partial}{\partial r}\frac{n^2}{2}$. Such "free-floating" particles under the influence of planetary gravitational fields are known to follow trajectories in which the *angular momentum* $\mathbf{r}\times\frac{d\mathbf{r}}{d\tau}$ *per unit mass* is conserved, that is,

$$\mathbf{r}\times\frac{d\mathbf{r}}{d\tau} = \text{const.} \;\Rightarrow\; n\mathbf{r}\times(\hat{r}\sin\psi + \hat{\theta}\cos\psi) = nr\cos\psi(\hat{r}\times\hat{\theta}) = \text{const.}$$

[3]Ray tracing in *anisotropic* media can also be handled with the set (4.9) so long as the dispersion relation $D(\mathbf{r},\omega,\mathbf{k})$ properly describes the anisotropy involved, namely the dependence of ω on vector $\mathbf{k}$ rather than on magnitude $|\mathbf{k}|$. In anisotropic media the vectors $d\mathbf{r}/d\tau$ and $\hat{k}$ are in general *not* aligned, and, as a result, energy transport along the ray-trajectories $\mathbf{r}(\tau)$ does not take place in the direction of $\mathbf{k}$, the direction of phase progression. See e.g., Weinberg, *Phys. Rev.*, **126**, 1899 (1962).

[4]Note that the analogy is more specifically with unity mass particles having a total energy of $E = K.E. + P.E. = 0$, since for a particle with speed $|n|$ and unity mass the kinetic energy $K.E. = \frac{1}{2}n^2$, and, furthermore, when the "gravitational force" is specified as $-\nabla P.E. = \hat{r}\frac{\partial}{\partial r}\frac{n^2}{2}$, it follows that the potential energy $P.E. = -\frac{1}{2}n^2$. See Arnaud, *Amer. J. Phys.*, **44**, 1067 (1976). Particle trajectories with a zero total energy are known as *parabolic* trajectories — as opposed to elliptical ($E < 0$) or hyperbolic ($E > 0$). Also note that while a central gravitational force invariably varies as $\propto r^{-2}$, a central force in optics due to spherical stratified $n(r)$ can have an arbitrary r dependence and lead to "zero energy" trajectories not encountered in planetary dynamics.

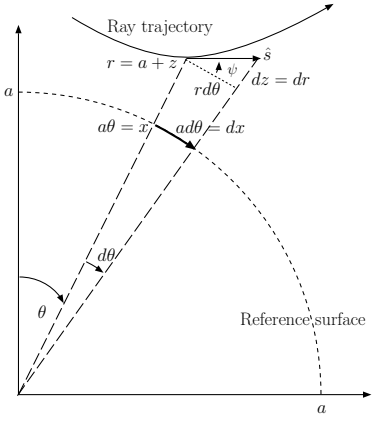


Figure 4.7: Geometry of ray tracing in a spherical stratified medium with a reference surface at $r = a$. Coordinate $x = a\theta$ is measured along the reference surface and ray height z is measured upward from the reference surface (i.e., $z < 0$ is below the reference surface).

Since the unit vector $\hat{r} \times \hat{\theta} = \hat{\phi}$ is a constant across the motion plane depicted in Figure 4.7, this condition reduces to

$$rn(r) \cos \psi(r) = \text{const.},$$

which is a counterpart of Snell's law known as *Bouguer's rule* appropriate for spherical-stratified ray-tracing problems.

4.4 Ray tracing in spherical stratified media

Consider a straight line ray trajectory with a closest approach of r_o distance to the origin as depicted in Figure 4.8a in a homogeneous region of space with some constant refractive index n_o. At an arbitrary point on the ray, when the distance to the origin is r, the angle of the trajectory measured from the local tangent plane is $\psi(r)$ such that $\cos \psi(r) = r_o/r$ as can be verified in Figure 4.8a . This can be stated as a constraint

$$r \cos \psi(r) = r_o \cos \psi_o$$

with $\psi_o \equiv \psi(r_o) = 0°$, or even as

$$rn_o \cos \psi(r) = r_o n_o \cos \psi_o$$

where n_o is the fixed refractive index in the region.

Now, if the refractive index n_o in the region were replaced by a spherical stratified refractive index function $n(r)$, the straight line trajectory of Figure 4.8a would be replaced by a curved trajectory like the one shown in Figure 4.8b, constrained by

$$rn(r) \cos \psi(r) = r_o n_o \cos \psi_o,$$

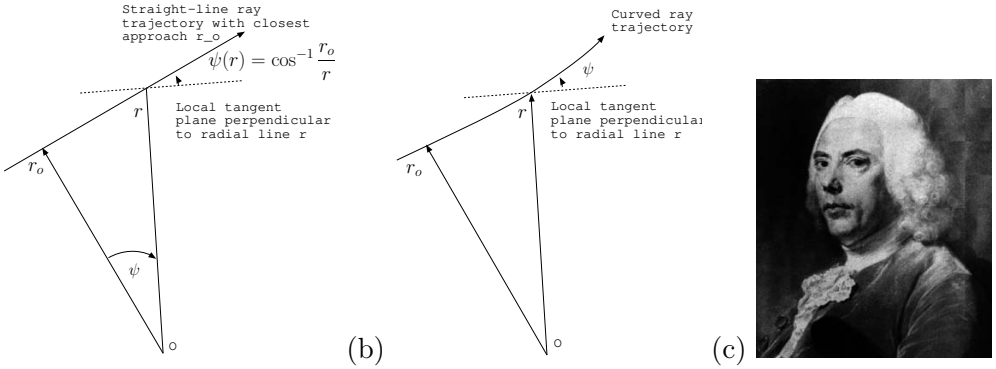


Figure 4.8: (a) Straight line ray geometry in the absence of refraction (n=constant), (b) curved trajectory in a spherical stratified setting $n = n(r)$, and (c) Pierre Bouguer (1698-1785) https://en.wikipedia.org/wiki/Pierre_Bouguer.

which can also be stated as

$$r_1 n(r_1) \cos\psi(r_1) = r_2 n(r_2) \cos\psi(r_2)$$

with r_1 and r_2 denoting the distances from the origin to any two points on the ray, however close or distant they may be. The justification for this constraint, known as *Bouguer's rule*[5], is that it reduces to[6] familiar *Snell's law*

$$n(r_1) \cos\psi(r_1) = n(r_2) \cos\psi(r_2)$$

when nearby points on the ray with $r_1 \approx r_2$ are considered, as it should because in that limit a plane stratified formulation well approximates the spherical one.

In summary, we will base our ray tracing equations for spherical stratified case on Bouguer's rule that can be stated as

$$r n(r) \cos\psi(r) = \text{const.}$$

just as we used Snell's law

$$n(z) \cos\psi(z) = \text{const.}$$

in plane stratified problems. Moreover, by using

$$r \equiv a + z$$

[5]While Snell's Law $n(z)\cos\psi(z) = \text{const.}$ conserves the horizontal momentum of light particles in vertically stratified media, Bouguer's Rule $rn(r)\cos\psi(r) = \text{const.}$ conserves their angular momentum in spherical stratified media — remember that $k = k_o n$ is photon momentum divided by $\hbar$ while angular mometum per unit mass is radial distance r times the transverse velocity componet $v\cos\psi$ — see http://hyperphysics.phy-astr.gsu.edu/hbase/amom.html — so that the conserved angular momentum in orbit is $rv(r)\cos\psi(r)$ per unit mass.

[6]in addition to reducing to the straight line ray constraint $r\cos\psi(r) = r_o$ when $n(r) = n_o$

in Bouguer's rule we have

$$(a+z)n(a+z)\cos\psi(a+z) = \text{const.} \quad \Rightarrow \quad \underbrace{(1+\frac{z}{a})n(a+z)}_{\equiv m(z)}\cos\psi(a+z) = \text{const.},$$

where a is a conveniently chosen radius measured from the origin and z is the "height" of r above a as depicted in Figure 4.7. Furthermore, labeling $\psi(a+z)$ as $\psi(z)$ for convenience, we can now re-cast Bouguer's rule as

$$m(z)\cos\psi(z) = \text{const.}$$

in terms of a so-called *modified index*

$$m(z) \equiv (1+\frac{z}{a})n(z+a).$$

Expressed in terms of the modified index $m(z)$ and "height z above a", Bouguer's rule appears identical to Snell's law[7].

As a consequence, ray tracing techniques developed earlier for the plane stratified geometry will apply directly in the spherical stratified case if $|z| \ll a$ and $n(z)$ of our earlier ray tracing equations is replaced by $m(z)$. With those conditions fulfilled, the trajectory solutions $z(x)$ will describe "height z above a" of the trajectories found in spherical stratified problems, with x measuring the distance along the curved surface of the $r = a$ sphere as illustrated in Figure 4.7.

Notice that in the spherical stratified case, Bouguer's rule

$$m(z)\cos\psi(z) = m_o\cos\psi_o$$

remains valid for all $z \equiv r - a$ along the ray trajectories. Since for $z \equiv r - a \ll a$, $dx \equiv ad\theta \approx rd\theta$ as shown in Figure 4.7, $dx \approx ds\cos\psi$ and $dz \equiv dr = ds\sin\psi$ as in a plane stratified medium. Thus, taking $ds = md\tau$ we can reach a set of ray tracing equations

$$\frac{d^2z}{d\tau^2} = \frac{d}{dz}\frac{m^2}{2}$$

subject to initial conditions

$$\frac{dz}{d\tau}|_{\tau=0} = m_o\sin\psi_o \;\text{ and }\; z(0) = z_o$$

that should work with all trajectories with $|z| = |r-a| \ll a$.

Example 1: Determine the constraint for the reflection height $z_r = r_r - a$ in a spherical stratified medium in terms of the corresponding modified index $m(z)$.

[7] and is identical to Snell's law in $z \ll a$ limit!

Solution: Since at the reflection height $\psi(z_r) = 0$ and $\cos\psi(z_r) = 1$, Bouguer's rule $m(z)\cos\psi(z) = m_o\cos\psi_o$ indicates that

$$m(z_r) = m_o\cos\psi_o$$

for any ray passing through a reference height z_o with an elevation angle ψ_o. Note that $m_o \equiv m(z_o)$.

When the modified index $m(z)$ in a spherical stratified medium is linear in z (or can be approximated linearly for small z) it becomes possible to apply a special ray tracing trick. We will describe the *trick* after showing in the next example how the modified index of the Earth's troposphere can be linearized for small z.

Example 2: Within the first few kilometers of the atmosphere refractive index varies[8] as

$$n(z) = n_o - \frac{z}{4a},$$

as a function of height z above the Earth's spherical surface of radius $a = 6371.2$ km. Here $n_o \approx 1.0003$ is a constant often approximated as unity. This formula is known as standard tropospheric refractive index model. Determine the modified index $m(z)$ for the troposphere and find a linear approximation for $m(z)$ which is good for small z.

Solution: Given that the modified index

$$m(z) \equiv (1 + \frac{z}{a})n(z),$$

and approximating the index $n(z)$ given above as $1 - \frac{z}{4a}$, it follows that for the standard troposphere

$$m(z) = (1 + \frac{z}{a})n(z) \approx (1 + \frac{z}{a})(1 - \frac{z}{4a}) = 1 - \frac{z}{4a} + \frac{z}{a} - \frac{z^2}{4a^2}.$$

The linear approximation for $m(z)$ is clearly

$$m(z) \approx 1 - \frac{z}{4a} + \frac{z}{a} = 1 + \frac{z}{\frac{4}{3}a}.$$

The use of linear approximation can be justified whenever

$$\frac{z^2}{4a^2} \ll \frac{z}{\frac{4}{3}a} \quad \text{or} \quad z \ll 3a.$$

Since the tropospheric refractive index model $n(z) \approx 1 - \frac{z}{4a}$ is itself only valid for $z \ll a$, the linear model for $m(z)$ can be used at all tropospheric heights where the $n(z)$ model is valid.

Now, to the description of the ray tracing trick mentioned above:

[8] A crude justification: $n(0) > 1$ because the atmosphere is the densest at the lowest altitude and therefore it behaves least like vacuum ($n = 1$). As the height increases atmospheric density gets reduced so $n(z)$ should decrease from the surface value — that accounts for the term $-\frac{z}{4a}$.

The linearized modified index formula for the troposphere obtained above can also be expressed as

$$m(z) = 1 + \frac{z}{a'} \quad \text{where} \quad a' \equiv \frac{4}{3}a$$

is known as the *modified Earth radius*. Note that if the actual radius of Earth were $a' = \frac{4}{3}a$ instead of a and if tropospheric refractive index were actually 1 instead of $1 - \frac{z}{4a}$, then tropospheric modified index would still be

$$m(z) = (1 + \frac{z}{a'})n(z) = (1 + \frac{z}{a'})1 = 1 + \frac{z}{a'},$$

the same as in the actual troposphere surrounding our actual Earth.

Ray tracing equations will therefore yield identical trajectories, e.g., $z(x)$, in the actual and fictitious tropospheres described above for the same set of initial conditions x_o, z_o, and ψ_o. But in the fictitious troposphere above an a'-radius Earth there is no refraction because the corresponding refractive index is a constant and all *rays follow straight line trajectories*. The upshot is, in standard tropospheric refraction problems we can avoid using the ray tracing equations altogether by pretending that the Earth radius is $a' = \frac{4}{3}a$ (instead of a) and the rays follow straight line trajectories. Here is an illustration:

Example 3: The horizon distance of an antenna positioned at $z = h$ above the surface of a planet with radius a' is given as

$$x_h = \sqrt{2a'h}$$

(see Chapter 4) provided that the refractive index in the atmosphere of the planet is some constant n_o — i.e., in the absence of any refraction. Determine the horizon distance of an antenna positioned at $z = h$ above Earth's surface taking into account the standard tropospheric refraction.

Solution: Note that $x_h = \sqrt{2a'h}$ can be interpreted as $x(h)$, where $x(z)$ is the straight line ray trajectory shown in Figure 4.9a through $(x, z) = (0, 0)$ with an elevation angle $\psi_o = 0$ above a spherical surface with radius a'. A second ray shown in Figure 4.9b through $(x, z) = (0, 0)$ with an elevation angle $\psi_o = 0$ above a radius a sphere will have the same trajectory equation $x(z)$ if the modified index above $r = a$ is $m(z) = 1 + \frac{z}{a'}$. Hence in Figure 4.9b $x(h) = x_h = \sqrt{2a'h}$ also. Since the modified index in the Earth's troposphere is $m(z) \approx 1 + \frac{z}{a'}$, where $a' = \frac{4}{3}a$, it follows that including the effects of tropospheric refraction the horizon distance of an antenna at height $z = h$ is

$$x_h = \sqrt{2(\frac{4}{3}a)h} = \sqrt{\frac{4}{3}}\sqrt{2ah} \approx 1.1547\sqrt{2ah}.$$

As shown in Example 3, tropospheric refraction extends the horizon distance of an antenna by about 15% assuming standard conditions.

Example 4: Consider a low flying airplane at a height of 100 m above the ground at a ground distance of 10 km from a ground based radar. A ray departing the radar with an elevation angle ψ_o passes through the airplane. Determine ψ_o with and without standard tropospheric refraction.

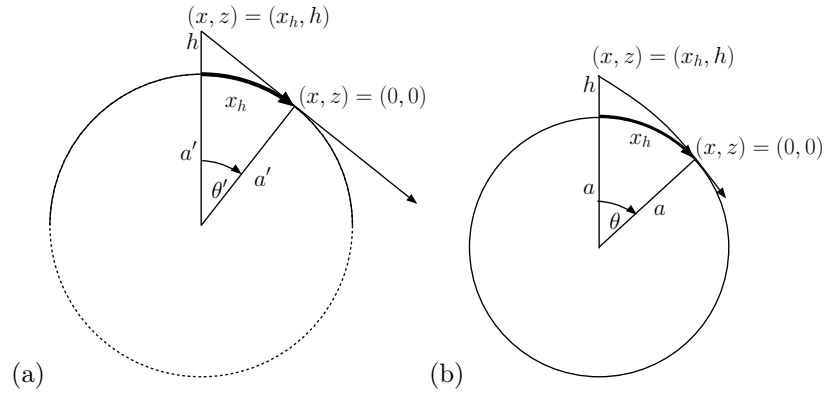


Figure 4.9: (a) An expanded radius non-refracting equivalent of (b) showing a refracted ray trajectory in a spherical stratified troposphere.

Solution: We will first determine ψ_o ignoring tropospheric refraction. Then the problem is essentially an exercise in geometry involving a triangle with sides a and $a + 1/10$ and included angle $\theta = 10/a$ (see Figure 4.10) where a is the Earth radius. Note that

$$d\cos\psi_o = (a + 1/10)\sin\theta,$$

where d is the distance from the radar to the airplane, and also the "law of sines" implies that

$$\frac{d}{\sin\theta} = \frac{a}{\sin(90^o - \theta - \psi_o)} = \frac{a}{\cos(\theta + \psi_o)}.$$

Combining the two expressions, we have

$$\frac{a\cos\psi_o}{\cos(\theta + \psi_o)} = a + \frac{1}{10},$$

and using this formula with $\theta = 10/a$, $a = 6371$ km, we find $\psi_o \approx 0.5279^o$. When we use $a' = \frac{4}{3}a$ in the same set of equations instead of a we find $\psi_o \approx 0.5392^o$ which takes into account standard tropospheric refraction. Since tropospheric refraction causes a downward bending of ray paths, the launch angle ψ_o required to detect the airplane is slightly larger than the 0.5279^o estimate obtained ignoring the refraction effect.

Example 5: In the troposphere of PlanetX with some radius a, the refractive index varies as

$$n(z) = n_o(1 + \frac{z}{K})$$

where n_o and K are unknown constants. Furthermore, it is found that trajectory of rays launched from $(x, z) = (0, h)$ with an elevation angle $\psi_o = 0$ is $z(x) = h$. In other words the rays circulate the planet at a fixed height. In addition, measurements show that at the surface of the planet the propagation speed is c. Determine n_o and K and find an explicit expression for $n(z)$.

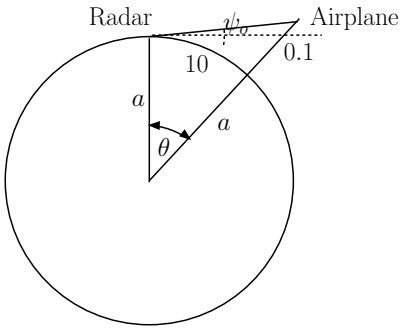


Figure 4.10: Geometry for Example 4. Note that the figure is grossly out of proportion.

Solution: This odd situation can only arise if the modified index $m(z)$ of PlanetX does not depend on z around $z = h$, i.e., $\frac{dm}{dz} = 0$ at $z = h$, because of ray tracing equation $\frac{d^2z}{d\tau^2} = \frac{d}{dz}\frac{m^2}{2} = m\frac{dm}{dz}$ and the fact that $m(h) = 0$ is not possible. Since

$$m(z) = (1 + \frac{z}{a})n(z) = n_o(1 + \frac{z}{a})(1 + \frac{z}{K}) = n_o(1 + \frac{z}{a} + \frac{z}{K} + \frac{z^2}{aK}),$$

it follows that

$$\frac{dm}{dz} = n_o(\frac{1}{a} + \frac{1}{K} + \frac{2z}{aK}).$$

Setting $\frac{dm}{dz}$ to zero at $z = h$, we find

$$n_o(\frac{1}{a} + \frac{1}{K} + \frac{2h}{aK}) = 0 \;\; \Rightarrow \;\; K = -a - 2h.$$

Furthermore $n(0) = n_o = 1$ is required because propagation speed at $z = 0$ is c. Hence,

$$n(z) = 1 - \frac{z}{a + 2h}.$$

Example 5 above indicates that if tropospheric refractive index here on Earth were $n(z) \approx 1 - \frac{z}{a}$ instead of $n(z) \approx 1 - \frac{z}{4a}$ we might have seen the back of our necks when we looked straight-on in any direction we pleased (with $\psi_o = 0$!). Fortunately, $|\frac{dn}{dz}|$ is four times too small for that to happen.

However, during temperature inversion events in the troposphere $n(z) \approx 1 - \frac{z}{a}$ can become locally valid near the surface. When that happens the horizon recedes dramatically and spherical Earth effectively (optically) flattens outs. It is said that the Viking king Eric the Red got a glimpse of some islands half-way to Iceland when looking out from the northern English shore during such a "tropospheric ducting" event (see next section for further details) — the event precipitated the Viking invasion of Iceland and the eventual discovery of America by Vikings way before Christopher Columbus. The Viking legends also make reference to mysterious "holes" that occasionally appear on the ocean surface and "nearly swallow up" the warriors and their longboats. These legendary holes are events of $n(z) \approx 1 - \frac{Kz}{a}$ with $K > 1$, again due to tropospheric temperature

inversions which can be very frequent and strong in Northern Atlantic region. The inversions typically occur when warm air above the relatively warm ocean surface moves laterally over a cold ice sheet or a patch of colder ocean surface.

Exercise 1: Suppose that tropospheric refractive index around Planet Naboo is $n(z) \approx 1 - \frac{2z}{a}$, where a is one Naboo radius. Find the linearized modified index $m(z)$ for Naboo for small z and express it in terms of a modified Naboo radius a'. What is a'? Derive and plot the trajectory $z(x)$ of a ray launched with $\psi_o = +0.1^o$ from a height $z = \frac{a}{100}$ above the surface of Naboo. Explain why the inhabitants of Naboo feel very "trapped".

Exercise 2: In a spherical stratified medium modified index $m(z) = 1 + Ae^{-\frac{z^2}{H^2}}$. Consider a ray launched from $(x, z) = (0, 0)$ with an elevation angle $\psi_o > 0$. If for a given ψ_o a reflection height z_r exists, then the ray will come back to the surface, reflect from the surface with the same elevation angle ψ_o and continue its journey around the spherical surface as a "ducted wave". Determine the range of ψ_o's in terms of A and/or H which for which ducting is possible. Hint: For ducted ψ_o's reflection condition $m(z_r) = m(0)\cos\psi_o$ will have a real valued solution.

4.5 Tropospheric ducting

The standard tropospheric modified index

$$m(z) = (1 + \frac{z}{a})(n_o - \frac{z}{4a}) \approx 1 + \frac{z}{a'},$$

where $a' = 4a/3$, has a positive gradient $dm/dz > 0$, which implies that under standard refraction conditions rays launched parallel to the ground will escape the troposphere. Under anomalous tropospheric conditions, however, $m(z)$ can exhibit downward gradients $dm/dz < 0$ over limited heights, in which case certain rays can be trapped and ducted to long distances well beyond the standard horizon.

Assuming a ray with an elevation angle ψ_o at a height z_o, Bouguer's rule implies that everywhere else on the same ray

$$m\cos\psi = m_o\cos\psi_o$$

in terms of $m_o \equiv m(z_o)$. Restricting our attention to small elevation angles ψ and ψ_o, the same condition can be written as

$$m - m\psi^2/2 = m_o - m_o\psi_o^2/2$$

after using the small angle expansion $\cos\psi \approx 1 - \psi^2/2$. If we further assume that $m\psi^2 \approx \psi^2$ and $m_o\psi_o^2 \approx \psi_o^2$ at the altitudes of interest, the expression above can be solved for the elevation angle ψ as

$$\frac{\psi}{\sqrt{2}} = \pm\sqrt{m - (m_o - \psi_o^2/2)}. \tag{4.13}$$

Evidently, the elevation angle ψ of the ray depends at each height z on the difference of the modified index m and the so called "characteristic" of the ray defined as

$$m_o - \psi_o^2/2.$$

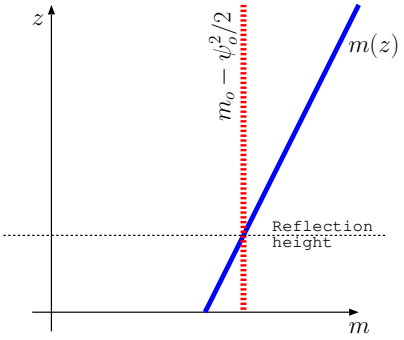


Figure 4.11: The modified index profile (in blue) in a *standard troposphere* and a "characteristic" of a ray having an elevation angle of $\pm\psi_o$ at a height z_o having a modified index $m_o \equiv m(z_o)$. The ray resides only at those altitudes where $m(z)$ exceeds the characteristic shown. The reference height z_o used in the specification of the ray characteristic can be located at any height above the intersection height.

Graphically, the characteristic of a ray appears as a vertical line on the mz-plane as shown in Figure 4.11 for a troposphere with a standard modified index profile. According to (4.13), the ray is forbidden at those altitudes where the modified index is smaller than the characteristic. Therefore, the ray depicted in Figure 4.11 only resides at heights above the intersection point of the characteristic and the $m(z)$ curve, and the intersection point corresponds to the reflection height of the ray (the ray is reflected upward). Note that the larger the ψ_o, the smaller the characteristic $m_o - \psi_o^2/2$ for a fixed m_o, and, therefore, the lower the reflection height. In fact for a sufficiently large ψ_o there is no upward reflection height and a downcoming ray hits the ground.

Figure 4.12a depicts a non-standard tropospheric index profile portraying a so-called a *surface duct*. An *elevated duct* with sample characteristics is depicted in Figure 4.12b.

Surface as well as elevated ducts associated with tropospheric temperature inversions with $dT/dh > 0.1$ K/m can facilitate propagation to distances well beyond the standard horizon. Typical tropospheric ducts may have vertical dimensions of about several m to a few 100 m and radio waves in the VHF-UHF band and at higher frequencies can be very effectively guided in such ducts to long distances. Ducting effects of course also occur in the visible frequency band, and give rise to a phenomenon known as "arctic mirage". Arctic mirage, as discussed in the last section, gives the impression of a flattened earth, or even occasionally earth with an upward curving surface, and occurs frequently at high latitudes particularly when high pressure systems are established upon cold surfaces. By contrast, the "desert mirage" phenomenon, common at low latitudes, is associated with sub-standard refraction when the optical horizon distance is reduced, and the sensation of earth's sphericity is enhanced. Desert mirage occurs when tropospheric temperature falls at a significantly faster rate than the standard 6.5 K/km rate.

Our discussion above does not explain the frequency dependence of tropospheric ducting phenomenon as it is exclusively based on ray tracing arguments (geometrical optics).

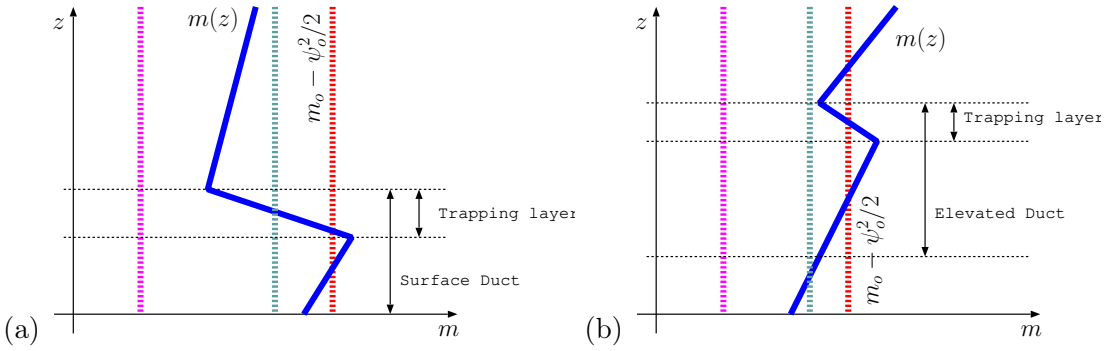


Figure 4.12: (a) The modified index profile of a *surface duct*. Rays with the red and green colored characteristics are "ducted" within the surface duct (provided frequency dependent guidance conditions are fulfilled). The ray with the green characteristic suffers ground reflections at the bottom edge while the ray with the red characteristic is reflected from the profile at both the top and the bottom. A third ray depicted by the magenta characteristic (in has the largest ψ_o of all the three rays for a fixed m_o) is not ducted at all as it never intersects the $m(z)$ curve and hence it never goes through a reflection height. (b) The modified index profile of a *elevated duct*. Only the red ray is ducted. Green ray is reflected upward and escapes the troposphere (like in Figure 4.11).

To understand the frequency dependence of propagation in a tropospheric duct one has to examine additional guidance conditions required for ducting. We shall conclude this section by examining the guidance condition of horizontally polarized modes in a surface duct. Recall, from Section 3.1, that below some reflection height z_r, y-polarized WKB fields can be expressed as

$$E_y(x,z) \propto \frac{\sqrt{k_z(0)}}{\sqrt{k_z(z)}} E_o(e^{-j\int_0^z k_z(z')dz'} + Re^{j\int_0^z k_z(z')dz'}),$$

where we can take

$$k_z(z) = \sqrt{k^2 - k_x^2} = k_o\sqrt{m^2(z) - m_o^2 c_o^2}$$

and

$$R = je^{-j2\int_0^{z_r} k_z(z')dz'} = e^{-j(2k_o\int_0^{z_r}\sqrt{m^2(z)-m_o^2c_o^2}dz' - \pi/2)}.$$

In these equations it is most convenient to consider m_o and c_o to be the modified index and the elevation cosine $\cos\psi$ evaluated at $z = 0$, respectively. Since we have seen that ducting will only be possible for small elevation angles ψ_o in practice, the corresponding ground reflection coefficient $R_h \approx -1$ and hence it is required that $E_y(x,0) = 0$ to satisfy the boundary condition at the bottom boundary of a surface duct. Consequently, $1 + R = 0$, or

$$R = e^{-j(2k_o\int_0^{z_r}\sqrt{m^2(z)-m_o^2c_o^2}dz' - \pi/2)} = -1,$$

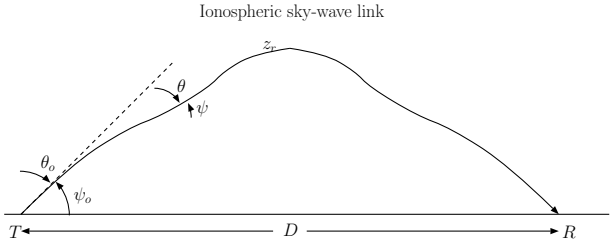


Figure 4.13: A "sky-wave" link between a ground based transmitter T and a ground based receiver at R a distance D away.

which leads to

$$2k_o \int_0^{z_r} \sqrt{m^2(z) - m_o^2 c_o^2}\, dz' - \pi/2 = \pi + 2\pi p,$$

or, equivalently,

$$\int_0^{z_r} \sqrt{m^2(z) - m_o^2 c_o^2}\, dz' = \frac{\lambda_o}{2}(p + \frac{3}{4}), \tag{4.14}$$

where $p = 0, 1, 2, \cdots$, $\lambda_o = 2\pi/k_o = c/f$ is the free space wavelength, and z_r is the reflection height where $\sqrt{m^2(z) - m_o^2 c_o^2} \to 0$. The left side of (4.14) represents the "phase height of reflection" of the trapped rays and (4.14) indicates that such rays can constitute guided modes provided that the phase height of reflection matches certain multiples of $\lambda_o/2$. At any frequency only a discrete set of trapped rays with discrete reflection heights below the top boundary of the surface duct can satisfy the guiding condition, and as the free space wavelength approaches the depth of the duct, the number of such rays or guided modes will decrease rapidly. Guidance condition for an elevated duct can be specified as

$$\int_{z_b}^{z_r} \sqrt{m^2(z) - m_o^2 c_o^2}\, dz' = \frac{\lambda_o}{2}(p + \frac{1}{2}), \tag{4.15}$$

where z_b and z_r denote reflection heights at the bottom and top of the duct.

4.6 Sky-wave links in a plane stratified ionosphere

Oblique propagating monochromatic waves in a plane stratified ionosphere will follow curved trajectories like one shown Figure 4.13. Such curved trajectories can be determined with the ray tracing techniques we have already learned provided that refractive index $n(z)$ is known at the pertinent wave frequency.

However, it is often the case that all we know about the ionosphere is the corresponding ionosonde curve, that is, some $h'(f_v)$ data collected by a vertical incidence sounder operated with a carrier frequency f_v. In this section we will find out what we can learn about ionospheric "sky-wave" links between ground based transmitters T and receivers R using only $h'(f_v)$ data. We will see that many aspects of the link can be determined using the ionogram $h'(f_v)$ and three simple rules to be derived in this section.

Secant Law, Breit and Tuve Theorem, and Martyn's Law

In plane-stratified ionospheric sky-link calculations Snell's law is usually expressed as

$$n(z)\sin\theta(z) = n_o \sin\theta_o \tag{4.16}$$

in terms of zenith angle θ shown in Figure 4.13. Also, the reference altitude z_o is always taken as the ground $z_o = 0$, such that $\theta_o = \theta(z_o) = \theta(0)$ corresponds to the *sub-ionospheric zenith angle* of the ray path. Finally, $n_o = n(z_o) = n(0)$ is always taken as unity since the ionosphere starts only above $z \approx 50$ km. This decision is a consequence of the facts that

$$n(z) = \sqrt{1 - \frac{f_p^2(z)}{f^2}},$$

and $f_p(0) \approx \sqrt{80.6N(0)} \approx 0$. Thus, in general, we have

$$n(z)\sin\theta(z) = \sin\theta_o, \tag{4.17}$$

and, at the reflection height $z = z_r$ of a sky-wave trajectory,

$$n(z_r)\sin\theta(z_r) = \sqrt{1 - \frac{f_p^2(z_r)}{f^2}} = \sin\theta_o \quad \Rightarrow \quad f_p(z_r) = f\cos\theta_o.$$

The above result, namely,

$$f_p(z_r) = f\cos\theta_o, \tag{4.18}$$

indicates that if a transmission with frequency f and sub-ionospheric zenith angle θ_o is to reflect from some ionospheric height, then the plasma frequency at the reflection height, namely $f_p(z_r)$, has to equal the product $f\cos\theta_o$. The same result is sometimes written as

$$f = f_p(z_r)\sec\theta_o \tag{4.19}$$

and called the *Secant Law*.

Consider next a vertical incidence signal transmitted into the ionosphere, at a frequency, say, f_v, which reflects from the *same height* $z = z_r$ as the oblique transmission with wave parameters f and θ_o. Since at the reflection height of the vertical incidence signal $f_p(z_r) = f_v$, it follows that we can re-write (4.18) as

$$f_v = f\cos\theta_o, \tag{4.20}$$

an important relation between the frequencies of two different transmissions which reflect from the same ionospheric height z_r. Furthermore, the *virtual reflection height* of the vertical transmission signal with frequency f_v is of course

$$h'(f_v) = h'(f\cos\theta_o), \tag{4.21}$$

where $h'(\cdot)$ is the ionogram function discussed in the previous chapter.

The very last observation — that is, $h'(f_v) = h'(f\cos\theta_o)$ — is important, because it immediately tells us that an oblique propagating signal with a carrier frequency f and direction θ_o will be reflected from the ionosphere *if and only if* $h'(f\cos\theta_o)$ is finite (because round-trip group delay of the vertical incidence signal at frequency $f_v = f\cos\theta_o$ is proportional to $h'(f\cos\theta_o)$). Conversely, the transmission will penetrate through the ionosphere if $h'(f\cos\theta_o)$ is infinite.

Having established how we can choose f and θ_o to ensure an ionospheric reflection — use

$$f\cos\theta_o < f_c,$$

where f_c is the maximum plasma frequency of the ionosphere known as the *critical* frequency — let us next see how we can calculate the link distance D for a given f and θ_o combination.

Consider the x and z components of the group velocity of the oblique signal given by

$$\frac{dx}{d\tau} = v_g\sin\theta = cn\sin\theta \quad \text{and} \quad \frac{dz}{d\tau} = v_g\cos\theta = cn\cos\theta,$$

where we used $v_g = cn$ which is true for plasmas; also $d\tau$ stands here for an infinitesimal group delay. Since angle θ in these expressions varies with height z as constrained by Snell's law

$$n\sin\theta = \sin\theta_o,$$

the first equation above[9] reads as

$$\frac{dx}{d\tau} = c\sin\theta_o \quad \Rightarrow \quad x = c\tau\sin\theta_o,$$

assuming that $x = 0$ for $\tau = 0$ at the transmission site T in Figure 4.13. Thus, we have a simple description of the x-component of the trajectory of a wave packet propagating in the ionosphere.

If the wave packet reaches its reflection height z_r after a group delay of τ_r, the corresponding horizontal displacement is

$$x_r = c\tau_r\sin\theta_o$$

and equals the horizontal projection of a hypotenuse $c\tau_r$ shown in Figure 4.14a that corresponds to the distance that would have been covered at the speed of light over the duration τ_r had the pulse travelled on a straight line trajectory ignoring the presence of the ionosphere. Now, the vertical projection of the same "virtual" hypotenuse is

$$h = c\tau_r\cos\theta_o$$

and is called the *virtual reflection height* of the link as shown in Figure 4.14a. Dividing x_r by the virtual reflection height h we find

$$\frac{x_r}{h} = \tan\theta_o.$$

[9]The second equation above will be used in the proof of Martyn's Law later in the section.

Accordingly, the total link distance between T and R is

$$D = 2x_r = 2h \tan\theta_o \tag{4.22}$$

in terms of virtual reflection height h.

This final result, known as *Breit and Tuve Theorem,* is very useful because according to *Martyn's Law*[10] , proven below, the virtual reflection height of the link is

$$h = h'(f_v), \tag{4.23}$$

where the right-hand side is the ionogram function for the ionosphere evaluated at a frequency $f_v = f\cos\theta_o$.

Combining the Breit and Tuve Theorem, Martyn's Law, and Secant Law $f_v = f\cos\theta_o$, we find that the sky-wave link distance indicated in Figure 4.14b must be

$$D = 2\tan\theta_o h'(f\cos\theta_o). \tag{4.24}$$

Proof of Martyn's Law: Given that

$$\frac{dz}{d\tau} = v_g\cos\theta = cn\cos\theta,$$

it follows that an infinitesimal group path of ionospheric transmission is

$$\begin{aligned} cd\tau &= \frac{dz}{n\cos\theta} = \frac{dz}{n\sqrt{1-\sin^2\theta}} = \frac{dz}{\sqrt{n^2-n^2\sin^2\theta}} \\ &= \frac{dz}{\sqrt{1-\frac{f_p^2}{f^2}-\sin^2\theta_o}} = \frac{dz}{\sqrt{\cos^2\theta_o-\frac{f_p^2}{f^2}}} = \frac{dz}{\cos\theta_o\sqrt{1-\frac{f_p^2}{f_v^2}}} \end{aligned}$$

[10] *Proof of Martyn's Law:* Given that

$$\frac{dz}{d\tau} = v_g\cos\theta = cn\cos\theta,$$

it follows that

$$\begin{aligned} cd\tau &= \frac{dz}{n\cos\theta} = \frac{dz}{n\sqrt{1-\sin^2\theta}} = \frac{dz}{\sqrt{n^2-n^2\sin^2\theta}} \\ &= \frac{dz}{\sqrt{1-\frac{f_p^2}{f^2}-\sin^2\theta_o}} = \frac{dz}{\sqrt{\cos^2\theta_o-\frac{f_p^2}{f^2}}} = \frac{dz}{\cos\theta_o\sqrt{1-\frac{f_p^2}{f_v^2}}} \end{aligned}$$

with $f_v = f\cos\theta_o$. Thus

$$h = c\tau_r\cos\theta_o = \cos\theta_o\int_0^{z_r}\frac{dz}{\cos\theta_o\sqrt{1-\frac{f_p^2}{f_v^2}}} = \int_0^{z_r}\frac{dz}{\sqrt{1-\frac{f_p^2}{f_v^2}}} = h'(f_v)$$

according to the ionogram definition established in the last chapter. QED

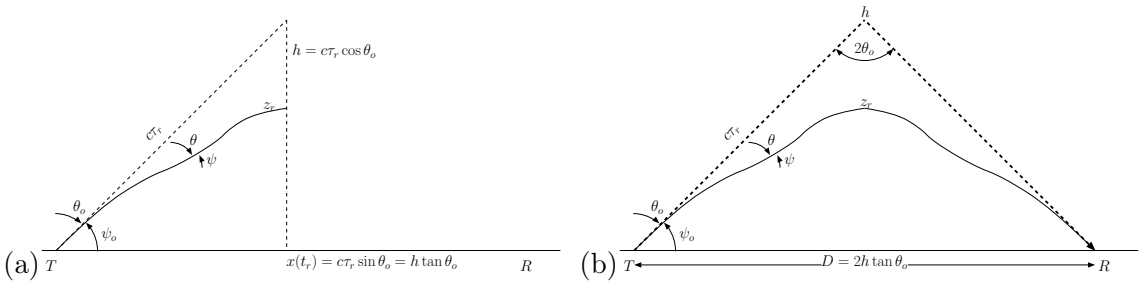


Figure 4.14: (a) Upleg portion of a sky-wave ray trajectory to the reflection height z_r, and (b) the full ray path and the virtual trajectory (dashed line) described by Breit and Tuve theorem.

with $f_v = f\cos\theta_o$. The vertical projection of the total group path to the reflection height is therefore

$$h = c\tau_r \cos\theta_o = \cos\theta_o \int_0^{z_r} \frac{dz}{\cos\theta_o \sqrt{1 - \frac{f_p^2}{f_v^2}}} = \int_0^{z_r} \frac{dz}{\sqrt{1 - \frac{f_p^2}{f_v^2}}} = h'(f_v)$$

according to the ionogram definition established in the last chapter. Thus the virtual reflection height of the link is the same as the virtual reflection height of the vertical transmission which is reflected from the same true height as the oblique link transmission. QED

Example 1: In a plane stratified ionosphere the ionogram $h'(f_v) = 50 + 10 f_v$ km where vertical incidence frequency f_v is specified in MHz units. Determine the T to R distance D of a sky-wave link with a transmission frequency $f = 10$ MHz and an initial zenith angle $\theta_o = 45^o$.

Solution: The virtual path for the link will have a height given by

$$h = h'(f\cos\theta_o) = h'(10\frac{1}{\sqrt{2}}) = 50 + 10\frac{10}{\sqrt{2}} \approx 120.71 \text{ km}.$$

Since x_r and h form a right-angle triangle with a 45^o apex angle, $x_r \approx 120.71$ km and $D \approx 241$ km.

Example 2: In a single-humped parabolic ionosphere

$$h'(f_v) = 50 + \frac{1}{2}50\frac{f_v}{10}\ln\left\{\frac{10 + f_v}{10 - f_v}\right\} \text{ km}.$$

What is the *maximum usable frequency* for a transmission with a zenith angle $\theta_o = 45^o$ to form a sky-wave link (i.e., be reflected from the ionosphere)?

Solution: For the given ionosphere the critical frequency $f_c = 10$ MHz (note that as $f_v \to 10$ MHz, $h'(f_v) \to \infty$). Hence the usable frequencies $f = f_v \sec\theta_o = \frac{f_v}{\cos 45^o} = \sqrt{2} f_v$ are constrained by $f_v < f_c = 10$ MHz. The maximum usable frequency is MUF $= \sqrt{2}10 \approx 14.14$ MHz. Transmissions with $f >$MUF will penetrate the ionosphere and fail to form a sky-wave link.

Example 3: Determine the link distance D if the transmission described in Example 2 uses $f = 8\sqrt{2}$ MHz.

Solution: The virtual reflection height will be $h = h'(8\sqrt{2}\cos 45^o) = h(8) \approx 93.94$ km. Since with $\theta_o = 45^o$, $x_r = h$, it follows that $D = 2h \approx 187.9$ km.

Example 4: What will be the group delay from T to R of a $f = 10$ MHz transmission with a $\theta_o = 60^o$ initial angle in the ionosphere described in Example 2?

Solution: The height of the virtual path triangle is

$$h'(10\cos 60^o) = h(5) = 63.73 \text{ km}.$$

Therefore the slant length to virtual reflection height is $c\tau_r = \frac{63.73\,\text{km}}{\cos 60^o} = 127.465$ km, implying that $\tau_r = \frac{127.465\times 10^3\,\text{m}}{3\times 10^8\,\text{m/s}} \approx 0.425$ ms. Therefore, T to R group delay is $2\tau_r \approx 0.85$ ms.

Skip distance and MUF

In general in forming a sky-link to a distance D we want to use as high a carrier frequency f as we can use in order to minimize D-region absorption effects described in the previous chapter. Of course an arbitrarily large f will penetrate the ionosphere, and, thus, there is a *maximum usable frequency* — a MUF — in order to reach a given D. The MUF of course also depends on the pertinent ionosphere $h(f_v)$ and is in fact the maximum value of f that satisfies the link formula

$$D = 2\tan\theta_o h'(f\cos\theta_o)$$

for a given D over all θ_o. Since

$$\tan\theta_o = \frac{\sin\theta_o}{\cos\theta_o} = \frac{f\sqrt{1-\cos^2\theta_o}}{f\cos\theta_o} = \frac{\sqrt{f^2-f_v^2}}{f_v},$$

the same equation can be re-arranged as

$$h'(f_v) = \frac{D}{2}\frac{f_v}{\sqrt{f^2-f_v^2}} \tag{4.25}$$

which is useful for graphical determinations of all usable f, including the MUF. The left side of this equation is an ionogram and the right side can be interpreted as a curve — known as *transmission curve* — in variable f_v that can be drawn on the ionogram for given values of f and D parameters.

For instance Figure 4.15a shows a transmission curve corresponding to the $f = 11$ MHz and $D = 500$ km case. Each intersection of the curve with the ionosonde curve $h'(f_v)$ represents a link solution and yields a virtual height of $h'(f_v) = h'(f\cos\theta_o)$ of a virtual sky-wave path to distance $D = 500$ km (and an initial angle $\theta_o = \cos^{-1}(f_v/f)$). Two intersections means that two distinct links are formed at 11 MHz to $D = 500$ km.

Transmission curve for $f = 13$ MHz and $D = 500$ km shown in Figure 4.15b has no intersection with the ionogram curve $h'(f_v)$ and thus frequency $f = 13$ MHz unusably

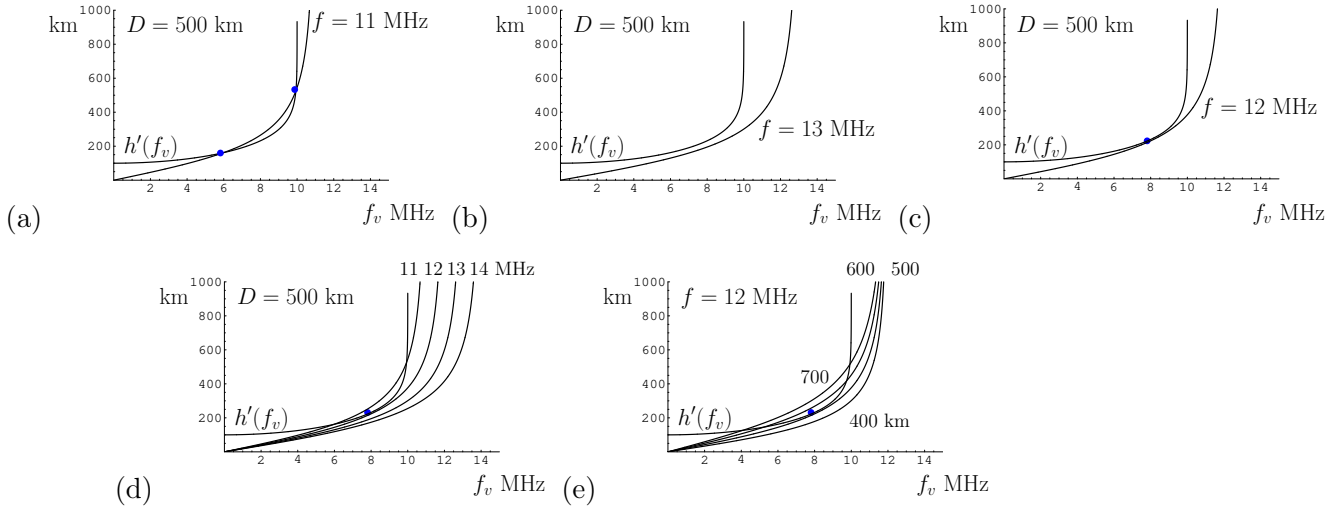


Figure 4.15: Transmission curves overlaid an ionogram $h'(f_v)$; (a), (b), and (c) depict $f = 11$ MHz, 13 MHz, and 12 MHz cases for $D = 500$ km, respectively; (d) shows a family of transmission curves for different f at a fixed $D = 500$ km and (e) shows a family for different D at a fixed $f = 12$ MHz.

high to reach $D = 500$ km under the given ionosphere. However, transmission curve for $f = 12$ MHz shown in Figure 4.15c touches $h'(f_v)$ curve at a single point indicating that at $f = 12$ MHz there is a single ionospheric channel to $D = 500$ km.

Figure 4.15d shows a family of transmission curves for $D = 500$ km and a range of f values labeling the individual transmission curves. Clearly, $f = 12$ MHz appears to be the *maximum usable frequency* (MUF) to reach distance $D = 500$ km — for larger f the transmission equation (4.25) has no solution.

Also, we can refer to $D = 500$ km as *skip distance* at $f = 12$ MHz, because, as shown Figure 4.15e, there is no solution of the transmission equation (4.25) for $D < 500$ km with a fixed $f = 12$ MHz — the transmission curves in Figure 4.15e are for a fixed $f = 12$ MHz and different D values labeling the individual curves.

In summary, when we use $f =$MUF to reach a certain distance D, our transmissions will also reach to $x > D$ (in fact via two or more paths, as implied by Figure 4.15e) but not to $x < D$. The *skip zone* $x < D$ at the MUF of link distance D is inaccessible because all reflected rays at the MUF fall back at $x > D$ while the remaining rays at the same frequency (with smaller values of θ_o penetrate the ionosphere (i.e., do not satisfy the reflection condition $f \cos\theta_o < f_c$). The only way to reach the skip zone is to reduce f to values less than MUF.

Exercise 1: Explain why if $f < f_c$ there is no skip zone. Skip zone only arises if transmission frequency f exceeds the ionospheric critical frequency f_c.

Skip distance and MUF under a parabolic ionosphere

Consider a parabolic electron density profile that leads to a plasma frequency profile described as

$$f_p^2(z) = f_c^2(1 - \frac{(z - z_m)^2}{s^2})$$

for $h_o \equiv z_m - s < z < z_m + s$ and zero elsewhere. The corresponding virtual height function or ionogram was calculated in the previous chapter as

$$h'(f) = h_o + s\frac{1}{2}\frac{f}{f_c}\ln(\frac{1 + f/f_c}{1 - f/f_c}),$$

where h_o can be viewed as *ionospheric base height* and s the "semi-thickness" of the ionosphere.

Even though a real ionosphere will in general be multi-humped and non-parabolic, the density variation near the peak altitude is usually well approximated by a parabolic model. Thus, for sufficiently large $\frac{f}{f_c}$ any reasonable $h'(f)$ will resemble the form given above with some s and h_o values which can be inferred from the slope and intercept of h plotted against $\frac{1}{2}\frac{f}{f_c}\ln(\frac{1+f/f_c}{1-f/f_c})$. Thus, the method to be developed below will apply even to non-parabolic ionospheres in an approximate sense.

We continue by substituting the parabolic $h'(f)$ expression above within the general link distance formula

$$D = 2h'(f\cos\theta_o)\tan\theta_o$$

to obtain

$$D = 2h_o\tan\theta_o + s\sin\theta_o\frac{f}{f_c}\ln(\frac{1 + f\cos\theta_o/f_c}{1 - f\cos\theta_o/f_c}).$$

Now, for a fixed f, the minimum value of D above corresponds to the skip distance under the given parabolic ionosphere (i.e., h_o, s, f_c), which, in turn, corresponds to some specific θ_o value that can be found by solving the equation $\frac{\partial D}{\partial \theta_o} = 0$. Since $\frac{\partial D}{\partial \theta_o} = 0$ implies that

$$2\frac{h_o}{s} + \underbrace{\frac{f_v}{f_c}\ln(\frac{1 + f_v/f_c}{1 - f_v/f_c})}_{\equiv A}\cos^2\theta_o \quad = \quad \underbrace{2\frac{(f_v/f_c)^2}{1 - (f_v/f_c)^2}}_{\equiv B}\sin^2\theta_o,$$

where

$$f_v \equiv f\cos\theta_o,$$

it follows that

$$\tan^2\theta_o = \frac{A + \frac{2h_o}{s}}{B - \frac{2h_o}{s}} \quad \text{and} \quad \frac{D}{h_o} = (2 + A\frac{s}{h_o})\tan\theta_o.$$

The equations above allow the construction of the parametric curves for fixed values of $\frac{h_o}{s}$ shown in Figure 4.16 which relate the MUF $\frac{f}{f_c}$ normalized by ionospheric critical

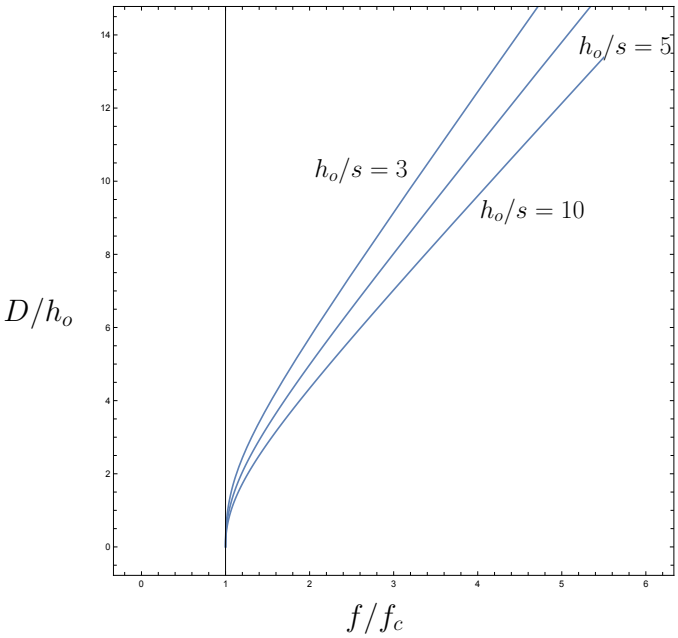


Figure 4.16: Normalized skip distance and MUF curves for a parabolic ionosphere with the indicated $\frac{h_o}{s}$ ratios, first developed by *Appleton and Benyon* (1947)

frequency f_c to the skip distance $\frac{D}{h_o}$ normalized by ionospheric base height h_o. For a fixed $\frac{h_o}{s}$ the solution curve is obtained after calculating A, B, θ_o, $\frac{f}{f_c}$, and $\frac{D}{h_o}$ (in that order) as a function of $\frac{f_v}{f_c}$. The curves simplify MUF and skip distance calculations under all ionospheres best fitted by a parabolic ionosphere.

Exercise 2: Assume $\frac{f_v}{f_c} = 0.94$ and $\frac{h_o}{s} = 5$, and calculate the corresponding $\frac{f}{f_c}$ and $\frac{D}{h_o}$ (using equations given above).

Exercise 3: Consider a parabolic ionosphere with $\frac{h_o}{s} = 5$, $h_o = 250$ km, and $f_c = 9$ MHz. What is the MUF to reach a receiver at a distance $D = 750$ km?

4.7 Effect of spherical geometry on sky-wave links

In a spherical stratified ionosphere the reflection condition is described by

$$m(z_r) = \sin\theta_o, \text{ where } m(z_r) = (1 + \frac{z_r}{a})\sqrt{1 - f_p^2(z_r)/f^2}.$$

Hence, for $z_r \ll a$,

$$(1 + \frac{2z_r}{a})(1 - \frac{f_p^2(z_r)}{f^2}) = \sin^2\theta_o \;\Rightarrow\; 1 - \frac{f_p^2(z_r)}{f^2} = \sin^2\theta_o(1 - \frac{2z_r}{a}),$$

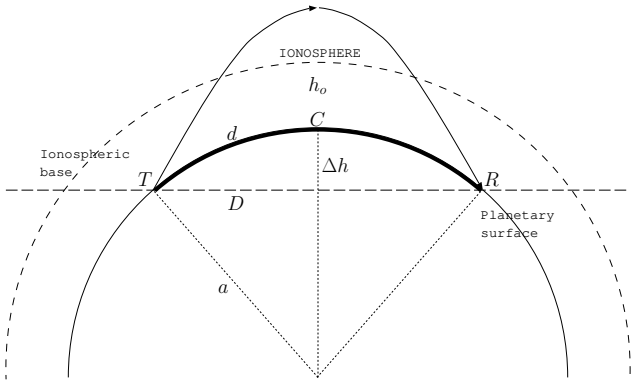


Figure 4.17: Given the arc length d between stations T and R, calculate length D and "planetary bulge" Δh using the formulae given in text. Then use Figure 4.16 with h_o replaced by $h_o + \Delta h$ in order to determine the normalized MUF $\frac{f}{f_c}$ appropriate for the single-hop link between T and R. This method is reasonably accurate for d up to ~3000 km.

leading to

$$f_p(z_r) = f \cos\theta_o \sqrt{1 + \frac{2z_r}{a} \tan^2\theta_o}.$$

This expression reduces to the Secant law form if and only if $\frac{2z_r}{a} \tan^2\theta_o \ll 1$ or

$$\tan\theta_o < \sqrt{\frac{a}{2z_r}}.$$

With $a = 6371.2$ km and $z_r = 300$ km (considering an F region reflection), for instance, the inequality above leads to $\theta_o < 73^o$. Hence, sky-wave calculations for links with $\theta_o > 70^o$ cannot be very accurate. Also, link distance formula $D = 2\tan\theta_o h'(f\cos\theta_o)$ from the previous section can only be trusted for $\theta_o < 70^o$ when

$$D < \sqrt{\frac{2a}{z_r}} h'(f\cos\theta_o).$$

In general, plane stratified results for MUF are suspect for $D > 500$ km or so.

However, in MUF calculations for single-hop links with $D > 500$ km we can still make use of plane stratified results as follows: Let d denote the shortest distance along the spherical surface of Earth between two points T and R, while D denote the straight line distance between the same two points as illustrated in Figure 4.17. We can make use of the curves in Figure 4.16 provided we take

$$D = 2a \sin(\frac{d}{2a}),$$

which can easily be shown to be correct using Figure 4.17, where a denotes the Earth radius as usual. Also, an increment

$$\Delta h = a(1 - \cos(\frac{d}{2a}))$$

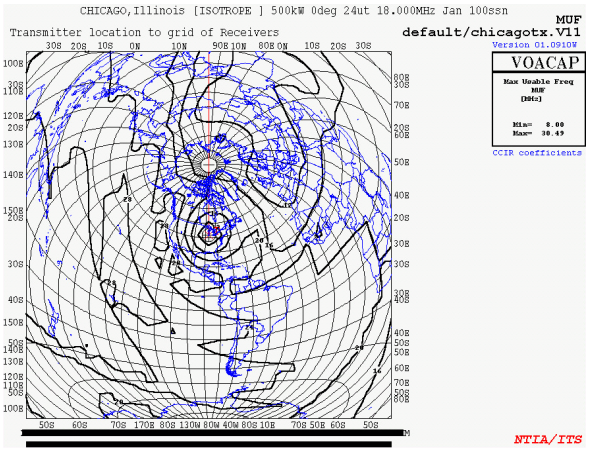


Figure 4.18: MUF map generated by VOACAP model for a transmitter located in Chicago. Transmission time is 19:00 LT and the standard ionosphere for the month of January is assumed. The MUFs towards the sunlit Pacific region are high while MUFs towards the nighttime European region are low.

shown in Figure 4.17 needs to be added to ionospheric base parameter h_o obtained from an ionogram operated at the central point C between stations T and R. This method of MUF calculation is valid and reasonably accurate as long as the curved part of the ray trajectory within the ionosphere has a horizontal extent no larger than ~500 km.

Exercise 1: Verify the formulae above given for D and Δh defined in Figure 4.17.

Sky-wave link calculations for even larger distances and multi-hop cases are best handled with numerical ray tracing models. Such models are based on empirical models of the ionosphere (i.e., $f_p(z)$) based on large amounts of ionospheric data inputs (global ionosonde networks as well as other remote and in-situ measurement techniques of $f_p(z)$). For instance, the VOACAP model (a PC freeware) produces global MUF maps to reach out from a given transmitter site to receivers located anywhere around the globe based on ray tracing calculations. The maps take into account multi-hop sky-wave paths some of which traverse the nighttime ionosphere.

Figure 4.18 shows a VOACAP MUF map for a Chicago based transmitter operating at 19:00 local time. To east of the transmitter is a weak nighttime ionosphere with relatively low f_c values; as a consequence MUFs to reach Europe are quite low and some portions of Eastern Europe and Western Asia may in fact be difficult to contact. By contrast MUFs into the sunlit Pacific region as well as into South America are quite large. The map above and the discussion of ionospheric absorption below show why a single transmitting frequency cannot be optimal for global radio coverage.

4.8 Earth-ionosphere waveguide and ELF propagation

Near the ionospheric base, where $\frac{\nu}{2\pi}$ is in 0.3-30 MHz range, depending on the altitude of ionospheric base (which varies from day to night), ionospheric conductivity (3.49)

$$\sigma = \frac{Ne^2}{m}\frac{\nu - j\omega}{\nu^2+\omega^2} \approx \frac{Ne^2}{m\nu} = \epsilon_o\frac{\omega_p^2}{\nu} = 2\pi\epsilon_o\frac{f_p^2}{\nu/2\pi} \tag{4.26}$$

for radiowaves at VLF and below. For ELF (30 Hz to 3 kHz) this expression is valid up till nearly 100 km altitude. Therefore, ELF waves propagating between the ground and ionospheric base are bounded both from above and below by Ohmic conductors. This situation is analogous to propagation within a parallel plate waveguide since the free-space wavelength at ELF (e.g., 300 km for $f = 1$ kHz) is larger than the ground-ionospheric base distance z_b. Therefore, the main effect of the ionosphere for ground based ELF emissions is to form the upper boundary of an effective Earth-ionosphere waveguide which will be examined below.

Note that the Earth-ionosphere waveguide concept is not valid at frequencies above VLF: First, above VLF σ has a non-negligible and increasing inductive component with increasing f. Second, at larger frequencies f the transition into the ionosphere occurs slowly in comparison to the shorter wavelengths of the radiowaves; hence, ionospheric reflections at higher frequencies — if they occur at all — happen as a consequence of slow refraction processes. Finally, at higher frequencies wavelength $\lambda_o = \frac{c}{f} \ll z_b$ and a guided wave approach is unnecessary.

The properties of Earth-ionosphere waveguide for ELF waves can be examined with varying degrees of detail. The simplest approach is to model it as a parallel-plate waveguide with perfectly conducting boundaries. In this idealization ELF signals propagate within the waveguide as TEM waves (see below) with no attenuation while TE_m and TM_m modes, $m \geq 1$, are cut-off because of the relatively small value of z_b compared to ELF wavelengths (100 to 1000 km). In practice, propagation occurs with some amount of attenuation since the boundaries of the waveguide are actually imperfect conductors. As we will see below, small energy leakage from the Earth-ionosphere waveguide has useful applications.

Recall that in an air filled ideal parallel-plate waveguide with a plate separation z_b, the boundary constraints (mainly, zero tangential $\mathbf{E}$ at the boundaries) dictate that TM_m fields

$$\begin{aligned}\mathbf{E}_m &= \frac{-\hat{x}k_z + \hat{z}k_x}{k_o}\frac{E_o}{2}e^{-j(k_x x + k_z z)} + \frac{\hat{x}k_z + \hat{z}k_x}{k_o}\frac{E_o}{2}e^{-j(k_x x - k_z z)} \\ &= j\hat{x}\frac{k_z}{k_o}E_o\sin(k_z z)e^{-jk_x x} + \hat{z}\frac{k_x}{k_o}E_o\cos(k_z z)e^{-jk_x x}\end{aligned}$$

can propagate if and only if

$$k_z z_b = m\pi \;\Rightarrow\; k_z = \frac{m\pi}{z_b} \;\Rightarrow k_x^2 = k_o^2 - k_z^2 = \frac{\omega^2}{c^2} - (\frac{m\pi}{z_b})^2 > 0 \;\Rightarrow\; \omega > \frac{mc\pi}{z_b} \;\Rightarrow\; f > \frac{mc}{2z_b}.$$

For an average value of $z_b = 70$ km, the cut-off frequencies of TE_m modes are therefore

$$f_m = \frac{m3 \times 10^8}{2 \times 70 \times 10^3} \approx 2143m\,\mathrm{Hz} \approx 2.1m\,\mathrm{kHz}.$$

Hence, ELF waves with $f < 2\,\mathrm{kHz}$ can only propagate as TM_0 =TEM modes; the only other propagating modes are TE_1 and TM_1 for higher ELF waves in $2.1 < f < 3$ kHz range (depending on the value of z_b). Since for the principle ELF mode within the waveguide (i.e., TEM) $k_x = k_o = \frac{\omega}{c}$, ELF waves propagate without dispersion at the speed of light.

In a lossless parallel-plate waveguide with a plate separation z_m in z direction, the TEM mode fields

$$\mathbf{E}_o = \hat{z}E_o e^{-jk_x x} \quad \text{and} \quad \mathbf{H}_o = -\hat{y}\frac{E_o}{\eta}e^{-jk_x x}$$

have uniform and plane wave characteristics between the plates while the tangential $\mathbf{H}_o$ at the boundaries are matched with surface current densities flowing in the propagation direction with magnitudes $\frac{E_o}{\eta}$. In a lossy waveguide with imperfect conducting walls the currents penetrate the conductor by a skin depth δ with an average current density of $\frac{E_o}{\eta\delta}$. The corresponding energy loss rate per unit volume is $\frac{1}{2\sigma}|\mathbf{J}|^2 = \frac{1}{2\sigma}|\frac{E_o}{\eta\delta}|^2$, while the energy loss rate per unit area of the wall is $\delta\frac{1}{2\sigma}|\mathbf{J}|^2 = \frac{1}{\eta\sigma\delta}\frac{|E_o|^2}{2\eta} = \frac{1}{\eta\sigma\delta}S$, where $S \equiv \frac{|E_o|^2}{2\eta}$ is the flux of the guided wave and σ is Ohmic wall conductivity. As a result of wall losses (on both walls) transmitted power per unit length Sz_b at $x = 0$ drops to $Sz_b - 2\frac{1}{\eta\sigma\delta}Sx = Sz_b(1 - \frac{2}{\eta\sigma\delta z_b}x) = Sz_b e^{-\frac{2}{\eta\sigma\delta z_b}x}$ after an infinitesimal distance x. Hence, the guided field attenuates at a rate

$$A = 10\log_{10}(e^{|-\frac{2}{\eta\sigma\delta z_b}|1\,\mathrm{km}}) = \frac{10000 \times 2}{\eta\sigma\delta z_b}\log_{10} e \approx 4343\frac{2}{\eta\sigma\delta z_b}\frac{\mathrm{dB}}{\mathrm{km}},$$

or,

$$A \approx 4343\frac{1}{\eta z_b}(\frac{1}{\sigma_g\delta_g} + \frac{1}{\sigma_i\delta_i})\frac{\mathrm{dB}}{\mathrm{km}}$$

in a waveguide with unequal bottom and upper wall conductivities σ_g and σ_i. In high conductivity materials where $\frac{\sigma}{\omega\epsilon} \gg 1$ the skin depth $\delta = \frac{1}{\sqrt{\pi f\sigma\mu}}$ (see homework) and hence $\sigma\delta = \sqrt{\frac{\sigma}{\pi f\mu}}$. Therefore, with $\eta = \eta_o$, $\mu = \mu_o$, and assuming $\sigma_g \gg \sigma_i$, the attenuation rate

$$A \approx 4343\frac{\sqrt{\pi f\mu_o}}{\eta_o z_b}(\frac{1}{\sqrt{\sigma_g}} + \frac{1}{\sqrt{\sigma_i}})\frac{\mathrm{dB}}{\mathrm{km}} \approx 0.0229\frac{\sqrt{f/\sigma_i}}{z_b}\frac{\mathrm{dB}}{\mathrm{km}}. \tag{4.27}$$

Indeed, typical ground conductivities $\sigma_g \sim 10^{-3} - 10^{-2}\,\frac{\mathrm{S}}{\mathrm{m}}$ are much larger than ionospheric conductivity σ_i at ELF. For instance, for $\frac{\nu}{2\pi} = 3$ MHz and $f_p = 0.06$ MHz, (4.26) implies $\sigma_i \sim 10^{-6}\,\frac{\mathrm{S}}{\mathrm{m}}$. In that case, according to (4.27) attenuation rates of guided ELF waves are about 0.01 and $0.003\,\frac{\mathrm{dB}}{\mathrm{km}}$ for $f = 1\,\mathrm{kHz}$ and 100 Hz signals with $z_b = 70\,\mathrm{km}$. For a 100 Hz signal the attenuation is only 30 dB over a 10000 km distance. These calculations show that extremely low frequency signals can be used to reach nearly the

antipode point[11] via the Earth-ionosphere waveguide. However, extremely low carrier frequencies imply extremely low signal bandwidths and thus extremely slow information rates. Hence, applications of ELF are mainly limited to communicating with submerged submarines which are otherwise very difficult to contact because of high attenuation rates of TEM waves in sea water (except in ELF band).

Another application of the Earth-ionosphere waveguide concerns timing signals broadcast by WWVB from Fort Collins, Colorado, with a carrier frequency of 60 kHz in LF band. In LF band the waveguide has multiple modes of propagation and attenuation rates are much higher than in ELF. Nevertheless, the range of WWVB covers the entire continental U.S. and some more. So-called "atomic wristwatches" in the market rely on WWVB signals to make time corrections to maintain an accuracy of better than a second (WWVB signals include a "beginning of minute" marker each minute, and also contain a special "phase signature" to identify itself unambiguously[12]).

[11] Since all propagation paths from a transmitting antenna converge at the antipode ideally the transmission gain to the antipode is independent of the great circle distance of about 20000 km.In other words the usual r^{-2} dependence of Friis formula cancels out at the antipode because of the focusing effect signals arriving from all directions. This makes ~ 60 dB attenuation at $f = 100$ Hz to reach the antipode tolerable.

[12] See http://www.boulder.nist.gov/timefreq/stations/wwvb.html

5 Radiowave scattering and radar remote sensing

Radar is an acronym for *ra*dio *d*etection *a*nd *r*anging. The term describes *active* RF systems which operate by reflecting or scattering radiowaves from targets of interest that stand out electrically from the background in which they are immersed — airplanes, meteor trails, birds and insects, rain drops, CAT[1], free electrons. The term *backscatter* radar (or *monostatic* radar) is used to refer to systems which use the same antenna for transmission and reception. By contrast, radars which use two or more antennas located at different sites are known as *forward* scatter or *bistatic* radars. In this chapter we will examine the basics of radiowave scattering and focus on backscatter radars used in probing the atmosphere and the ionosphere.

5.1 Radiowave scattering and backscatter radars

Consider propagation in a medium where $\sigma = 0$, $\mu = \mu_o$, and ϵ is expressed as

$$\epsilon = \epsilon_b + \delta\epsilon$$

in terms of, for instance, a slowly varying ϵ_b and a small remainder $\delta\epsilon = \epsilon - \epsilon_b$ which could include some fast variations. The fields in the described medium can be expressed as

$$\mathbf{E} = \mathbf{E}_b + \delta\mathbf{E} \quad \text{and} \quad \mathbf{H} = \mathbf{H}_b + \delta\mathbf{H},$$

where

$$\nabla \times \mathbf{E}_b \equiv -j\omega\mu_o\mathbf{H}_b \quad \text{and} \quad \nabla \times \mathbf{H}_b \equiv j\omega\epsilon_b\mathbf{E}_b$$

and

$$\nabla \times \mathbf{E} = -j\omega\mu_o\mathbf{H} \quad \text{and} \quad \nabla \times \mathbf{H} = j\omega\epsilon\mathbf{E},$$

which are of course source-free Faraday's and Ampere's laws for media with permittivities ϵ_b and $\epsilon = \epsilon_b + \delta\epsilon$, respectively. Clearly, if $\mathbf{E}_b$ and $\mathbf{H}_b$ were known (or easily calculated because ϵ_b is slow varying or even a constant), $\mathbf{E}$ and $\mathbf{H}$ could be obtained by adding $\delta\mathbf{E}$ and $\delta\mathbf{H}$ to $\mathbf{E}_b$ and $\mathbf{H}_b$, respectively.

The field "corrections" $\delta\mathbf{E}$ and $\delta\mathbf{H}$ needed because of non-zero $\delta\epsilon$ are regarded as fields "scattered" by $\delta\epsilon$ in response to "incident fields" $\mathbf{E}_b$ and $\mathbf{H}_b$. The corrections or *scattered fields* $\delta\mathbf{E}$ and $\delta\mathbf{H}$ can be identified by "solving" the *difference* of the equations given above, namely,

$$\nabla \times \delta\mathbf{E} = -j\omega\mu_o\delta\mathbf{H} \quad \text{and} \quad \nabla \times \delta\mathbf{H} = j\omega\delta\epsilon\mathbf{E} + j\omega\epsilon_b\delta\mathbf{E},$$

[1] Clear air turbulence

which are of course Faraday's and Ampere's laws for $\delta\mathbf{E}$ and $\delta\mathbf{H}$ excited by an equivalent *source current* $j\omega\delta\epsilon\mathbf{E}$. Note that the equations are non-linear — a doubling of $\delta\epsilon$ does not lead to a doubling of $\delta\mathbf{E}$ and $\delta\mathbf{H}$ — because of the presence of $\delta\mathbf{E}$ within $\mathbf{E} = \mathbf{E}_b + \delta\mathbf{E}$. However, if the permittivity irregularities were small, i.e., if $|\delta\epsilon| \ll \epsilon_b$, then $\delta\mathbf{E}$ and $\delta\mathbf{H}$ would also be small, in which case $j\omega\delta\epsilon\mathbf{E}_b$ could be used in place of $j\omega\delta\epsilon\mathbf{E}$. Exchanging $j\omega\delta\epsilon\mathbf{E}$ above with $j\omega\delta\epsilon\mathbf{E}_b$ — a procedure known as first-order *Born approximation* — leads to a set of linearized equations

$$\nabla \times \delta\mathbf{E} = -j\omega\mu_o\delta\mathbf{H} \quad \text{and} \quad \nabla \times \delta\mathbf{H} = j\omega\delta\epsilon\mathbf{E}_b + j\omega\epsilon_b\delta\mathbf{E}$$

for the scattered fields.

The linearized equations above can be used to solve for $\delta\mathbf{E}$ and $\delta\mathbf{H}$ when $\mathbf{E}_b$ and $\delta\epsilon$ are known, or for $\delta\epsilon$ when $\mathbf{E}_b$ and $\delta\mathbf{E}$ are known. We refer to the latter problem of inferring $\delta\epsilon$ from $\mathbf{E}_b$ and $\delta\mathbf{E}$ as *inverse scattering* and focus here on the former problem of modeling $\delta\mathbf{E}$ (and $\delta\mathbf{H}$) in terms of $\delta\epsilon$ and $\mathbf{E}_b$ under the Born approximation. The solution of this *forward scattering* problem — inferring $\delta\mathbf{E}$ from $\delta\epsilon$ and $\mathbf{E}_b$ — is the key for inferring the statistics of $\delta\epsilon$ from the statistics of $\delta\mathbf{E}$, which constitutes a major objective of the atmospheric/ionospheric backscatter radars described later in this chapter.

5.1.1 Backscattered field in Born approximation

In a medium where the background permittivity ϵ_b is some constant ϵ, the Born approximation solution of forward scattering problem can be obtained with the help of radiation equations introduced in Chapter 4. In particular, the retarded potential $\delta\mathbf{A}(\mathbf{r})$ due to equivalent radiating current $\delta\mathbf{J}(\mathbf{r}) \equiv j\omega\delta\epsilon(\mathbf{r})\mathbf{E}_b(\mathbf{r})$ can be written as

$$\begin{aligned}\delta\mathbf{A}(\mathbf{r}) &= \int_{V'} \mu_o\{j\omega\delta\epsilon(\mathbf{r})\mathbf{E}_b(\mathbf{r})\}\frac{e^{-jk|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|}dV' \\ &= j\omega\mu_o\int_{V'} \delta\epsilon(\mathbf{r}')\mathbf{E}_b(\mathbf{r}')\frac{e^{-jk|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|}dV',\end{aligned}$$

where $k \equiv \omega\sqrt{\mu_o\epsilon}$ is the wavenumber of the scattered field in the "background" propagation medium with permittivity ϵ. For $|\mathbf{r}| \ll |\mathbf{r}'|$ we can use the paraxial approximation $|\mathbf{r}-\mathbf{r}'| \approx r' - \hat{r}'\cdot\mathbf{r}$ to obtain

$$\delta\mathbf{A}(\mathbf{r}) = j\omega\mu_o\int_{V'} \delta\epsilon(\mathbf{r}')\mathbf{E}_b(\mathbf{r}')e^{jk\hat{r}'\cdot\mathbf{r}}\frac{e^{-jkr'}}{4\pi r'}dV',$$

which is valid for small $\mathbf{r}$ — see Figure 5.1 — and it is in the form of a plane-wave superposition. Thus the scattered fields $\delta\mathbf{B}(\mathbf{r}) = \nabla \times \delta\mathbf{A}(\mathbf{r})$ and $\delta\mathbf{E}(\mathbf{r}) = \frac{\nabla\times\delta\mathbf{B}(\mathbf{r})}{j\omega\mu_o\epsilon}$ can be readily evaluated near $\mathbf{r} = 0$ using the result above. We find, in particular,

$$\begin{aligned}\delta\mathbf{E}(\mathbf{r}) &= \int_{V'} \frac{\delta\epsilon(\mathbf{r}')}{\epsilon}\nabla \times \nabla \times \mathbf{E}_b(\mathbf{r}')e^{jk\hat{r}'\cdot\mathbf{r}}\frac{e^{-jkr'}}{4\pi r'}dV' \\ &= -\mu_o\omega^2\int_{V'} \delta\epsilon(\mathbf{r}')\hat{r}' \times \hat{r}' \times \mathbf{E}_b(\mathbf{r}')e^{jk\hat{r}'\cdot\mathbf{r}}\frac{e^{-jkr'}}{4\pi r'}dV'\end{aligned}$$

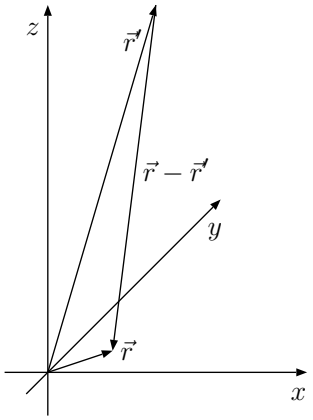


Figure 5.1: Position $\mathbf{r}$ is in the far field of position $\mathbf{r}'$, where a *polarization current* $\sim j\omega\delta\epsilon\mathbf{E}_b(\mathbf{r}')$ radiates a scattered field $\delta\mathbf{E}$ in response to an incident field $\mathbf{E}_b$.

after replacing ∇ with $jk\hat{r}'$ — because of the plane-wave form of the integrand $\propto e^{jk\hat{r}'\cdot\mathbf{r}}$ — and $\frac{k^2}{\epsilon}$ with $\mu_o\omega^2$. An antenna located at the origin of our coordinate system will then respond to a scattered field

$$\delta\mathbf{E}(0) = -\mu_o\omega^2\int_{V'}\delta\epsilon(\mathbf{r}')\hat{r}'\times\hat{r}'\times\mathbf{E}_b(\mathbf{r}')\frac{e^{-jkr'}}{4\pi r'}dV',$$

which can be termed the *backscattered field* if $\mathbf{E}_b(\mathbf{r})$ is itself the radiation field of the very same antenna. But if $\mathbf{E}_b(\mathbf{r})$ is a TEM wave radiated away from the origin then $\hat{r}\times\hat{r}\times\mathbf{E}_b(\mathbf{r}) = -\mathbf{E}_b(\mathbf{r})$ and consequently backscattered fields can be expressed as

$$\delta\mathbf{E} = \mu_o\omega^2\int\delta\epsilon(\mathbf{r}')\mathbf{E}_b(\mathbf{r}')\frac{e^{-jkr'}}{4\pi r'}d\mathbf{r}' \tag{5.1}$$

after a slight change of notation.

Although the backscattered field (5.1) was derived assuming $\sigma = 0$, it can be easily modified with the replacements

$$\epsilon \to \epsilon - j\frac{\sigma}{\omega} \quad \Rightarrow \quad \delta\epsilon \to \delta\epsilon - j\frac{\delta\sigma}{\omega}$$

familiar from Chapter 1 to account for scattering from a conducting medium. For instance in a medium with $\delta\epsilon = 0$ but $\delta\sigma \neq 0$ the backscattered field can be written as

$$\delta\mathbf{E} = -j\mu_o\omega\int\delta\sigma(\mathbf{r}')\mathbf{E}_b(\mathbf{r}')\frac{e^{-jkr'}}{4\pi r'}d\mathbf{r}';$$

furthermore, within a plasma, where

$$\sigma = \frac{Ne^2}{mj\omega} \quad \Rightarrow \quad \delta\sigma = \frac{\delta Ne^2}{mj\omega},$$

we obtain

$$\delta\mathbf{E} = -r_e \int \delta N(\mathbf{r}')\mathbf{E}_b(\mathbf{r}')\frac{e^{-jkr'}}{r'}d\mathbf{r}', \tag{5.2}$$

where

$$r_e \equiv \frac{e^2\mu_o}{4\pi m} = \frac{e^2}{4\pi\epsilon_o mc^2} \approx 2.8\times 10^{-15}\ \text{m} \tag{5.3}$$

denotes the *classical electron radius* encountered in Chapter 2. In (5.2) the perturbation $\delta N(\mathbf{r})$ corresponds to the fluctuating part of ionospheric electron density N.

5.1.2 Thomson scattering, backscatter RCS, and radar equation

Let us first apply the backscattered field equation (5.2) to the case of a limiting "single electron ionosphere" where

$$\delta N(\mathbf{r}) = \delta(\mathbf{r}-\mathbf{r}_p)$$

and the background electron density $N_b = 0$, meaning that k and $\mathbf{E}_b(\mathbf{r})$ are to be calculated using free-space parameters. In that scenario

$$\delta\mathbf{E} = -r_e \int \delta(\mathbf{r}'-\mathbf{r}_p)\mathbf{E}_b(\mathbf{r}')\frac{e^{-jkr'}}{r'}d\mathbf{r}' = -r_e\mathbf{E}_b(\mathbf{r}_p)\frac{e^{-jkr_p}}{r_p}, \tag{5.4}$$

clearly a co-polarized field echo from the electron at $\mathbf{r} = \mathbf{r}_p$ back to the antenna that radiates the incident field $\mathbf{E}_b(\mathbf{r})$. We will refer to distance $r_p \equiv |\mathbf{r}_p|$ as "radar range" of the electron causing the backscatter $\delta\mathbf{E}$.

In general, the backscattering *radar cross-section* (RCS) of a radar target is *defined as*

$$\sigma \equiv \lim_{r\to\infty} 4\pi r^2 \frac{|\delta\mathbf{E}|^2}{|\mathbf{E}_b|^2}, \tag{5.5}$$

where r denotes the radar range (or the physical distance) of the target and $\delta\mathbf{E}$ the field that it scatters back to the radar antenna in response to an incident field $\mathbf{E}_b$ (arriving from the same antenna). Clearly, (5.5) and (5.4) together imply that the backscatter RCS of a free electron must be

$$\sigma = 4\pi r_e^2 \equiv \sigma_e. \tag{5.6}$$

Now, in terms of the RCS (5.5), the available power in $\delta\mathbf{E}$ backscattered from any discrete target (like a single electron) at a far-field location $\mathbf{r}$ is obtained as

$$\begin{aligned} P_a &= \overbrace{\frac{|\delta\mathbf{E}|^2}{2\eta_o}}^{\text{Backscattered Poynting}} \underbrace{A(\hat{r})}_{\frac{\lambda^2}{4\pi}G(\hat{r}')} = \frac{\sigma}{4\pi r^2}\underbrace{\frac{|\mathbf{E}_b(\mathbf{r})|^2}{2\eta_o}}_{\text{Incident Poynting } \frac{P_t}{4\pi r^2}G(\hat{r})}\frac{\lambda^2}{4\pi}G(\hat{r}) \\ &= P_t\frac{\lambda^2 G^2(\hat{r})}{(4\pi)^3 r^4}\sigma, \end{aligned}$$

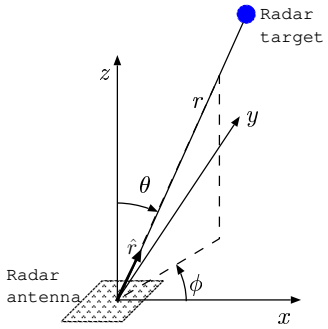


Figure 5.2: Radar geometry showing a target at a radar range r in direction $\hat{r}$.

Target	σ	Explanation/Condition
Free electron	$4\pi r_e^2$	$r_e = \frac{e^2}{4\pi\epsilon_o mc^2} \approx 2.8 \times 10^{-15}$ m
Small dielectric sphere	$4\pi a^2 (ka)^4 (\frac{\epsilon-\epsilon_o}{\epsilon+2\epsilon_o})^2$	Radius $a \ll \lambda = \frac{2\pi}{k}$
Large dielectric sphere	$\pi a^2 (\frac{\sqrt{\epsilon}-\sqrt{\epsilon_o}}{\sqrt{\epsilon}+\sqrt{\epsilon_o}})^2$	Radius $a \gg \lambda = \frac{2\pi}{k}$
PEC plate	$\frac{4\pi}{\lambda^2} A_p^2$	Plate area $A_p \gg \lambda^2$

Table 5.1: Backscatter RCS of simple radar targets immersed in free space.

where P_t represents the transmitted power of a backscatter radar system with an antenna[2] gain $G(\hat{r})$.

The available power formula

$$P_a = P_t \frac{\lambda^2 G^2(\hat{r})}{(4\pi)^3 r^4} \sigma \tag{5.7}$$

for any discrete far-field target with a backscatter RCS σ — see Figure 5.2 — is known as *radar equation.* It is useful in radar power budget and SNR (signal-to-noise ratio) calculations when the target RCS is known. Alternatively, it can be used to determine the RCS of unknown targets by experimental means. The backscatter RCS of some simple objects are listed in Table 5.1. The list entries show that RCS is in general different from the physical cross-sectional area of the radar target[3].

Radiowave scattering by free-electrons — known as *Thomson scattering* — is a simple consequence of oscillatory electron motion caused by an oscillating incident field $\mathbf{E}_b$. The oscillating electron radiates like a Hertzian dipole, generating the scattered field $\delta\mathbf{E}$ with a gain pattern that maximizes at a value of 1.5 in the propagation direction of incident $\mathbf{E}_b$. The total power radiated out and carried by $\delta\mathbf{E}$ is of course at the expense of power

[2]Backscatter radars use the same antenna for transmission and reception by maintaining a pulsed operation where the antenna is switched periodically between a transmitter and receiver circuit. Details of pulsed operations will be discussed later on.

[3]For most targets the RCS σ is also *aspect sensitive*; that is, σ depends on the orientation of the target with respect to the radar. For instance the RCS of a metal plate is less than that given in Figure 5.2 if it is not oriented normal to the radar viewing direction.

taken out from the incident wave $\mathbf{E}_b$. The power loss from $\mathbf{E}_b$ is not

$$4\pi r^2 \times \frac{|\delta\mathbf{E}|^2}{2\eta_o}$$

as it would have been if the electron were an isotropic radiator, instead it is

$$\int d\Omega r^2 \frac{|\delta\mathbf{E}|^2 \sin^2\theta}{2\eta_o} = r^2 \frac{|\delta\mathbf{E}|^2}{2\eta_o} \underbrace{\int d\Omega \sin^2\theta}_{4\pi/1.5} = 4\pi r^2 \frac{|\delta\mathbf{E}|^2}{2\eta_o \times 1.5} = \frac{|\mathbf{E}_b|^2}{2\eta_o}\frac{\sigma_e}{1.5} = \frac{|\mathbf{E}_b|^2}{2\eta_o}\overbrace{\frac{2}{3}\sigma_e}^{\equiv \sigma_t}$$

when the Hertzian gain pattern $\propto \sin^2\theta$ of the scattered field is taken into account.

The parameter

$$\sigma_t = \frac{2}{3}\sigma_e = \frac{8\pi}{3} r_e^2 \approx 6.6 \times 10^{-29}\ \mathrm{m}^2 \tag{5.8}$$

above that converts the incident Poynting flux $\frac{|\mathbf{E}_b|^2}{2\eta_o}$ into the *total scattered power* is known as *total cross-section* of free electron. The total cross-section σ_t of any target equals the total *scattered power per unit incident flux*, which is in general different from σ. The latter, *backscatter* RCS, can be interpreted as 4π times the *power scattered per unit solid angle and per unit incident flux* in the direction of the radar antenna. This is illustrated in Figure 5.3 where the scattered power per unit solid angle towards the radar antenna is $\frac{|\mathbf{E}_b|^2}{2\eta_o}\frac{\sigma}{4\pi}$, and $\frac{A(\hat{r})}{r^2}$ is the solid angle of the antenna aperture seen from the target. Their product is an alternate form of the radar equation shown in the figure.

Exercise 1: Find the expression for the total cross-section σ_t of an infinitely conducting square plate with dimensions $L \times L$ oriented perpendicular to the direction of an illuminating source. Assume $L \gg \lambda$. Hint: The plate will radiate like a uniformly illuminated aperture antenna.

According to what we have seen above, a free electron illuminated by 1 W/m^2 of Poynting flux scatters a total of 6.6×10^{-29} W average power, a truly negligible amount which justifies the use of first-order Born approximation in ionospheric radar problems without worrying about the extinction of incident field $\mathbf{E}_b(\mathbf{r})$ or secondary scattering of scattered fields.

5.2 Atmospheric backscatter fundamentals

5.2.1 Volumetric backscatter from free-electrons

In this section we will generalize what we learned about single-electron scattering to describe the operation principles of ionospheric backscatter radars. More specifically, the fundamental concept of Bragg scattering from electron density waves of a certain spatial scale will be developed.

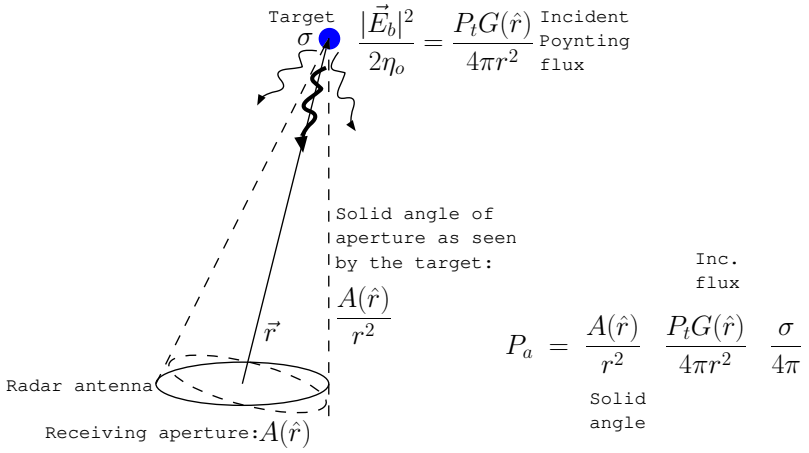


Figure 5.3: Target scatters a total power of $\frac{|\mathbf{E}_b|^2}{2\eta_o}\sigma_t$ W in all directions and a power density of $\frac{|\mathbf{E}_b|^2}{2\eta_o}\frac{\sigma}{4\pi}\frac{\text{W}}{\text{ster.}}$ back towards the radar antenna. The available power at the radar is $\frac{|\mathbf{E}_b|^2}{2\eta_o}\frac{\sigma}{4\pi} \times \frac{A(\hat{r})}{r^2}$ where $\frac{A(\hat{r})}{r^2}$ steradians is the solid angle angle spanned by the receiving aperture. Solid angle $\propto A(\hat{r}) \propto G(\hat{r})$ depends on direction $\hat{r}$ of the target.

We will start by first modeling the scattered field from a small ionospheric volume V containing a total of P free-electrons (and other particles with negligible cross-sections which we will ignore). The model is based on

$$\delta\mathbf{E} = -r_e \mathbf{E}_b(\mathbf{r}_p)\frac{e^{-jk_o r_p}}{r_p}, \tag{5.9}$$

the backscattered field of a single electron at a frequency ω_o and wavenumber $k_o = \omega_o/c$, where

$$\mathbf{E}_b(\mathbf{r}_p) = j\eta_o I_o k_o \mathbf{l}(\hat{r}_p)\frac{e^{-jk_o r_p}}{4\pi r_p} \tag{5.10}$$

is the field incident on the electron in terms of the input current I_o and effective length $\mathbf{l}(\hat{r}_p)$ of the radar antenna illuminating the electron.

First, we note that vectors $\delta\mathbf{E}$ and $\mathbf{E}_b$ are parallel[4], and, therefore, we have the freedom to relate them to one another using their scalar amplitudes δE and E_b — we will do that. Second, by choosing the volume V small enough (more on how small later on) we can justify the use of an incident field model

$$E_b(\mathbf{r}_p) \equiv E_i e^{-jk_o r_p}, \tag{5.11}$$

where we are dropping all the distinctions between $\mathbf{E}_b(\mathbf{r}_p)$'s at different electron locations $\mathbf{r}_p$ (within V) except for the propagation phase delay $k_o r_p$. And finally, we will replace

[4] An exception when $\delta\mathbf{E}$ Faraday rotates with respect to $\mathbf{E}_b$ due to magneto-ionic effects discussed in Chapter 6 can be handled as an additional propagation effect in a separate calculation.

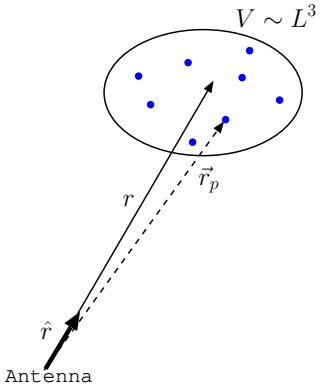


Figure 5.4: A sketch of the relative geometry of a radar antenna and a small scattering volume $V \sim L^3$ such that it is assumed $r > L^2/\lambda_o/2$ to satisfy the Rayleigh criterion. The volume contains P free-electrons with an average density of $N = P/V$.

all the occurrences of r_p in the equations with r — the distance of the center of volume V with respect to the radar antenna as depicted in Figure 5.4 — except in the phase term.

With those assumptions the scattered field amplitude from a single electron anywhere within volume V can be written as

$$\delta E = -r_e E_i \frac{e^{-j2k_o r_p}}{r}, \tag{5.12}$$

and the total scattered field from the volume is a simple superposition

$$E_s = -\frac{r_e}{r} E_i \sum_{p=1}^{P} e^{-j2k_o r_p}. \tag{5.13}$$

All the approximations above inherent in (5.13) are identical in nature to simplifying assumptions commonly employed in antenna array calculations in the far-field. In array theory one more approximation is typically used, namely the paraxial approximation of replacing $e^{-j2k_o r_p}$ with some $e^{-j2k_o \hat{r}\cdot \mathbf{r}_p}$, where $-\hat{r}$, in our case, is the direction of the radar antenna seen from volume V (and, conversely, $\hat{r}$ the direction of volume V seen from the antenna). With that replacement, and using

$$\mathbf{k} \equiv -2k_o \hat{r}, \tag{5.14}$$

which will be referred to as *Bragg wavevecto*r, we re-write E_s, in the paraxial approximation, as

$$E_s = -\frac{r_e}{r} E_i \sum_{p=1}^{P} e^{j\mathbf{k}\cdot \mathbf{r}_p}. \tag{5.15}$$

We can now establish the permissible size of volume V in terms of Rayleigh criterion discussed in the last chapter. Basically, to justify the paraxial approximation we need the distance r from volume $V \equiv L^3$ to the radar antenna to exceed the Rayleigh distance, as in

$$r > R_{ray} = \frac{2L^2}{\lambda_o} \quad \Rightarrow \quad L < \sqrt{\frac{r\lambda_o}{2}}.$$

Thus, our scattered field model (5.15) is valid if the characteristic dimension L of volume V is, for instance, less than about 500 m when we take $r = 100$ km and $\lambda_o/2 = 3$ m, pertinent for a 50 MHz radar.

Our next objective is to obtain an equivalent backscatter RCS for volume V. We can achieve that by relating $\langle |E_s|^2 \rangle$ to $|E_i|^2$ via (5.15). Alternatively, we can use (5.15) to relate the *auto-correlation function* (ACF) of the scattered field to $|E_i|^2$, which can also achieve the same aim as a special case, but also provide a model for the backscattered signal spectrum needed later on.

The ACF of scattered field E_s is defined as the expected value $\langle E_s^* E_s^{'} \rangle$, where $E_s^{'}$ refers to the scattered field detected some time delay τ after detecting the field E_s. Thus, using (5.15) we can write

$$\langle E_s^* E_s^{'} \rangle = \frac{r_e^2 |E_i|^2}{r^2} \langle \sum_{p=1}^{P} e^{-j\mathbf{k}\cdot\mathbf{r}_p} \sum_{q=1}^{P} e^{j\mathbf{k}\cdot\mathbf{r}_q^{'}} \rangle, \tag{5.16}$$

where $\mathbf{r}_q^{'}$ refers to the new position of particle q a time τ after it has moved from its prior position at $\mathbf{r}_q$. The above ACF equation simplifies if we can assume that individual particles follow independent trajectories so that

$$\langle e^{-j\mathbf{k}\cdot\mathbf{r}_p} e^{j\mathbf{k}\cdot\mathbf{r}_q^{'}} \rangle = \langle e^{-j\mathbf{k}\cdot\mathbf{r}_p} \rangle \langle e^{j\mathbf{k}\cdot\mathbf{r}_q^{'}} \rangle = 0,$$

unless $q = p$. With that assumption, we re-write (5.16) as

$$\langle E_s^* E_s^{'} \rangle = \frac{r_e^2 |E_i|^2}{r^2} \sum_{p=1}^{P} \langle e^{j\mathbf{k}\cdot(\mathbf{r}_p(t+\tau) - \mathbf{r}_p(t))} \rangle = \frac{r_e^2 |E_i|^2}{r^2} P \langle e^{j\mathbf{k}\cdot(\mathbf{r}_p(t+\tau) - \mathbf{r}_p(t))} \rangle, \tag{5.17}$$

where we are now using $\mathbf{r}_p(t+\tau)$ and $\mathbf{r}_p(t)$ to refer to $\mathbf{r}_p^{'}$ and $\mathbf{r}_p$, respectively. This ACF model is specific to scattering from a collection of *non-interacting* electrons, a concept which provides a useful baseline for modeling scattering from ionospheric plasmas where electron interactions are to be expected.

For the non-interacting electron gas, the ACF (5.17) at $\tau = 0$ yields

$$\langle |E_s|^2 \rangle = \frac{r_e^2 |E_i|^2}{r^2} P = 4\pi r_e^2 \frac{|E_i|^2}{4\pi r^2} \frac{P}{V} V = \sigma_e N V \frac{|E_i|^2}{4\pi r^2}, \tag{5.18}$$

where $N \equiv P/V$ is the *average* electron density of volume V. Clearly, (5.18) indicates the backscatter RCS of the volume is

$$\sigma = \sigma_e N V, \tag{5.19}$$

which in turn indicates that we can think of

$$\sigma_v \equiv \sigma_e N \tag{5.20}$$

as the corresponding RCS *per unit volume*. This, so-called volumetric RCS, σ_v, is just the single-electron RCS $\sigma_e = 4\pi r_e^2$ times the average number of electrons per unit volume.

The model results above, specifically, (5.17)-(5.20), will be modified in minor or major ways in practical ionospheric radar experiments, depending on whether the ionosphere is in thermal equilibrium (minor changes) or not (major changes). In either case, the required changes in (5.17)-(5.20) will reflect the fact that ionospheric electrons do interact with one another to varying degrees as a consequence of Coulomb forces. Before describing how the interactions take place, our immediate objectives will be: (i) to show how $\langle e^{j\mathbf{k}\cdot(\mathbf{r}_p(t+\tau)-\mathbf{r}_p(t))}\rangle$ can be evaluated in terms of a probability density function (pdf) for particle velocities, and (ii) to show how the same expression is related to the ACF of electron density waves to be expected in a gas of non-interacting electrons.

Let us start with objective (ii): We can express the *electron density function* of our volume V as

$$n(t,\mathbf{r}) = \sum_{p=1}^{P} \delta(\mathbf{r} - \mathbf{r}_p(t)), \tag{5.21}$$

where $\mathbf{r}_p(t)$ denotes the trajectory function of each particle. In this density definition the individual electrons are portrayed as point particles, a standard practice of classical electrodynamics which can be justified on quantum mechanical grounds as long as de Broglie[5] wavelength of an average electron is sufficiently small. The key scale to compare to is the mean interparticle distance $\sim (V/P)^{1/3}$ of P particles spread out over volume V. Spatial Fourier transforming the density (18.8) over volume V gives us

$$n(t,\mathbf{k}) = \int_V d\mathbf{r}\, n(t,\mathbf{r}) e^{j\mathbf{k}\cdot\mathbf{r}} = \sum_{p=1}^{P} e^{j\mathbf{k}\cdot\mathbf{r}_p(t)} \tag{5.22}$$

which can be viewed as the complex amplitude of a density wave of position with a wavevector $\mathbf{k}$. Hence, we have, for the case of non-interacting electrons with independent trajectories $\mathbf{r}_p(t)$,

$$\langle n^*(t,\mathbf{k}) n(t+\tau,\mathbf{k})\rangle = \sum_{p=1}^{P} \langle e^{j\mathbf{k}\cdot(\mathbf{r}_p(t+\tau)-\mathbf{r}_p(t))}\rangle = NV\langle e^{j\mathbf{k}\cdot(\mathbf{r}_p(t+\tau)-\mathbf{r}_p(t))}\rangle \tag{5.23}$$

where we used $P = NV$. Clearly, (18.10) indicates that we can re-express the scattered

[5] de Broglie wavelength for an average electron is $\Delta x = \frac{h}{m\sqrt{2C}}$, where $\sqrt{2}C = \sqrt{2KT/m}$ is the most probable electron thermal speed. Requiring Δx to be much less than $(V/P)^{1/3} = N^{-1/3}$ leads to the condition $\frac{h^2}{2m} N^{2/3} \ll KT$, where the right hand side is recognized as Fermi level to within a numerical constant. As long as electron thermal energy KT is large compared to the Fermi level the electron gas is non-degenerate and can be treated classically.

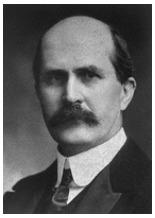
(a)  (b)  (c) 

Figure 5.5: (a) J. J. Thomson (1856-1940), who discovered the electron, (b) his student Max Born (1882-1970), and (c) William H. Bragg (1862-1942), who pioneered the X-ray scattering method to study crystals.

field ACF (5.17) as

$$\begin{aligned}\langle E_s^* E_s^{'}\rangle &= \frac{r_e^2|E_i|^2}{r^2}\sum_{p=1}^{P}\langle e^{j\mathbf{k}\cdot(\mathbf{r}_p(t+\tau)-\mathbf{r}_p(t))}\rangle = \frac{r_e^2|E_i|^2}{r^2}\langle n^*(t,\mathbf{k})n(t+\tau,\mathbf{k})\rangle \\ &= \frac{|E_i|^2}{4\pi r^2}\sigma_e\langle n^*(t,\mathbf{k})n(t+\tau,\mathbf{k})\rangle. \qquad (5.24)\end{aligned}$$

We have reached a key result in (5.24). What this equation says is, the backscattered field ACF is proportional to the ACF of electron density waves in volume V — which is exactly what $\langle n^*(t,\mathbf{k})n(t+\tau,\mathbf{k})\rangle$ denotes — at, specifically, the Bragg scale represented by wavevector $\mathbf{k} = -2k_o\hat{r}$ defined in (5.14). If volume V has zero-amplitude density waves at the Bragg scale, then the backscattered field will be zero. The backscattered field and its statistics depend only and only on the existence and statistics of Bragg scale density waves in the scattering volume. Note that $|\mathbf{k}| = 2k_o$ and thus Bragg wavelength equals $\lambda_o/2$, where λ_o is the radar carrier wavelength. For example, for a radar with a 50 MHz carrier frequency, $\lambda_o = 6$ m, and thus the intensity of radar backscatter depends solely on the intensity of 3 m density waves in the scattering volume.

Another key point is this: although we derived (5.24) in the context of a gas of non-interacting electrons, the result itself applies to any electron gas with any degree or complexity of interactions permitted by our classical density model (18.8). In other words, the scattered field ACF always scales with the ACF $\langle n^*(t,\mathbf{k})n(t+\tau,\mathbf{k})\rangle$ of Bragg scale density waves, whether or not the electrons composing the waves are interacting — this is a straightforward consequence of the fact that (5.15) and (18.9) imply

$$E_s = -\frac{r_e}{r}E_i\, n(t,\mathbf{k}). \qquad (5.25)$$

Of course, how, and to what extent, electrons are interacting are crucial for how $\langle n^*(t,\mathbf{k})n(t+\tau,\mathbf{k})\rangle$ varies with τ, and it is only in the absence of the interactions we know how to compute the variation as in (18.10), i.e., using

$$\langle n_t^*(t,\mathbf{k})n_t(t+\tau,\mathbf{k})\rangle = NV\langle e^{j\mathbf{k}\cdot(\mathbf{r}_p(t+\tau)-\mathbf{r}_p(t))}\rangle.$$

To evaluate $\langle e^{j\mathbf{k}\cdot(\mathbf{r}_p(t+\tau)-\mathbf{r}_p(t))}\rangle$ — which brings us to our first objective set above — let us assume straight-line trajectories for the non-interacting particles so that

$$\mathbf{r}_p(t+\tau) - \mathbf{r}_p(t) = \mathbf{v}_p\tau.$$

Then, it follows that

$$\langle e^{j\mathbf{k}\cdot(\mathbf{r}_p(t+\tau)-\mathbf{r}_p(t))}\rangle = \langle e^{j\mathbf{k}\cdot\mathbf{v}_p\tau}\rangle = \langle e^{jkv\tau}\rangle = \int dv G(v) e^{jkv\tau} = e^{jk\bar{v}_{ze}\tau} e^{-\frac{C_e^2 k^2\tau^2}{2}} \tag{5.26}$$

if we assume a *Maxwellian* velocity distribution (1-D)

$$G(v) = \sqrt{\frac{m}{2\pi KT}} e^{-\frac{\frac{1}{2}m(v-\bar{v})^2}{KT}} = \frac{e^{-\frac{(v-\bar{v})^2}{2C^2}}}{\sqrt{2\pi}C} \tag{5.27}$$

as the pdf of v, the electron velocity component in scattering direction along $\mathbf{k}$. Above, $C \equiv \sqrt{KT/m}$ is the electron thermal speed, $\bar{v}$ is the first moment of $G(v)$, and a Maxwellian pdf is what is expected for an electron gas in thermal equilibrium.

Substituting (18.11) in (5.24), we obtain

$$\langle E_s^* E_s^{'}\rangle = \frac{|E_i|^2}{4\pi r^2}\sigma_e \sum_{p=1}^{P} \langle e^{j\mathbf{k}\cdot(\mathbf{r}_p(t+\tau)-\mathbf{r}_p(t))}\rangle = \frac{|E_i|^2}{4\pi r^2}\sigma_e NV e^{jk\bar{v}\tau} e^{-\frac{C^2k^2\tau^2}{2}}, \tag{5.28}$$

which describes radar backscatter from an ionosphere of non-interacting electrons in thermal equilibrium, which is, in fact, an accurate description for an unmagnetized[6] ionosphere at sufficiently high radar operation frequencies (as described in Appendix VI).

Exercise: Suggest a method to measure $\bar{v}$ based on result (5.28) and discuss the significance of velocity $\bar{v}$.

We will close this section with a minor detail that needs to be addressed better now than later: Our general ACF model (5.24), i.e.,

$$\langle E_s^* E_s^{'}\rangle = \frac{|E_i|^2}{4\pi r^2}\sigma_e \langle n^*(t,\mathbf{k}) n(t+\tau,\mathbf{k})\rangle \tag{5.29}$$

to be specific, concerns scattering from an ionospheric volume V which is required to satisfy the Rayleigh criterion. However, it turns out that the actual *scattering volume* V_s pertinent to real radar experiments — established by the support volume of the transmitted field pulse $|E_i|$, which is in turn determined by the antenna beamwidth and the duration of the radar pulse[7] — always exceeds the maximum size allowed for V by Rayleigh criterion. Thus, ACF models for realistic V_s need to be *reconstructed* in terms of properly weighted volume integrals of differentials

$$\frac{|E_i|^2}{4\pi r^2}\sigma_e \frac{\langle n^*(t,\mathbf{k}) n(t+\tau,\mathbf{k})\rangle}{V} dV.$$

We will carry on our discussions in the next section using (5.29) and only apply the required volume averaging to ultimate results obtained at the end of our discussions.

[6] In the presence of an ambient magnetic field, particles follow helical, rather than straight line trajectories in between collisions, and therefore $\langle E_s^* E_s^{'}\rangle$ takes a different form — see Appendix VI.

[7] In general $V_s \sim r^2 \delta\Omega \delta r$ for a pulsed backscatter radar, where $\delta\Omega$ is the solid angle of "two-way beam" of radar antenna and $\delta r = \frac{c\delta t}{2}$ is *range resolution* related to the transmitted pulse duration δt.

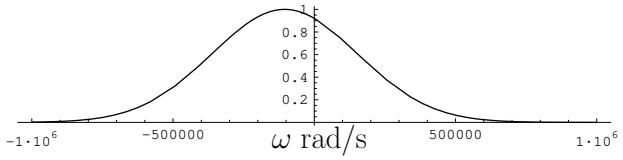


Figure 5.6: Doppler spectrum example for backscattering at 50 MHz from a gas of-non interacting electrons at temperature 1000 K moving away from a radar at an unrealistically large velocity of 50 km/s (by terrestrial norms). Notice that motion away from the radar causes a shift of the peak of the spectrum into the negative ω region, a response commonly known as *red shift*.

5.2.2 Frequency spectrum of backscattered fields

Let

$$\langle |E_s(\omega)|^2 \rangle \equiv \int d\tau e^{-j\omega\tau} \langle E_s^* E_s^{'} \rangle \tag{5.30}$$

denote the Fourier transform of the backscattered field ACF. We will refer to $\langle |E_s(\omega)|^2 \rangle$ as radar *Doppler spectrum* for reasons to become clear in this section.

Substituting (5.29) in (5.30) we find that the spectrum

$$\langle |E_s(\omega)|^2 \rangle = \frac{|E_i|^2}{4\pi r^2} V \sigma_e \langle |n(\omega, \mathbf{k})|^2 \rangle, \tag{5.31}$$

where

$$\langle |n(\omega, \mathbf{k})|^2 \rangle \equiv \int d\tau e^{-j\omega\tau} \frac{\langle n^*(t, \mathbf{k}) n(t+\tau, \mathbf{k}) \rangle}{V} \tag{5.32}$$

is the *space-time power spectrum* of electron density variations within V. It is convenient to write (18.18) as

$$\langle |E_s(\omega)|^2 \rangle = \frac{|E_i|^2}{4\pi r^2} V \sigma(\omega) \tag{5.33}$$

where

$$\sigma(\omega) \equiv \sigma_e \langle |n(\omega, \mathbf{k})|^2 \rangle \tag{5.34}$$

is, by definition, the volumetric RCS *spectrum* of the probed volume of space. Finally, since from the inverse Fourier transform of (5.30) we have

$$\langle |E_s|^2 \rangle \equiv \int \frac{d\omega}{2\pi} \langle |E_s(\omega)|^2 \rangle = \frac{|E_i|^2}{4\pi r^2} V \int \frac{d\omega}{2\pi} \sigma(\omega), \tag{5.35}$$

it follows that volumetric RCS satisfies the relation

$$\sigma_v = \int \frac{d\omega}{2\pi} \sigma(\omega) \tag{5.36}$$

with the RCS spectrum. In view of these relations — which are unconditionally valid in the ionosphere as long as the Born approximation holds — it should be clear that

an overarching aim of backscattering experiments is to measure the Doppler spectrum $\langle |E_s(\omega)|^2 \rangle \propto \sigma(\omega)$ to learn about $\sigma(\omega)$ describing the medium and the dynamics of scattering particles within the medium. Of course, an appropriate volume averaging of these concepts is needed in practical applications as explained at the end of the last section.

Example: Consider again an ionosphere of non-interacting electrons in thermodynamic equilibrium where

$$\frac{\langle n^*(t,\mathbf{k}) n(t+\tau,\mathbf{k})\rangle}{V} = N e^{jk\bar{v}\tau} e^{-\frac{C^2k^2\tau^2}{2}}$$

According to (5.32)

$$\langle |n(\omega,\mathbf{k})|^2 \rangle \equiv \int d\tau e^{-j\omega\tau} N e^{jk\bar{v}\tau} e^{-\frac{C^2k^2\tau^2}{2}} = N\frac{\sqrt{2\pi}}{Ck} e^{-\frac{(\omega-k\bar{v})^2}{2k^2C^2}},$$

and, therefore,

$$\sigma(\omega) \equiv \sigma_e \langle |n(\omega,\mathbf{k})|^2 \rangle = \sigma_e N \frac{\sqrt{2\pi}}{Ck} e^{-\frac{(\omega-k\bar{v})^2}{2k^2C^2}};$$

the total volumetric RCS is

$$\sigma_v \equiv \int \frac{d\omega}{2\pi}\sigma(\omega) = \sigma_e N \int \frac{d\omega}{2\pi}\frac{\sqrt{2\pi}}{Ck} e^{-\frac{(\omega-k\bar{v})^2}{2k^2C^2}} = \sigma_e N.$$

The important result from above example is

$$\langle |E_s(\omega)|^2 \rangle \propto \sigma(\omega) \propto e^{-\frac{(\omega-k\bar{v})^2}{2k^2C^2}},$$

which indicates that the spectrum $\langle |E_s(\omega)|^2 \rangle$ — see Figure 5.6 for an example — peaks at a frequency

$$\omega = \omega_d \equiv k\bar{v},$$

where, $\bar{v}$ denotes the mean velocity of the Maxwellian electron population in the direction of $\mathbf{k} = -2k_o\hat{r}$, pointing from the scattering volume towards the radar antenna. Clearly, the observed radar spectrum $\langle |E_s(\omega)|^2 \rangle$ provides a means of computing

$$\bar{v} = \frac{\omega_d}{k},$$

the mean velocity of scattering electrons towards the probing radar, and, clearly, the frequency ω_d of the peak of $\langle |E_s(\omega)|^2 \rangle$ is the *Doppler shift* of the backscattered radiowave frequency from its original transmission value ω_o due to velocity $\bar{v}$ of an average electron in the probed volume. Since the electrons have a spread of velocities about $\bar{v}$ by $\pm C$, a spread of Doppler shifted returns constitute the observed Doppler spectrum like in Figure 5.6. Thermal speed C of the electrons can be readily inferred from the width of the observed spectrum.

While the results discussed above were specific to the case of non-interacting electrons, the same ideas apply in many other situations. The mean velocity of scattering targets towards the radar can usually be determined with $\bar{v} = \omega_d/k$ from the peak frequency ω_p

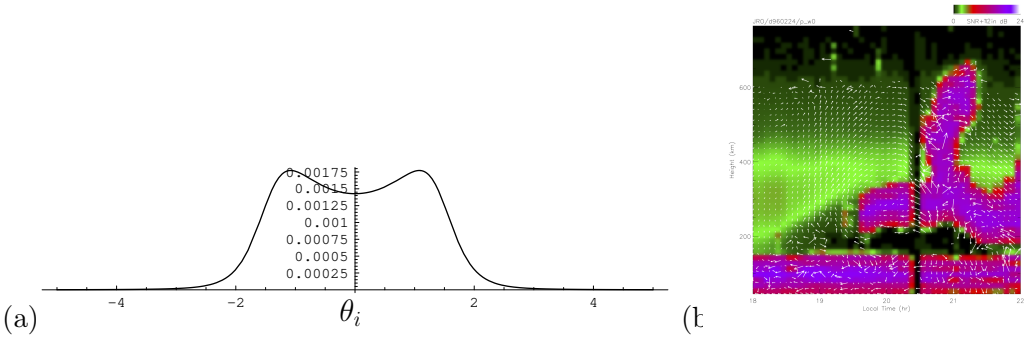


Figure 5.7: (a) Incoherent scatter Doppler spectrum at 50 MHz for an ionosphere in thermal equilibrium with $T_e = T_i = 1000$ K and zero mean drift. The spectrum is plotted as a function of normalized frequency $\theta_i \equiv \frac{\omega/k}{\sqrt{2C_i}}$, where C_i is the ion-thermal speed for O_2^+. (b) A backscatter radar power map exemplifying ionospheric coherent and incoherent scatter. Power (color scale) and drift velocity (arrows) data were obtained with the 50 MHz Jicamarca incoherent scatter radar located near Lima, Peru. Intense echoes depicted in purple are coherent scatter from plasma instabilities.

of the Doppler spectrum $\langle |E_s(\omega)|^2 \rangle$. For instance backscatter from a real ionosphere in thermodynamic equilibrium — for which the RCS spectrum is a complicated expression like

$$\sigma(\omega) = \sigma_e 2N \frac{|j(k^2\lambda_e^2 + \mu) + \mu\theta_i J(\theta_i)|^2 \frac{\mathrm{Re}\{J(\theta_e)\}}{k\sqrt{2C_e}} + |j + \theta_e J(\theta_e)|^2 \frac{\mathrm{Re}\{J(\theta_i)\}}{k\sqrt{2C_i}}}{|j(k^2\lambda_e^2 + 1 + \mu) + \theta_e J(\theta_e) + \mu\theta_i J(\theta_i)|^2} \tag{5.37}$$

as explained in Appendix VI — known as ionospheric *incoherent scatter* behaves that way. The volumetric RCS of incoherent scatter obeys

$$\sigma_v \approx \frac{\sigma_e N}{1 + T_e/T_i}, \tag{5.38}$$

where T_e and T_i denote the temperatures to the electron and ion gases which can be different. This result is valid under a condition $k\lambda_e \ll 1$ but does not apply in a direction perpendicular to the ambient magnetic field if $T_e > T_i$ — see, e.g., *Farley* [1971] and the discussion in Appendix VI. Figure 5.7a shows an example of a possible incoherent scatter Doppler spectrum plotted according to (18.23).

In backscatter from a real ionosphere *not* in thermal equilibrium, the RCS σ_v can be orders of magnitude larger than (5.38) — that case is usually referred to as *coherent scatter*. Examples of coherent scatter include radar backscatter from ionospheric plasma instabilities such as spread F and the equatorial electrojet (see Figure 5.7b), as well as backscatter from the neutral atmosphere, the so-called MST region, below 100 km altitude (see Figure 5.8).

The RCS spectrum model (18.22) for scattering from a gas of free electrons can be readily modified to describe radar backscattering from the neutral atmosphere in terms of the

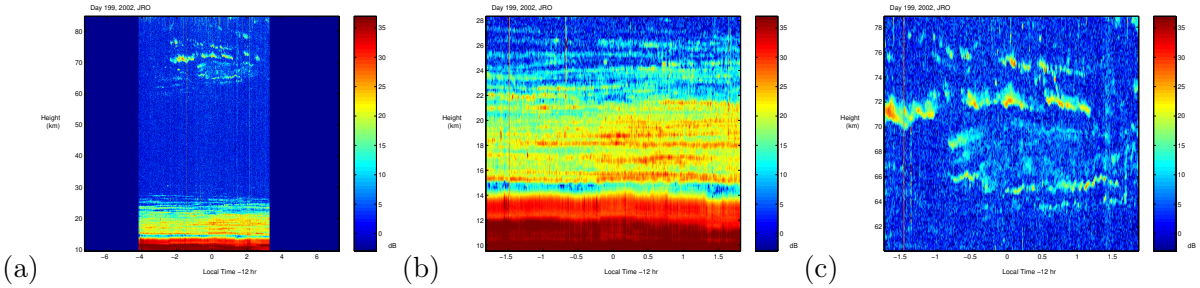


Figure 5.8: (a) Backscattered power data (SNR) collected in an MST experiment conducted with the 50 MHz Jicamarca radar located. Panels (b) and (c) show details from the troposphere/stratosphere and mesosphere regions.

frequency spectrum $\langle|\delta\epsilon(\omega,\mathbf{k})|^2\rangle$ of Bragg scale permittivity waves replacing $\langle|n(\omega,\mathbf{k})|^2\rangle$ due to electron density fluctuations. This is accomplished by using the transformations

$$\epsilon \rightarrow \epsilon - j\frac{\sigma}{\omega} \quad \Rightarrow \quad \delta\epsilon \rightarrow \delta\epsilon - j\frac{\delta\sigma}{\omega} = \delta\epsilon - \frac{\delta N e^2}{m\omega^2}.$$

from Section 1 in reverse direction. Clearly, an electron density fluctuation δN is equivalent to a permittivity fluctuation $\delta\epsilon$ times a factor $m\omega_o^2/e^2$ assuming a frequency $\omega = \omega_o$. It follows that

$$\langle|n(\omega,\mathbf{k})|^2\rangle = \frac{m^2\omega_o^4}{e^4}\langle|\delta\epsilon(\omega,\mathbf{k})|^2\rangle,$$

and thus,

$$\begin{aligned}
\sigma(\omega) &= \sigma_e\langle|n(\omega,\mathbf{k})|^2\rangle = 4\pi r_e^2\frac{m^2\omega_o^4}{e^4}\langle|\delta\epsilon(\omega,\mathbf{k})|^2\rangle \\
&= 4\pi r_e^2\frac{m^2\omega_o^4}{e^4}\langle|\delta\epsilon(\omega,\mathbf{k})|^2\rangle = 4\pi(\frac{e^2}{4\pi\epsilon_o mc^2})^2\frac{m^2\omega_o^4}{e^4}\langle|\delta\epsilon(\omega,\mathbf{k})|^2\rangle \\
&= \frac{k_o^4}{4\pi}\langle|\frac{\delta\epsilon(\omega,\mathbf{k})}{\epsilon_o}|^2\rangle,
\end{aligned}$$

a well-known result known as *Booker-Gordon formula* relevant to atmospheric radar remote sensing. The cause of $\delta\epsilon$ fluctuations below about 50 km altitude is fluctuations in atmospheric temperature and/or relative humidity (dominant below 20 km).

Exercise: Find an equivalent expression for $\sigma(\omega)$ defined above in terms of $\langle|\delta n(\omega,\mathbf{k})|^2\rangle$ where δn denotes the fluctuating component of radio refractive index — due to electron density and/or permittivity fluctuations.

5.2.3 Doppler spectrum estimation

A radar antenna will convert the backscattered field E_s into some output voltage $V \propto E_s$ as discussed in the previous chapter (also see Section 3 for radar receiver details) which is typically sampled and stored in a digital form for further processing. Let a collection

of such samples V_n represent a time sequence of scattered fields E_s received from some scattering volume. We will now describe how the V_n data can be processed to estimated the Doppler spectrum $\langle |E_s(\omega)|^2 \rangle \propto \sigma(\omega)$ defined in the last section.

An N-point FFT (fast-Fourier transform) of N consecutive V_n samples provides us with

$$v_q \equiv v(q\Delta\omega) = \sum_{n=0}^{N-1} V_n e^{-jq\Delta\omega nT} = \sum_{n=0}^{N-1} V_n e^{-j\frac{2\pi nq}{N}}$$

at discrete frequencies $\omega = q\Delta\omega = q\frac{2\pi}{NT}$. We can call $|v_q|^2$ a Doppler *periodogram*, which has an expected value

$$\begin{aligned} & & \overbrace{\langle V_n^* V_{n+p} \rangle} & \equiv p \\ \langle |v_q|^2 \rangle &= \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} \overbrace{\langle V_n^* V_m \rangle} \, e^{-jq\Delta\omega T} \overbrace{(m-n)} \\ &= \sum_{p=-N}^{N} (N - |p|) \langle V_n^* V_{n+p} \rangle e^{-jq\Delta\omega pT}, \end{aligned} \tag{5.39}$$

where $\langle V_n^* V_{n+p} \rangle$ is the ACF of the voltage data, independent of n and proportional to the scattered field ACF $\langle E_s^* E_s' \rangle$ at a delay $\tau = pT$, which, in turn, is related to the Doppler spectrum $\langle |E_s(\omega)|^2 \rangle$, according to (5.30), via

$$\langle |E_s(\omega)|^2 \rangle \equiv \int d\tau \langle E_s^* E_s' \rangle e^{-j\omega\tau}. \tag{5.40}$$

Clearly, (5.39) multiplied on both sides by an appropriate scaling constant related to the radar antenna length, turns into a Riemann sum version of (5.40) in the limit of increasing N. Therefore, for sufficiently large N, and, small T, so that frequency aliasing is avoided, the average periodogram (5.39) can be regarded a practical estimator of the shape of the radar Doppler spectrum at discrete Doppler frequencies $\omega_q = q\frac{2\pi}{NT}$. To avoid aliasing T should be short enough to satisfy the Nyquist criterion — i.e., $T \leq \frac{1}{2B}$, where B is the bandwidth of spectrum $\langle |E_s(\omega)|^2 \rangle$ in Hertz units — and N large enough to ensure that $\langle V_n^* V_{n \pm N} \rangle \ll \langle |V_n|^2 \rangle$.

Exercise 1: The MATLAB file "data.mat" in our class web site contains radar data samples V_n which are used in this example problem. After downloading the file and placing it in your working directory issue the MATLAB command **load('data.mat').** The **load** command will read in three separate arrays — called **Va**, **Vb**, and **Vc** — of complex numbers with dimensions of 64×10. Verify the dimensions with the command **size(Va)**, for instance. Array **Va** contains 640 consecutive V_n samples (plus noise) of backscatter radar data obtained with a sampling interval of $T = 0.2001$ s (exactly). The data is for a radar range of $r = 65.05$ km in the *mesosphere* above the Jicamarca radar located near Lima in Peru (on 07/19/2002, starting with 13:39:01 LT, with a nearly vertical pointed antenna beam and with a range resolution of $\delta r = \frac{c\delta t}{2} = 150$ m). Array **Vb** for $r = 77.7$ km and **Vc** for $r = 46.95$ km are organized in a similar manner to **Va**. Think of **Va** as a collection of 10 column vectors of 64 elements each, with each element representing a receiver output sample corresponding to $r = 65.05$ km.

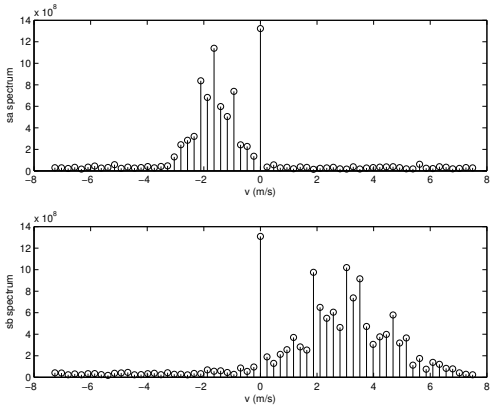


Figure 5.9: The power spectra computed for data sets **Va** and **Vb** discussed in Example 1. The horizontal axes of the spectra are labeled in terms of the equivalent Doppler velocity $v = \frac{\omega}{k}$.

a) Issue the command **plot(Va,'o')** in MATLAB to see a scatter plot of all the elements of array **Va** on the complex plane; different colored circles denote data samples from different columns of the array. What is the impression you get from the scatter plot about the data contained in **Va**? Compare the scatter plots of **Vb** and **Vc** with **Va** — what are the main differences? Which set has the largest average power — i.e., $\langle |V_n|^2 \rangle$ — according to the scatter plots?

b) Issue commands like **plot(real(Va(:,1)))** and **stem(imag(Vb(:,2)))** to examine the real and imaginary parts of sample time series contained in the individual columns of the arrays.

c) Issue the command **plot(abs(fft(Va)).^2)** to see the periodograms of 10 separate 64-point time-series contained in **Va** (each one is plotted with a different color). How do the periodograms of **Vb** and **Vc** compare with the periodograms of **Va**?

d) Issue the command **plot(sum(abs(fft(Va)).^2,2)/10)** to plot the average of the periodograms from **Va**. The result is an estimate of the Doppler spectrum of the backscatter data contained in **Va**. Make spectral plots of the data in **Vb** and **Vc** in a similar way. Which data set — **Va**, **Vb**, or **Vc** — has the largest average power? Explain using the spectral plots just made. Hint: array **Vc** was selected from a range with SNR=0!

e) The spectra of sets **Va** and **Vb** peak at different frequencies $\omega_p = q\frac{2\pi}{64T}$ and hence imply different values of average radar line of sight velocities $u_p = -\frac{\omega_p}{2k}$. What is your estimate of $|u_p|$ for the peak of the spectrum of **Va**? For the Jicamarca radar the carrier frequency is $f_o = \frac{\omega_o}{2\pi} = 50$ MHz, and therefore $k = \frac{2\pi}{\lambda} = \frac{2\pi}{6}$ rad/m. Hint: to calculate $|u_p|$ you need an estimate of $|q|$ associated with the peak of the spectrum. But, where is $q = 0$ — i.e., the DC point — along the horizontal axis of the spectrum plot? Try to find the DC point by doing a **sum(abs(fft(Va+1000)).^2,2)/10** plot that contains lots of power at DC, and then determine $|u_p|$.

f) Issue **help fft** and **help fftshift** commands and find out how to display the **fft** outputs and periodograms over the frequency range $|\omega| = |q\Delta\omega| < \frac{\pi}{T}$, with $\omega = 0$ at the center. Subsequently plot the spectral estimates from **Va** and **Vb** with the horizontal axes labeled

in terms of equivalent radial velocity $u = -\frac{\omega}{2k}$ ranging from negative to positive u's (see Figure 5.9).

g) Are the targets in sets **Va** and **Vb** moving in the same direction with respect to the radar? Explain using the spectral plots produced in the previous part. Which target has a wider spread of relative motions in its own frame of reference, **Va** or **Vb**?

Once an estimate $\tilde{S}_q$ of Doppler spectrum $S_q \equiv \langle |v_q|^2 \rangle$ is obtained we are usually interested in determining the central frequency of the spectrum and its width. While all that can be done by fitting $\tilde{S}_q$ to a spectral model equation S_q (see further discussion in Section 4), an alternative is to determine first and second central moments of $S_q$by using the $\tilde{S}_q$ estimate. The first moment of S_q is defined and estimated as

$$\langle q \rangle \equiv \frac{\sum q S_q}{\sum S_q} \approx \frac{\sum q \tilde{S}_q}{\sum \tilde{S}_q} \equiv \tilde{q},$$

and

$$\langle (q - \langle q \rangle)^2 \rangle = \langle q^2 \rangle - \langle q \rangle^2 \equiv \frac{\sum q^2 S_q}{\sum S_q} - \langle q \rangle^2 \approx \frac{\sum q^2 \tilde{S}_q}{\sum \tilde{S}_q} - \tilde{q}^2 \equiv \tilde{q}^2_{rms}.$$

Scaling $\tilde{q}$ and $\tilde{q}_{rms}$ by $\Delta\omega = \frac{2\pi}{NT}$ we obtain estimates of the spectral mean Doppler shift and width, respectively. These two parameters, in addition to the total received power, are the most important parameters obtained in radar experiments at specific radar ranges (see next section).

5.3 Pulsed radar operations

In the previous section we focused on the physics of soft-target scattering pertinent to atmospheric/ionospheric backscatter radars. In this section we will focus on systems properties of such radars, in particular on pulsed operations and spatial averaging over the radar beam.

5.3.1 Pulsed radar formulation

Assume an antenna input current

$$i(t) = Re\{f(t)e^{j\omega_o t}\} \tag{5.41}$$

where ω_o is the carrier frequency and $f(t)$ is some periodic waveform with period T which will be called interpulse period (IPP). The following discussion will focus on transmitted and backscattered signals received during one IPP since similar descriptions will apply to all periods T.

The current waveform will excite a transmitted field

$$\mathbf{E}_b(\mathbf{r}, t) = Re\{j\eta_o f(t - \frac{r}{c}) k_o \mathbf{l}(\hat{r}) \frac{e^{-jk_o r}}{4\pi r} e^{j\omega_o t}\} \tag{5.42}$$

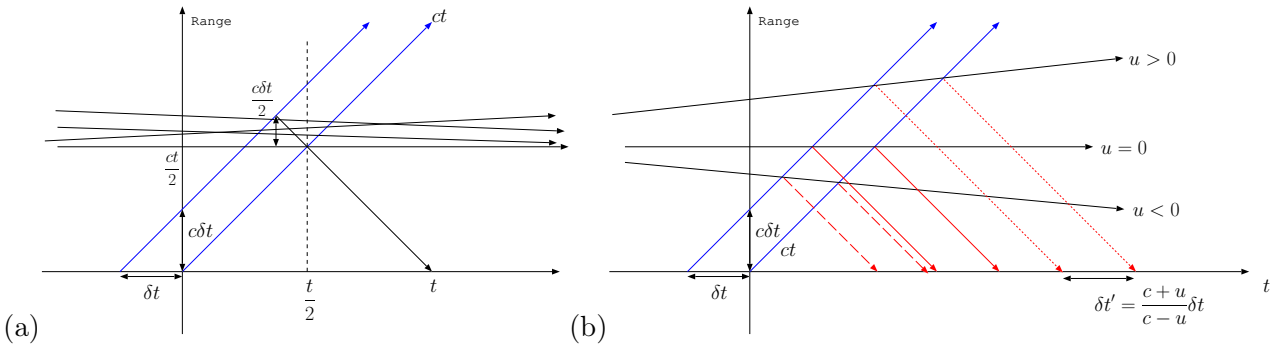


Figure 5.10: (a) Radar output envelope sampled at time t after pulse transmission is a superposition of contributions from targets between ranges $r = c\frac{t}{2}$ and $r + \delta r$, where $\delta r \equiv c\frac{\delta t}{2}$ is said to be range resolution. (b) A range-time diagram illustrating the interactions of a radar pulse of some duration δt with three distinct radar targets. The duration $\delta t'$ of pulse echoes depend on δt and target velocity u as shown, but in the non-relativistic cases $\delta t' = \delta t$ is a very good approximation. If the target moves a distance much less than a Bragg wavelength $\lambda_o/2$ in a pulse duration δt, then the phase variation of the echo can also be ignored.

where $f(t - \frac{r}{c})$ is a delayed[8] version of $f(t)$. It is assumed that time variation of pulse waveform $f(t)$ is slow compared to the oscillation period $\frac{2\pi}{\omega_o}$ of the radar carrier, and, therefore, in most of our calculations we treat $f(t)$ as if it were a phasor in the general spirit of narrowband signal representation and analysis.

A stationary hard target at some location $\mathbf{r}_p$ will be illuminated by $\mathbf{E}_b(\mathbf{r}_p, t)$, and produce a backscattered field $\propto \mathbf{E}_b(\mathbf{r}_p, t - \frac{|\mathbf{r}_p|}{c}) = \mathbf{E}_b(\mathbf{r}_p, t - \frac{r_p}{c})$ at the radar antenna, producing an open-circuit voltage phasor that can be denoted as

$$V(t) = Z_p f(t - \frac{2r_p}{c}) e^{-j2k_o r_p} = Z_p f(t - \frac{2r_p}{c}) e^{jkr_p}; \tag{5.43}$$

above, $r_p \equiv |\mathbf{r}_p|$ is the radar range of the target, $k = -2k_o$ is Bragg wavenumber, and Z_p is an appropriate scaling factor. To obtain a handle on Z_p, we note that the available power

$$P_a = \frac{|V(t)|^2}{8R_{rad}} = \frac{|Z_p|^2 |f(t - \frac{2r_p}{c})|^2}{8R_{rad}} = \frac{|Z_p|^2}{4R_{rad}^2} P_t(t - \frac{2r_p}{c}) = \frac{P_t(t - \frac{2r_p}{c})}{4\pi r_p^2} G_p \frac{\sigma_p}{4\pi r_p^2} A_p,$$

and therefore

$$|Z_p|^2 = 4R_{rad}^2 \frac{G_p A_p}{(4\pi r_p^2)^2} \sigma_p, \tag{5.44}$$

[8]We assume here that the carrier frequency ω_o is sufficiently large to ignore the distinctions between v_g, v_p, and c.

where G_p and A_p are the antenna gain and effective area in the direction of target σ_p. Naturally, if multiple targets were present we would use

$$V(t) = \sum_p Z_p f(t - \frac{2r_p}{c}) e^{jkr_p} \tag{5.45}$$

assuming that non-linear effects (multi-scatter, extinction) are negligible.

Although the results above are for stationary targets, they can certainly be used for targets in motion when the speeds involved are sufficiently slow, as argued graphically in Figure 5.10 depicting the time-history of the leading and trailing edges of pulsed radar transmissions and possible radar targets.

Also shown graphically in Figure 5.10 is the fact that radar *range resolution* δr scales with pulse length δt as $\delta r = \frac{c\delta t}{2}$. The resolution δr corresponds to the depth in radar range r that contains all the targets which are causing the returns sampled at any given instant. It is also the depth dimension of the radar scattering volume V_s pertinent to soft-target case. The exception to the fact that $\delta r = \frac{c\delta t}{2}$ occurs when coded pulses are used together with matched filtering as discussed in Section 4.

For ionospheric *soft-target* returns — as opposed to the *hard-target* case considered above — we adopt a linear and causal open-circuit voltage model

$$V(t) = \int d\mathbf{r} Z(\mathbf{r}) f(t - \frac{2r}{c}) n(\mathbf{r}, t - \frac{r}{c}) e^{jkr}, \tag{5.46}$$

where $n(\mathbf{r}, t)$ stands for the fluctuating component of ionospheric electron density and $k = -2k_o$ is the Bragg wavenumber again such that $\mathbf{k} = k\hat{r}$. Since, in the extreme case of $n(\mathbf{r}, t) = \delta(\mathbf{r} - \mathbf{r}_p)$, this model will reduce to

$$V(t) = Z(\mathbf{r}_p) f(t - \frac{2r_p}{c}) e^{jkr}, \tag{5.47}$$

it is clear that $Z(\mathbf{r})$ in (5.46) is constrained, in analogy to (5.44), by

$$|Z(\mathbf{r})|^2 = 4R_{rad}^2 \frac{G(\hat{r})A(\hat{r})}{(4\pi r^2)^2} \sigma_e. \tag{5.48}$$

With a straightforward re-scaling of (5.48), we can also choose to interpret n in (5.46) as a refractive index fluctuation in applications relevant to the lower atmosphere.

The main utility of soft-target antenna voltage model (5.46) is that $Z(\mathbf{r}) f(t - \frac{2r}{c})$ in the integrand helps specify the proper way of doing the volume averages discussed in the previous section. To see how this comes about it is useful calculate the voltage variance $\langle |V(t)|^2 \rangle$ based on (5.46). The result can also be generalized into a soft-target radar equation as we will see.

5.3.2 Soft-target radar equation

Let[9]

$$W(\mathbf{r},t) \equiv Z(\mathbf{r}) f(t - \frac{2r}{c}) \tag{5.49}$$

and use (5.46) to compute

$$\begin{aligned} V^*(t)V(t) &= |V(t)|^2 = \int d\mathbf{r} W^*(\mathbf{r},t) n^*(\mathbf{r}, t - \frac{r}{c}) e^{-jkr} \int d\mathbf{r} W(\mathbf{r},t) n(\mathbf{r}, t - \frac{r}{c}) e^{jkr} \\ &= \int d\mathbf{r} \int d\mathbf{r}' W^*(\mathbf{r},t) W(\mathbf{r}',t) n^*(\mathbf{r}, t - \frac{r}{c}) n(\mathbf{r}', t - \frac{r}{c}) e^{jk(r-r')}. \end{aligned}$$

Taking the expected value of both sides,

$$\langle |V(t)|^2 \rangle = \int d\mathbf{r} \int d\mathbf{r}' W^*(\mathbf{r},t) W(\mathbf{r}',t) \langle n^*(\mathbf{r}, t - \frac{r}{c}) n(\mathbf{r}', t - \frac{r'}{c}) \rangle e^{jk(r-r')}.$$

Within the integrand above $\langle n^*(\mathbf{r}, t - \frac{r}{c}) n(\mathbf{r}', t - \frac{r'}{c}) \rangle$ is the *space-time* ACF of density fluctuations $n(\mathbf{r},t)$. If the fluctuations have a homogeneous and stationary statistics[10] the ACF will have a form

$$\langle n^*(\mathbf{r}, t - \frac{r}{c}) n(\mathbf{r}', t - \frac{r'}{c}) \rangle = g(\mathbf{r}' - \mathbf{r}, -\frac{r' - r}{c}),$$

some function $g(\cdot,\cdot)$ that only depends on differences in space and time coordinates, primed and unprimed. Also if $g(\mathbf{r}' - \mathbf{r}, -\frac{r'-r}{c})$ vanishes rapidly with an increasing magnitude for $\mathbf{x} \equiv \mathbf{r}' - \mathbf{r}$, we can proceed as

$$\langle |V(t)|^2 \rangle = \int d\mathbf{r} \int d\mathbf{r}' |W(\mathbf{r},t)|^2 g(\mathbf{r}' - \mathbf{r}, 0) e^{jk(r'-r)},$$

where $\mathbf{r}'$ has been replaced with $\mathbf{r}$ everywhere except for in two places where the distinction between $\mathbf{r}'$ and $\mathbf{r}$ still matters. Applying a change of variables $\mathbf{r}' = \mathbf{r} + \mathbf{x}$, we notice

$$g(\mathbf{r}' - \mathbf{r}, 0) e^{jk(r'-r)} \approx g(\mathbf{x}, 0) e^{jk\hat{r}\cdot\mathbf{x}} = \langle n^*(\mathbf{r},t) n(\mathbf{r} + \mathbf{x}, t) \rangle e^{j\mathbf{k}\cdot\mathbf{x}}$$

for $|\mathbf{x}| \ll r$, and, therefore, obtain

$$\langle |V(t)|^2 \rangle \approx \int d\mathbf{r} |W(\mathbf{r},t)|^2 \int d\mathbf{x} \langle n^*(\mathbf{r},t) n(\mathbf{r} + \mathbf{x}, t) \rangle e^{j\mathbf{k}\cdot\mathbf{x}}. \tag{5.50}$$

Finally, combining the fact that $\int d\mathbf{r} = \int dr \int r^2 d\Omega$ with

$$|W(\mathbf{r},t)|^2 \equiv |Z(\mathbf{r})|^2 |f(t - \frac{2r}{c})|^2 = 4R_{rad}^2 \frac{G(\hat{r}) A(\hat{r}) \sigma_e}{(4\pi r^2)^2} |f(t - \frac{2r}{c})|^2$$

[9] It can be shown rigorously that $W(\mathbf{r},t) = -j\eta_o k_o r_e f(t - \frac{2r}{c}) \frac{|\mathbf{l}(\hat{r})|^2}{4\pi r^2}$ in terms of effective antenna length $\mathbf{l}(\hat{r})$ in transmission and $\mathbf{l}^*(\hat{r})$ in reception; the antenna is reconfigured for reception as $\mathbf{l}^*(\hat{r})$ specifically when circular polarization is employed.

[10] In practice $n(\mathbf{r},t)$ cannot be expected to have homogeneous and stationary statistics over all positions and times. However it is sufficient that these conditions are satisfied over a scattering volume and an observation interval to justify this derivation.

within (5.50) we obtain[11]

$$\frac{\langle |V(t)|^2 \rangle}{8R_{rad}} \approx \int dr \int d\Omega \frac{G(\hat{r})A(\hat{r})}{(4\pi r)^2} P_t(t - \frac{2r}{c})\sigma_v, \tag{5.51}$$

where[12]

$$P_t(t - \frac{2r}{c}) = \frac{1}{2} R_{rad} |f(t - 2r/c)|^2 \tag{5.52}$$

is the transmitted radar power and

$$\sigma_v \equiv \sigma_e \int d\mathbf{x} \langle n^*(\mathbf{r}, t) n(\mathbf{r} + \mathbf{x}, t) \rangle e^{j\mathbf{k}\cdot\mathbf{x}} \tag{5.53}$$

the Bragg scatter RCS per unit volume. In addition, we can write

$$\frac{\langle |V(t)|^2 \rangle}{8R_{rad}} \approx \int \frac{d\omega}{2\pi} \int dr \int d\Omega \, \frac{G(\hat{r})A(\hat{r})}{(4\pi r)^2} P_t(t - \frac{2r}{c})\sigma(\omega), \tag{5.54}$$

where

$$\sigma(\omega) \equiv \sigma_e \int d\tau \int d\mathbf{x} \langle n^*(\mathbf{r}, t) n(\mathbf{r} + \mathbf{x}, t + \tau) \rangle e^{-j(\omega\tau - \mathbf{k}\cdot\mathbf{x})} \tag{5.55}$$

is the RCS spectrum.

Clearly, (5.51) above is a *radar equation* for soft-target radars, and, its expansion (5.54) embodies a volume integrated Doppler spectrum model. The Fourier transforms

$$\int d\mathbf{x} \langle n^*(\mathbf{r}, t) n(\mathbf{r} + \mathbf{x}, t) \rangle e^{j\mathbf{k}\cdot\mathbf{x}}$$

and

$$\int d\tau \int d\mathbf{x} \langle n^*(\mathbf{r}, t) n(\mathbf{r} + \mathbf{x}, t + \tau) \rangle e^{-j(\omega\tau - \mathbf{k}\cdot\mathbf{x})}$$

in (5.53) and (5.55) represent the *spatial* and *space-time* power spectra of the *random process* $n(\mathbf{r}, t)$ relevant to scattering, and, as such, we can use for scattering cross-sections σ_v and $\sigma(\omega)$ the results developed in Section 2. In fact, (5.51) and (5.54) above are the properly volume averaged models promised back in Section 2.

It should be noted that both σ_v and $\sigma(\omega)$ are in general functions of position $\mathbf{r}$ within the scattering volume V_s. However, for a homogeneous and isotropic scattering process $n(\mathbf{r}, t)$, the RCS σ_v will not depend on $\mathbf{r}$, in which case the radar equation (5.51) simplifies as

$$\langle P_a \rangle = \int dr \frac{P_t(t - \frac{2r}{c})}{4\pi r^2} \int d\Omega \frac{G(\hat{r})}{4\pi} A(\hat{r})\sigma_v \approx \frac{P_t A_{BS} \delta R}{4\pi R^2} \sigma_v, \tag{5.56}$$

[11] The approximation sign reminds us that the result is only valid for density fluctuations with stationary and homogeneous statistics and only for $L_r \ll r\Omega_o^{1/2}$ and $L_t \ll \min\{\lambda^{1/3} r^{2/3}, \ \delta r\}$; here L_r and L_t stand for the widths of $R(\mathbf{x}, 0)$ in radial and transverse directions, δr is range resolution, and $r\Omega_o^{1/2}$ denotes the width of the antenna beam at range r. The inequalities are needed to justify the replacements of $\mathbf{r}'$ by $\mathbf{r}$ and $r' - r$ by $\hat{r} \cdot \mathbf{x}$.

[12] Assuming the signal bandwith is less than the system bandwidth. See Section 4 how possible exceptions may be treated with a suitable replacement of $f(t)$ in (5.52).

where $R \equiv \frac{ct}{2}$ is the pertinent radar range, P_t the average power output during pulse transmission,

$$\delta R \equiv \int dr \frac{P_t(t - \frac{2r}{c})}{P_t} \tag{5.57}$$

the range resolution (for uncoded pulses — see Section 4 for pulse coding and modified δR), and

$$A_{BS} \equiv \int d\Omega \frac{G(\hat{r})}{4\pi} A(\hat{r}) \tag{5.58}$$

is known as *backscatter aperture* of the radar antenna.

Thus, in summary and in a very compact form, we have

$$\langle P_a \rangle = \frac{P_t A_{BS} \delta R}{4\pi R^2} \sigma_v, \quad \text{and} \quad P_a = \frac{P_t A(\hat{r}) G(\hat{r})}{(4\pi)^2 R^4} \sigma, \tag{5.59}$$

for soft- and hard-target radars, respectively. These *radar equations* highlight the fact that for soft-target radars the received power varies with range as R^{-2} as opposed to R^{-4}. Also the dependence on antenna gain has a peculiar form for soft-target radars which is expressed in terms of the A_{BS} parameter introduced above.

Exercise 1: Show that backscatter aperture $A_{BS} \leq \frac{\lambda^2}{4\pi} D \leq A_p$ where A_p is the physical area of the radar antenna and D the antenna directivity. Hint: $\int d\Omega \frac{G(\hat{r})}{4\pi} = 1$.

Exercise 2: Show that for a square array antenna with a physical area $A_p \gg \lambda^2$ and a uniform illumination, $A_{BS} = \frac{4}{9} A_p$. Hint: for the described array $D = \frac{4\pi}{\lambda^2} A_p$ and the gain pattern has a sinc squared form.

5.3.3 ACF and ambiguity function for Bragg scatter:

Starting with

$$V(t) = \int d\mathbf{r} W(\mathbf{r}, t) n(\mathbf{r}, t - \frac{r}{c}) e^{jkr},$$

where $W(\mathbf{r}, t) \equiv Z(\mathbf{r}) f(t - \frac{2r}{c})$ and $k = -2k_o$, we can model the Bragg scattered signal ACF as

$$\begin{aligned}
\langle V^*(t) V(t+\tau) \rangle &= \int d\mathbf{r} \int d\mathbf{r}' W^*(\mathbf{r}, t) W(\mathbf{r}', t+\tau) \langle n(\mathbf{r}, t - \frac{r}{c}) n^*(\mathbf{r}', t + \tau - \frac{r'}{c}) \rangle e^{jk(r'-r)} \\
&\approx \int d\mathbf{r} \int d\mathbf{x} W^*(\mathbf{r}, t) W(\mathbf{r} + \mathbf{x}, t + \tau) R(\mathbf{x}, \tau) e^{jk\hat{r} \cdot \mathbf{x}} \\
&\approx \int d\mathbf{r} W^*(\mathbf{r}, t) W(\mathbf{r}, t + \tau) \int d\mathbf{x} R(\mathbf{x}, \tau) e^{j\mathbf{k} \cdot \mathbf{x}}
\end{aligned}$$

using the same assumptions about the ACF of $n(\mathbf{r}, t)$ as in our earlier derivation of the voltage variance. The inner integral above defines the temporal ACF

$$R(\mathbf{k}, \tau) \equiv \int d\mathbf{x} R(\mathbf{x}, \tau) e^{j\mathbf{k} \cdot \mathbf{x}}$$

of Bragg scale PSD, the volume sum of which, weighted by an *ambiguity function*

$$A(\mathbf{r},t,\tau) \equiv W^*(\mathbf{r},t)W(\mathbf{r},t+\tau),$$

yields the radar ACF $\langle V^*(t)V(t+\tau)\rangle$.

Note that within the ambiguity function $W(\mathbf{r},t) \propto f(t-\frac{2r}{c})$, but when the effect of radar receiver is taken into account, $f(t)$ needs to replaced with $h(t)*f(t)$, as explained later in the chapter. The Bragg-scatter ACF and ambiguity function are primarily used in broadband or "overspread" radar applications when the radar IPP T "undersamples" the process $n(\mathbf{r},t)$ in violation of the Nyquist criterion (to avoid "range aliasing"). In overspread applications time lags τ of ACF estimates are chosen such that $\tau \ll \delta t \ll T$, meaning that the sampling interval of radar antenna output is shorter than the transmitter pulse length.

5.3.4 Partial reflections and Fresnel scatter

Consider once more the Born approximation model for backscattered signal

$$V(t) = \int d\mathbf{r} W(\mathbf{r},t) n(\mathbf{r}, t-\frac{r}{c}) e^{jkr},$$

where k is Bragg wavenumber $-2k_o$ and $W(\mathbf{r},t) \equiv Z(\mathbf{r})f(t-\frac{2r}{c})$ highlights a radar "resolution volume" centered about radar range

$$R = \frac{ct}{2}.$$

Using a representation

$$\mathbf{r} \equiv \hat{z}R + \mathbf{r}' \equiv \hat{z}R + (x,y,z')$$

and

$$\begin{aligned} r &= |\mathbf{r}| = |\hat{z}R + \mathbf{r}'| = \sqrt{R^2 + 2R\hat{z}\cdot\mathbf{r}' + r'^2} = R\sqrt{1 + 2\frac{\hat{z}\cdot\mathbf{r}'}{R} + \frac{r'^2}{R^2}} \\ &= R(1 + \frac{1}{2}(\frac{2z'}{R} + \frac{x^2+y^2+z'^2}{R^2}) - \frac{1}{8}(\frac{2z'}{R} + \frac{x^2+y^2+z'^2}{R^2})^2 + \cdots) \\ &\approx R + z' + \frac{x^2+y^2}{2R} = z + \frac{x^2+y^2}{2R}, \end{aligned}$$

where we neglected higher order expansion terms starting with $z'r'^2/R^3$, we can write

$$V(t) \approx \int d\mathbf{r} e^{jk\frac{x^2+y^2}{2R}} W(\mathbf{r},t)\delta n(\mathbf{r}, t-\frac{r}{c}) e^{jkz}.$$

If $n(\mathbf{r},t)$ is slowly varying this result can be interpreted as a 1-D Fourier transform over z of a windowed projection

$$\int dx \int dy e^{jk\frac{x^2+y^2}{2R}} W(x,y,z,t) n(x,y,z,\frac{t}{2})$$

of $n(\mathbf{r}, \frac{t}{2})$ evaluated at Bragg wavenumber $k = -2k_o$. Also, for a stratified and time-invariant $n(\mathbf{r}, t) = n(z)$ the result

$$V(t) \approx \int dx \int dy e^{jk\frac{x^2+y^2}{2R}} \int dz W(x, y, z, t) n(z) e^{jkz}$$

can be interpreted the same way. A variance expression $\langle |V(t)|^2 \rangle$ proportional to a k sample of a 1-D spatial PSD of $n(z)$ can be obtained from this expression if $n(z)$ is a sample function of a homogeneous random process, a situation described as "Fresnel scatter". The result also describes non-stochastic partial reflections — so-called "Fresnel reflections" — in terms of a Bragg scale Fourier sample of the $n(z)$ function.

5.4 Radar receiver, SNR, and parameter estimation

Let

$$Re\{V(t)e^{j\omega_o t}\} + v_n(t) \tag{5.60}$$

denote the open-circuit voltage of a radar antenna during reception, where $V(t)$ is the radar signal model developed in the previous section and $v_n(t)$ is additive broadband antenna noise (see Section 4 in last chapter). A modern radar receiver will sample and process the noisy antenna voltage in order to detect $V(t)$ and its statistics in the presence of $v_n(t)$.

Let us refer to the receiver output in the absence of antenna noise $v_n(t)$ as

$$s(t) \equiv h(t) * V(t). \tag{5.61}$$

We will then call the noisy receiver output including $s(t)$ and the response to antenna noise $v_n(t)$ as

$$y(t) = s(t) + n(t), \tag{5.62}$$

where

$$n(t) \equiv h(t) * \{v_n(t) 2e^{-j\omega_o t}\}. \tag{5.63}$$

In above definitions $h(t)$ denotes the baseband filter of the radar receiver applied to the open antenna output after it has been heterodyned into the baseband via mixing with $2e^{-j\omega_o t}$ (digitally, in most state-of-the-art radar receivers). The purpose of filter $h(t) \leftrightarrow H(\omega)$ is of course to limit the noise content of the output, namely, $n(t)$.

In selecting $h(t)$ we consider a number of factors including the effective range resolution of filtered signal $s(t)$ and the output signal to noise ratio defined as

$$\text{SNR} \equiv \frac{|s(t)|^2}{\langle |n(t)|^2 \rangle} \quad \text{or} \quad \frac{\langle |s(t)|^2 \rangle}{\langle |n(t)|^2 \rangle}. \tag{5.64}$$

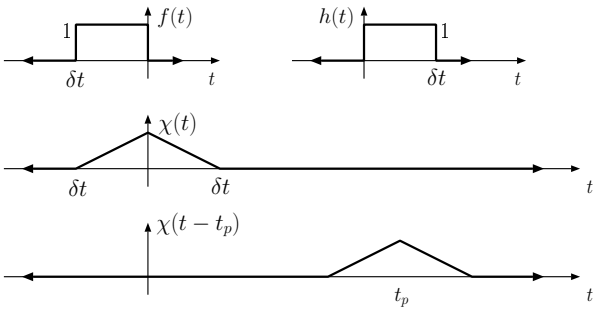


Figure 5.11: A simple radar pulse $f(t)$, a possible impulse response $h(t)$ as well as shifted and unshifted versions of $\chi(t) = h(t) * f(t)$. Radar output signal consists of delayed and amplitude scaled copies of $\chi(t)$ which can only be *resolved* one by one in the output if the individual delays exceeds a pulse length δt.

5.4.1 Discrete target SNR and matched filter

Let us first examine the output SNR pertinent to a stationary point-target located at some range r_p. In that case

$$V(t) = Z_p f(t - t_p) e^{jkr_p} \tag{5.65}$$

and

$$s(t) \equiv h(t) * V(t) = Z_p h(t) * f(t - t_p) e^{jkr_p} = \chi(t - t_p) Z_p e^{jkr_p}, \tag{5.66}$$

where $f(t)$ is the transmitted pulse envelope, $t_p \equiv 2r_p/c$ is the two-way delay to the target, $k = -2k_o$ the Bragg wavenumber, Z_p the appropriate scaling constant, and

$$\chi(t) \equiv h(t) * f(t). \tag{5.67}$$

Figure 5.11 illustrates $\chi(t)$ and $\chi(t - t_p)$ for a point-target assuming for the radar pulse $f(t)$ a simple rectangular shape and a *matched filter* impulse response $h(t)$, which is described below.

By definition, a matched filter with an impulse response $h(t)$ maximizes the peak value of the signal-to-noise ratio of a radar receiver output, namely

$$\mathrm{SNR} \equiv \frac{|s(t)|^2}{\langle |n(t)|^2 \rangle} = \frac{|Z_p|^2 |\chi(t - t_p)|^2}{\langle |n(t)|^2 \rangle} = |Z_p|^2 \frac{|h(t) * f(t - t_p)|^2}{\langle |n(t)|^2 \rangle}, \tag{5.68}$$

where the mean-square noise output $\langle |n(t)|^2 \rangle$ is obtained as follows:

Since

$$n(t) = 2h(t) * \{v_n(t) e^{-j\omega_o t}\}, \tag{5.69}$$

and, furthermore,

$$\langle v_n^2(t) \rangle = v_{rms}^2 = 4R_{rad} KT df \tag{5.70}$$

over an infinitesimal bandwidth $df = \frac{d\omega}{2\pi}$, it follows that the mean-squared $n(t)$ over the same bandwidth is

$$\langle n^2(t) \rangle \quad = \quad 4R_{rad} KT \frac{d\omega}{2\pi} 4|H(\omega)|^2. \tag{5.71}$$

Clearly, then, the noise power at the receiver output counting all the bandwidths over positive ω is

$$\langle n^2(t)\rangle \;=\; 4R_{rad}KT\int_0^\infty \frac{d\omega}{2\pi}4|H(\omega)|^2 = 8R_{rad}KT\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}|H(\omega)|^2. \qquad (5.72)$$

Substituting for $\langle n^2(t)\rangle$ in (5.68), we obtain

$$\begin{aligned}\text{SNR} &= \frac{|Z_p|^2|h(t)*f(t-t_p)|^2}{\langle|n(t)|^2\rangle} = \frac{|Z_p|^2}{8R_{rad}KT}\frac{|\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}H(\omega)F(\omega)e^{-j\omega t_p}e^{j\omega t}|^2}{\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}|H(\omega)|^2}\\ &\le \frac{|Z_p|^2}{8R_{rad}KT}\frac{\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}|H(\omega)|^2\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}|F(\omega)|^2}{\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}|H(\omega)|^2} = \frac{|Z_p|^2}{8R_{rad}KT}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}|F(\omega)|^2,\end{aligned}$$

where we used the *Schwarz inequality* for the last step. Since the condition for the equality option in the same step is

$$H(\omega)\propto F^*(\omega) \qquad (5.73)$$

for $t=t_p$, we conclude that:

1. Matched filter $h(t)$ is specified (to an arbitrary multiplicative constant) by

$$h(t)=f^*(-t) \;\leftrightarrow\; H(\omega)=F^*(\omega), \qquad (5.74)$$

2. Peak SNR at the radar receiver output occurs at $t=t_p$ with a value of

$$\frac{|s(t_p)|^2}{\langle|n(t)|^2\rangle} = \frac{|Z_p|^2|\chi(0)|^2}{\langle|n(t)|^2\rangle} = \frac{|Z_p|^2}{8R_{rad}KT}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}|F(\omega)|^2 = \frac{|Z_p|^2}{4R_{rad}^2}\frac{\mathcal{E}_t}{KT} \qquad (5.75)$$

where

$$\mathcal{E}_t = \frac{1}{2}R_{rad}\int_{-\infty}^{\infty}|F(\omega)|^2\frac{d\omega}{2\pi} = \frac{1}{2}R_{rad}\int_{-\infty}^{\infty}dt|f(t)|^2 \qquad (5.76)$$

denotes the transmitted *energy* of radar pulse (since $f(t)$ is the instantaneous current envelope during pulse transmission).

Clearly, the peak SNR seems to depend on transmitted pulse energy $\mathcal{E}_t$ but not the shape of radar pulse $f(t)$. Therefore, two different $f(t)$'s depicted in Figures 5.12a and b with the same $\mathcal{E}_t$ would yield equal peak SNR's even though their *matched filter response*

$$\chi(t)=f^*(-t)*f(t) \qquad (5.77)$$

have dramatically different shapes. The pulse shown in Figure 5.12b is an example of a "coded" radar pulse, with the effect of coding being a "time compression" in $\chi(t)$ as compared to an uncoded pulse of equal duration. Pulse coding is a means of improving the radar resolution δR without reducing the peak SNR and an is active area of research. Further topics on pulse coding and the *radar ambiguity function*

$$\chi(t,\omega)\equiv f^*(-t)*f(t)e^{j\omega t} \qquad (5.78)$$

pertinent to range and Doppler resolutions of moving targets are covered in ECE 558.

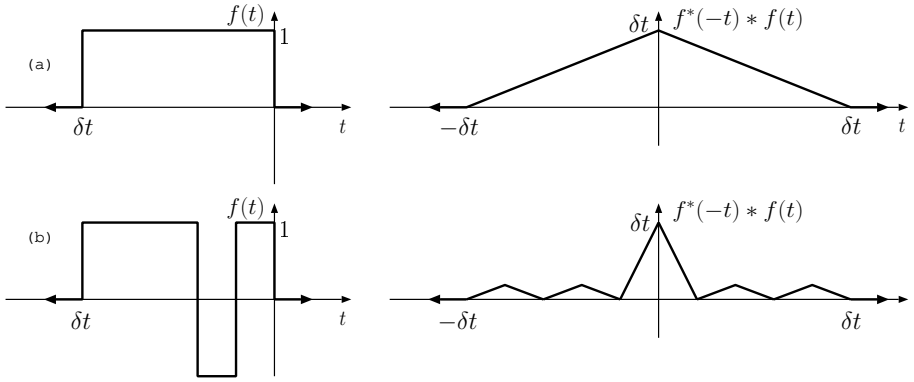


Figure 5.12: (a) A rectangular transmitter pulse $f(t)$ and the corresponding matched filter response $\chi(t) = f^*(-t) * f(t)$, and (b) a 5-baud Barker coded (+++-+) transmitter pulse $f(t)$ and matched filter response $f^*(-t) * f(t)$. Both signals have equal energy $\frac{1}{2}R_{rad}\int |f(t)|^2 dt$ and therefore yield equal peak SNR's. However, Barker coded pulse provides a better range resolution since the matched filter response $f^*(-t) * f(t)$ is compressed in range $\frac{ct}{2}$ with respect to the output due to the uncoded pulse.

5.4.2 Soft-target SNR

In the soft-target case with an open-circuit voltage response

$$V(t) = \int d\mathbf{r} Z(\mathbf{r}) f(t - \frac{2r}{c}) n(\mathbf{r}, t - \frac{r}{c}) e^{jkr} \tag{5.79}$$

we have

$$s(t) = h(t) * V(t) = \int d\mathbf{r} Z(\mathbf{r}) h(t) * f(t - \frac{2r}{c}) n(\mathbf{r}, t - \frac{r}{c}) e^{jkr} \tag{5.80}$$

at the receiver output accompanied by output noise $n(t)$. In (5.80) the impulse response $h(t)$ convolves the product $f(t - \frac{2r}{c}) n(\mathbf{r}, t - \frac{r}{c})$. But, if $n(\mathbf{r}, t)$ is slowly varying in t as compared to $h(t)$, the convolution can be carried only between $h(t)$ and $f(t - \frac{2r}{c})$ to obtain

$$s(t) \approx \int d\mathbf{r} Z(\mathbf{r}) \chi(t - \frac{2r}{c}) n(\mathbf{r}, t - \frac{r}{c}) e^{jkr}. \tag{5.81}$$

where $\chi(t) \equiv h(t) * f(t)$ once again.

Thus a model for $s(t)$ is obtained with $f(t) \to \chi(t)$ substitution in models for $V(t)$, and, furthermore, a model for $\langle |s(t)|^2 \rangle$ can be obtained from $\langle |V(t)|^2 \rangle$ using the same

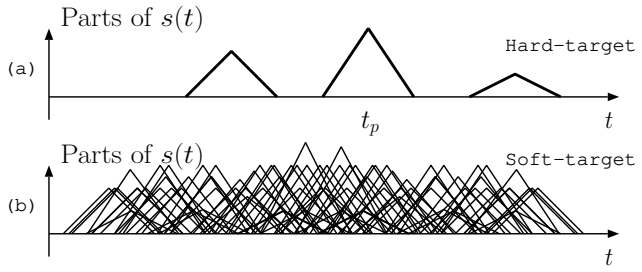


Figure 5.13: A cartoon showing the distinction between (a) hard targets and (b) soft-targets. In hard target returns $s(t) \propto$ a few distinct $\chi(t - t_p)$ responses, while for soft targets $s(t)$ is a superposition uncountably many $\chi(t - t_p)$ with small amplitude and random phased weights.

substitution. Thus, from (5.51) we obtain[13]

$$\langle |s(t)|^2 \rangle \approx 4R_{rad}^2 \int dr \int d\Omega \frac{G(\hat{r})A(\hat{r})}{(4\pi r)^2} |\chi(t - \frac{2r}{c})|^2 \sigma_v. \tag{5.82}$$

Dividing this result with noise variance $\langle |n(t)|^2 \rangle$, and remembering from (5.75) that in the matched filtered case

$$\frac{|\chi(0)|^2}{\langle |n(t)|^2 \rangle} = \frac{\mathcal{E}_t/KT}{4R_{rad}^2},$$

we find an SNR

$$\frac{\langle |s(t)|^2 \rangle}{\langle |n(t)|^2 \rangle} \approx \frac{\mathcal{E}_t}{KT} \int dr \int d\Omega \frac{G(\hat{r})A(\hat{r})}{(4\pi r)^2} \frac{|\chi(t - \frac{2r}{c})|^2}{|\chi(0)|^2} \sigma_v \tag{5.83}$$

for matched filtered soft-target radars. Comparing this result with a hard-target peak SNR

$$\frac{|s(t_p)|^2}{\langle |n(t)|^2 \rangle} = \frac{|Z_p|^2}{4R_{rad}^2} \frac{\mathcal{E}_t}{KT} = \frac{\mathcal{E}_t}{KT} \frac{G_p A_p}{(4\pi r_p^2)^2} \sigma_p \tag{5.84}$$

the following observations can be made:

1. Soft-target SNR $\propto r^{-2}$ while hard-target SNR $\propto r_p^{-4}$ — this difference is due to the fact that the number of individual targets within the radar beam increases with r^2 in the soft-target case.

[13] In cases when $n(\mathbf{r}, t)$ variation is not slow enough, it can be shown from (5.80) that the proper model is

$$\langle |s(t)|^2 \rangle \approx 4R_{rad}^2 \int \frac{d\omega}{2\pi} \int dr \int d\Omega \frac{G(\hat{r})A(\hat{r})}{(4\pi r)^2} |\chi(t - \frac{2r}{c}, \omega)|^2 \sigma(\omega),$$

where $\chi(t, \omega) \equiv h(t) * f(t)e^{j\omega t}$ is the radar ambiguity function. Note the extra integration over Doppler frequency in this revised model.

2. Hard-target SNR is independent of pulse shape $f(t)$ while that is not true for soft-target SNR — the factor

$$\int dr \frac{|\chi(t-\frac{2r}{c})|^2}{r^2|\chi(0)|^2} \approx \frac{1}{R^2}\int dr|\frac{\chi(t-\frac{2r}{c})}{\chi(0)}|^2 \equiv \frac{\delta R}{R^2}$$

above is clearly smaller for compressed pulses such as in Figure 5.12b as compared to uncompressed pulses. Thus pulse compression or coding reduces the soft target SNR with respect to uncoded pulses of equal length, but it is still worth using coded pulses instead of uncoded ones to achieve a given range resolution δR, because energy $\mathcal{E}_t$ will be larger for a coded long pulse than for an uncoded short pulse used to achieve the same resolution.

Exercise: For $\mathcal{E}_t = P_t\delta t$, where δt is the transmitted pulse length, and also for a range resolution of $\delta R \equiv \int dr|\frac{\chi(t-\frac{2r}{c})}{\chi(0)}|^2$ and mean range $R = ct/2$, show that the matched-filtered soft-target SNR (5.83) with a constant $\sigma_v(\mathbf{r}) = \sigma_v$can be simplified as

$$\frac{\langle|s(t)|^2\rangle}{\langle|n(t)|^2\rangle} \approx \frac{P_t\delta t A_{BS}\delta R}{4\pi R^2 KT}\sigma_v. \tag{5.85}$$

5.4.3 Soft-target detectability

Radar detectability is a figure of merit — the bigger the better — that has to do with the ability to recognize the presence of a non-zero signal variance $\langle|s|^2\rangle$ from a set of, say, I, independent data samples

$$y_i = s_i + n_i$$

at the receiver output representing some radar range. The components s_i and n_i of y_i are realizations of independent random variables and furthermore $|s_i|^2$ and $|n_i|^2$ are assumed to have exponential pdf's of the form

$$f(|n_i|^2) = \frac{e^{-\frac{|n_i|^2}{\langle|n|^2\rangle}}}{\langle|n|^2\rangle}$$

compatible with the soft-target signal statistic discussed in the next section.

Given the exponential pdf, the expected value of $|n_i|^2$ is the noise variance $\langle|n|^2\rangle$, and the rms deviation of $|n_i|^2$ from $\langle|n|^2\rangle$, i.e., $\sqrt{\langle(|n_i|^2 - \langle|n|^2\rangle)^2\rangle}$, is also $\langle|n|^2\rangle$ (see the Exercise problem below). However, the rms deviation of $\frac{1}{I}\sum_i |n_i|^2$ from its mean — again, it is $\langle|n|^2\rangle$ — is found to be $\frac{\langle|n|^2\rangle}{\sqrt{I}}$, which makes $\frac{1}{I}\sum_i |n_i|^2$ a better estimator of $\langle|n|^2\rangle$ than just a single $|n_i|^2$. The deviation is smaller in rms sense because of the cancellations between positive and negative deviations of the individual $|n_i|^2$ being averaged.

Consider next

$$\frac{1}{I}\sum_i |y_i|^2$$

as an estimator for $\langle|y|^2\rangle = \langle|s|^2\rangle + \langle|n|^2\rangle$. When a $\frac{1}{I}\sum_i |y_i|^2$ estimate exceeds $\langle|n|^2\rangle$ we will be tempted to conclude the presence of a non-zero backscattered signal variance

$\langle|s|^2\rangle$ in the data set. But, $\frac{1}{I}\sum_i |y_i|^2$ is likely to exceed $\langle|n|^2\rangle$ half the time even in the absence of any $\langle|s|^2\rangle$ because of the natural fluctuations of $\frac{1}{I}\sum_i |y_i|^2 = \frac{1}{I}\sum_i |n_i|^2$ with an rms value of $\frac{\langle|n|^2\rangle}{\sqrt{I}}$, known as *grass*.

Thus, in order to conclude the presence of a possible non-zero $\langle|s|^2\rangle$ within a $\frac{1}{I}\sum_i |y_i|^2$ estimate we need $\langle|y|^2\rangle$ to exceed $\langle|n|^2\rangle$ by many multiples of (grass) $\frac{\langle|n|^2\rangle}{\sqrt{I}}$. Writing out this condition as

$$\langle|y|^2\rangle - \langle|n|^2\rangle \gg \frac{\langle|n|^2\rangle}{\sqrt{I}},$$

we recognize that we need

$$D \equiv \sqrt{I}\frac{\langle|s|^2\rangle}{\langle|n|^2\rangle} \gg 1. \tag{5.86}$$

Parameter D defined above — the product of radar SNR with the square root of the number of independent data records for basing our decisions — is known as radar *detectability*. Number I is also known as data *integration length.* The required condition $D \gg 1$ indicates that it is not really the SNR, but rather the detectability, which has to be kept above unity in successful radar experiments. It is possible to have sufficient detectability even when SNR$<$1 if the integration length I is sufficiently long.

Given a radar observation period τ_{obs} and some signal decorrelation time τ_c for $s(t)$, it is possible to use an integration length of $I = \frac{\tau_{obs}}{\tau_c}$. Combining this value of I with the soft-target SNR (5.85) we find that

$$D = \sqrt{I}\text{SNR} \sim \sqrt{\frac{\tau_{obs}}{\tau_c}}\frac{P_t \delta t A_{BS}\delta R}{4\pi R^2 KT}\sigma_v.$$

To achieve this level of detectability the radar must be transmitting $\mathcal{E}_t$ Joules every τ_c seconds, putting out an average power of $P_{avg} \equiv \frac{\mathcal{E}_t}{\tau_c} = \frac{P_t \delta t}{\tau_c}$. Thus,

$$D \sim \sqrt{\frac{\tau_{obs}}{\tau_c}}\frac{P_{avg} A_{BS}\delta R}{4\pi R^2 KT\frac{1}{\tau_c}}\sigma_v \tag{5.87}$$

with the additional constraint that $\tau_c > \delta t$. Clearly, this result tells us that detectability goes *up* with the *power-aperture* product $P_{avg}A_{BS}$, and *down* with improved resolution — i.e., smaller τ_{obs} and δr. The ultimate figure of merit for any radar is its $P_{avg}A_{BS}$ product.

Exercise: Using the exponential pdf $p(P)$ given above, show that the expected value of P is $\langle P\rangle$. Also show that the expected value of $(P-\langle P\rangle)^2$ is $\langle P\rangle^2$, and hence the standard deviation equals $\langle P\rangle$.

5.4.4 Soft-target radar signal statistics

The noisy output of a soft-target radar receiver is in general a complex variable of the form

$$y \equiv I + jQ = \sum_p a_p e^{j\phi_p}$$

where I, Q, a_p, and ϕ_p are *by definition* real valued, a_p and ϕ_p are independent random variables for each p, ϕ_p has uniform pdf in $[0, 2\pi]$, and p runs over uncountably many contributions. It follows (see Exercise 1) that

$$\langle I \rangle = \langle Q \rangle = 0, \quad \langle I^2 \rangle = \langle Q^2 \rangle, \text{ and } \langle IQ \rangle = 0.$$

Furthermore, I and Q have Gaussian pdf's by the virtue of the Central Limit Theorem[14].

The joint pdf of I and Q — to be denoted as $f(y)$ as a short-hand for $f(I = Re\{y\}, Q = Im\{y\})$ — can therefore be expressed as (see Exercise 2)

$$f(y) = \frac{e^{-\frac{|y|^2}{\langle |y|^2 \rangle}}}{\pi \langle |y|^2 \rangle},$$

and an integral of $f(y)$ over the entire complex $y = |y|e^{j\phi}$ plane yields

$$\int_{\phi=0}^{2\pi} d\phi \int_{|y|=0}^{\infty} d|y| \; |y| f(y) = \int_{|y|=0}^{\infty} d|y| \frac{2|y| e^{-\frac{|y|^2}{\langle |y|^2 \rangle}}}{\langle |y|^2 \rangle} = \int_{|y|^2=0}^{\infty} d|y|^2 \frac{e^{-\frac{|y|^2}{\langle |y|^2 \rangle}}}{\langle |y|^2 \rangle} = 1.$$

This result indicates that:

1. The pdf of *magnitude* $|y|$ is given by Rayleigh distribution

$$f(|y|) = \frac{2|y| e^{-\frac{|y|^2}{\langle |y|^2 \rangle}}}{\langle |y|^2 \rangle} \quad \text{for } |y| \geq 0,$$

2. The pdf of *phase* ϕ is uniform in the range, say, $0 \leq \phi \leq 2\pi$,

3. Random variables $|y|$ and ϕ are statistically independent, and finally,

4. The pdf of squared magnitude $|y|^2$ is given by exponential distribution

$$f(|y|^2) = \frac{e^{-\frac{|y|^2}{\langle |y|^2 \rangle}}}{\langle |y|^2 \rangle} \quad \text{for } |y|^2 \geq 0.$$

Exercise 1: If $y \equiv I + jQ = \sum_p a_p e^{j\phi_p}$, where I and Q are real, show that $\langle I \rangle = \langle Q \rangle = 0$, $\langle I^2 \rangle = \langle Q^2 \rangle$, and $\langle IQ \rangle = 0$, provided that a_p and ϕ_p are independent random variables, ϕ_p has uniform pdf in $[0, 2\pi]$, and ϕ_p and $\phi_{q \neq p}$ are independent.

Exercise 2: For $y = I + jQ$, show that if $\langle I \rangle = \langle Q \rangle = 0$, $\langle I^2 \rangle = \langle Q^2 \rangle$, and $\langle IQ \rangle = 0$, then the zero-mean Gaussian expression given above for $f(y)$ reduces to the product of two zero-mean Gaussian pdfs in I and Q with variances $\langle I^2 \rangle$ and $\langle Q^2 \rangle$, respectively.

Exercise 3: For complex zero-mean random variable $y = |y|e^{j\phi}$ show that $\langle y^2 \rangle = \langle (y^*)^2 \rangle = 0$.

[14] This important theorem states that the probability density function of the sum of P independent random variables approaches a Gaussian when P is large. The mean and variance of the Gaussian are P times the mean and the variance of the individual random variables when they are identically distributed. For zero-mean, but non-identically distributed random variables with finite variances, the variance of the Gaussian equals the sum of the variances [e.g., *Helstrom*, 1991].

5.4.5 Spectral fitting and parameter estimation

Let

$$Y_q \equiv \sum_{n=0}^{N-1} y_n e^{-j\frac{2\pi nq}{N}}$$

denote the FFT of a radar receiver samples y_n representing some radar range gate. Since Y_q are linear combinations of zero-mean complex Gaussian random variables y_n, they are also described by Gaussian pdfs of the form

$$f(Y_q) = \frac{e^{-\frac{|Y_q|^2}{\langle |Y_q|^2\rangle}}}{\pi\langle |Y_q|^2\rangle},$$

where, by definition, $\langle |Y_q|^2\rangle$ is the frequency spectrum of y_n data.

Now, if Y_q and $Y_{p\neq q}$ were independent random variables, the joint pdf of all Y_q could be written as

$$f(Y_{-N/2}, \cdots Y_{N/2-1}) = \frac{e^{-\sum_{q=-N/2}^{N/2-1}\frac{|Y_q|^2}{\langle |Y_q|^2\rangle}}}{\pi^N \prod_{q=-N/2}^{N/2-1}\langle |Y_q|^2\rangle}.$$

Using this joint pdf expression together with the measured periodogram $|Y_q|^2$ it is possible to infer the most likely values of scattering medium parameters which control the form of spectrum $\langle |Y_q|^2\rangle$. Say, $\langle |Y_q|^2\rangle$ depends on three parameters A, B, and C, via known dependencies $\langle |Y_q|^2\rangle = f_q(A, B, C)$. The maximum-likelihood estimates[15] of these parameters can be obtained from the radar data by finding out which values of A, B, and C maximize the joint pdf when it is evaluated with the measured periodogram $|Y_q|^2$ and $f_q(A, B, C)$ values for the spectrum $\langle |Y_q|^2\rangle$. An equivalent procedure is the minimization of

$$-\ln f(Y_{-N/2}, \cdots Y_{N/2-1}) = \sum_{q=-N/2}^{N/2-1} \frac{|Y_q|^2 + \langle |Y_q|^2\rangle \ln(\pi\langle |Y_q|^2\rangle)}{\langle |Y_q|^2\rangle}$$

over parameters A, B, and C.

A comment about the independence of random variables Y_q and $Y_{p\neq q}$: This assumption is not universally valid. However, when spectrum $\langle |Y_q|^2\rangle$ is broad or, alternatively, under low SNR conditions the assumption is very good. Since the consequences of non-optimal parameter estimation are not critical when the SNR is high it makes sense to use the method described above under all types of conditions. Unfortunately the described maximization and minimization tasks are computationally expensive and therefore an alternate method described below is often used.

Suppose the radar periodogram $|Y_q|^2$ is integrated over I independent measurements and the result, referred to as radar spectrum, is denoted by S_q. We then have $\langle S_q\rangle =$

[15] For a detailed reading on maximum-likelihood parameter estimation see *Menke* [1989].

$I\langle|Y_q|^2\rangle$ and $\delta S_{q,rms} = \sqrt{I}\langle|Y_q|^2\rangle = \langle S_q\rangle/\sqrt{I}$. For large I, the Central Limit Theorem can be invoked to argue for a Gaussian pdf for S_q of the form

$$f(S_q) = \frac{e^{-\frac{(S_q-\langle S_q\rangle)^2}{2\langle S_q\rangle^2/I}}}{\sqrt{2\pi}\langle S_q\rangle/\sqrt{I}}.$$

Therefore,

$$-\ln f(S_q) = \frac{I(S_q-\langle S_q\rangle)^2}{2\langle S_q\rangle^2} + \ln(\frac{\sqrt{2\pi}\langle S_q\rangle}{\sqrt{I}}) \xrightarrow[\text{large } I]{} \frac{I}{2}\frac{(S_q-\langle S_q\rangle)^2}{\langle S_q\rangle^2},$$

and

$$-\ln f(S_{-N/2},\cdots,S_{N/2-1}) \approx \frac{I}{2}\sum_{q=-N/2}^{N/2-1}(\frac{S_q}{\langle S_q\rangle}-1)^2 \equiv \frac{1}{2}\chi^2.$$

The minimization of

$$\chi^2 = I\sum_{q=-N/2}^{N/2-1}(\frac{S_q}{\langle S_q\rangle}-1)^2$$

over the parameters controlling $\langle S_q\rangle = I\langle|Y_q|^2\rangle$ provides their approximate maximum likelihood (ML) estimates when I is large. The advantage of this approach is that efficient algorithms for χ^2 minimization already exist (e.g., Marquardt-Levenberg). Notice that $\langle\chi^2\rangle \sim N$, and therefore minimization algorithms can terminate their search as $\chi^2 \to N$.

There is an alternate way of writing χ^2 and interpreting the parameter estimation procedure: For sufficiently large I the relative rms error of $\delta S_q = S_q - \langle S_q\rangle$ will be sufficiently small so that

$$\ln S_q = \ln(\langle S_q\rangle + \delta S_q) = \ln\langle S_q\rangle + \ln(1+\frac{\delta S_q}{\langle S_q\rangle}) \approx \ln\langle S_q\rangle + \frac{\delta S_q}{\langle S_q\rangle}.$$

Thus,

$$\ln S_q - \ln\langle S_q\rangle \approx \frac{\delta S_q}{\langle S_q\rangle} = \frac{S_q}{\langle S_q\rangle} - 1,$$

and

$$\chi^2 \approx I\sum_{q=-N/2}^{N/2-1}(\ln S_q - \ln\langle S_q\rangle)^2.$$

Therefore χ^2 minimization can be viewed as unweighted least-squares fitting of the logarithm of the integrated radar spectrum "curve" S_q to the logarithm of its expected value. General nonlinear curve fitting algorithms can be employed to do the parameter estimation task.

Finally, we point out that in ionospheric incoherent scatter applications the model equation for $\langle S_q\rangle = I\langle|Y_q|^2\rangle$ can be formulated in terms of ionospheric state parameters including the particle temperatures and densities as well as drift velocities and ionic

compositions. In radar backscatter from unstable ionosphere or turbulent neutral atmosphere that is not possible and the focus to motion measurements. In such applications a *generalized* Gaussian model such as

$$\langle S_q \rangle = Ae^{-|\frac{q-\hat{q}}{Q}|^p} + B, \;\; q \in [-N/2, N/2-1]$$

is often adequate. In this model B denotes a spectral noise floor, $\hat{q}$ is the mean Doppler frequency shift normalized by $\Delta\omega = \frac{2\pi}{N\Delta t}$, whereas A, Q, and p describe the distribution of the backscattered signal power over the Doppler frequency space.

5.5 Lidar detection

Lidar is an acronym for *li*ght *d*etection *a*nd *r*anging and it is used to refer to systems analogous to backscatter radars operated in the optical frequency range. The lidar transmitter is a pulsed laser operating at a frequency ν. The receiving system consists of an optical telescope with some aperture area A to collect the backscattered light from the atmosphere and route it to a photon counter. The photon counter is the lidar receiver. In the absence of atmospheric motions the backscattered light from the atmosphere will consist of photons with an average energy $h\nu$, where h is Planck's constant.

In view of the discussion of incoherent soft-target radars covered in the previous section we can write the expected photon count rate of a lidar system as

$$\langle \frac{P_r}{h\nu} \rangle = \frac{P_t}{h\nu} \frac{A\delta r}{4\pi r^2} \sigma n,$$

where P_t is the average power of the transmitted laser pulse, $\delta r = c\delta t/2$ is the depth of the lidar scattering volume (lidar resolution), $r = ct/2$ is the lidar range, σ is the backscattering cross-section of the atmospheric constituent(s) that can scatter photons of frequency ν, and n is the mean number density of the constituent(s) in the atmosphere.

Suppose that the incoming photons are counted with some counting efficiency η_c over a duration δt equal to the pulse length, and the resulting count is denoted by C_r. Multiplying the count rate expression above by $\eta_c \delta t$ and using $\delta r = c\delta t/2$ we obtain the lidar equation

$$\langle C_r \rangle = P_t \frac{\eta_c A c \delta t^2}{4\pi r^2 2h\nu} \sigma n$$

that relates the expected value of photon count per scattering volume to the number density n of the scattering particles in the atmosphere.

In lidar applications external noise originates from background light collected by the lidar telescope in addition to the backscattered light in response to the laser pulse. During daytime external noise due to scattered sunlight is blindingly strong and makes lidar operations extremely difficult. By contrast, at nighttime external noise due to scattered starlight and atmospheric airglow[16] is generally weak enough that it can be neglected in

[16] Weak natural optical emissions from atmospheric atoms and molecules returning from excited states to lower energy states.

comparison to internal noise. Internal noise is by definition the number of photons C_d counted in the interval δt with zero input from the lidar telescope.

There is a third type of noise in lidar applications which is known as *self-noise* or shot-noise. These terms refer to random deviations of photon count C_r from its mean or expected value $\langle C_r \rangle$. The lidar scattering volume V_s contains $N_v = nV_s$ particles, and, therefore, according to lidar equation each particle contributes $\langle C_r \rangle / N_v$ photons on the average to the final count. Since $\langle C_r \rangle / N_v \ll 1$ (typical numbers for $\langle C_r \rangle$ and N_v could be $\sim 10^2 - 10^4$ and $\sim 10^{20} - 10^{30}$, respectively), not every particle in the volume gets a chance to make a contribution to C_r. If we were to assume that each particle contributes either 0 or 1 photon with probabilities of $1 - P$ and $P = \langle C_r \rangle / N_v \ll 1$, respectively, then the probability that a total of C_r photons are counted from the entire volume of N_v particles would be given by the binomial probability distribution

$$p_{N_v}(C_r) = \frac{N_v!}{C_r!(N_v - C_r)!} P^{C_r} (1 - P)^{(N_v - C_r)}.$$

For $N_v \gg 1$ and $P = \langle C_r \rangle / N_v \ll 1$ binomial distribution reduces to Poisson distribution

$$p(C_r) = \frac{(N_v P)^{C_r}}{C_r!} e^{-N_v P} = \frac{\langle C_r \rangle^{C_r}}{C_r!} e^{-\langle C_r \rangle}.$$

Therefore lidar photon count C_r is a Poisson random variable, and it deviates from its mean or expected value $\langle C_r \rangle$ by an amount $\delta C_r \equiv C_r - \langle C_r \rangle$ with an rms value $\delta C_{r,rms} = \langle (C_r - \langle C_r \rangle)^2 \rangle^{1/2} = \langle C_r \rangle^{1/2}$. Note that, by definition, $\delta C_{r,rms}$ is the *standard deviation* of random variable C_r and $\delta C_{r,rms}^2$ is the corresponding *variance*.

Lidar system parameters are usually selected such that self-noise $\delta C_{r,rms} = \langle C_r \rangle^{1/2} \gg C_d$. Thus, during nighttime lidar experiments the dominant noise is self-noise and internal noise can be neglected. Also, photon counts of independent lidar pulse returns can be summed up to reduce the relative strength of self-noise. Let C_I denote such a sum, a sum of I independent and identically distributed (IID) random variables C_r. The mean and the variance of a sum of IID random variables can be calculated by simply summing up the means and the variances, respectively. Therefore, $\langle C_I \rangle = I \langle C_r \rangle$, variance $\delta C_{I,rms}^2 = I \delta C_{r,rms}^2 = I \langle C_r \rangle = \langle C_I \rangle$, and standard deviation $\delta C_{I,rms} = \langle C_I \rangle^{1/2}$. Lidar SNR is then defined as the ratio of the expected value of the integrated count C_I and its standard deviation $\delta C_{I,rms}$, and is given by the formula

$$\text{SNR} = \frac{\langle C_I \rangle}{\delta C_{I,rms}} = \langle C_I \rangle^{1/2} = \sqrt{I} \langle C_r \rangle^{1/2} = \sqrt{I P_t \frac{\eta_c A c \delta t^2}{4\pi r^2 2h\nu} \sigma n}. \tag{5.88}$$

Clearly, SNR increases with integration length I, but of course at the cost of time resolution.

Exercise 1: In a lidar experiment I laser pulses are emitted and the corresponding I photon counts $C_{r,i}$, $i \in [0, I-1]$ are added in order to obtain an estimate of the particle density n in the scattering volume. Assuming that internal noise $C_d = 0$, show that the rms error in the estimate for n equals n/SNR. This result shows why it is desirable to increase

lidar SNR by using long integration lengths I. Hint: Begin with the assumption that the estimate gives $n + \delta n$ where δn is the estimation error. Then find the relation between the rms values of δn and self-noise in the integrated photon count.

Exercise 2: In a certain lidar experiment with no integration (i.e., I=1) it is found that SNR=40 for returns from a scattering volume at some range $R = R_1$. It is known that the number density n in the scattering volume at range $R = 2R_1$ is only 10% of the density at $R = R_1$. How long should the counts from $R = 2R_1$ be integrated so that SNR=10 for the returns detected from $R = 2R_1$?

6 Magnetoionic theory

We return in this chapter to ionospheric propagation and examine the impact of the Earth's DC magnetic field $\mathbf{B}_o$ on the propagation of radiowaves. We will find out that a non-zero $\mathbf{B}_o$ causes the plasma conductivity to take a tensor character — $\mathbf{J}$ is no longer parallel to its driver $\mathbf{E}$ — which in turn causes the medium to be anisotropic. In Section 1 we will take a general look at wave propagation in anisotropic media and subsequently focus our efforts on anisotropic propagation effects in a "magnetoplasma" such as the ionosphere.

6.1 Plane waves in anisotropic media

Consider Faraday's and Ampere's laws in plane wave form,

$$-j\mathbf{k}\times\mathbf{E} = -j\omega\mathbf{B}$$

and

$$-j\mathbf{k}\times\mathbf{H} = \mathbf{J} + j\omega\mathbf{D}.$$

Assuming $\mathbf{B} = \mu_o\mathbf{H}$, but $\mathbf{D} = \bar{\epsilon}\cdot\mathbf{E}$ and $\mathbf{J} = \bar{\sigma}\cdot\mathbf{E}$, where $\bar{\epsilon}$ and $\bar{\sigma}$ are permittivity and conductivity matrices, the equations take the form

$$\mathbf{k}\times\mathbf{E} = \omega\mu_o\mathbf{H}$$

and

$$\mathbf{k}\times\mathbf{H} = -\omega\epsilon_o(\frac{\bar{\epsilon}}{\epsilon_o} - j\frac{\bar{\sigma}}{\omega\epsilon_o})\cdot\mathbf{E},$$

respectively. Notice that in isotropic media where $\bar{\epsilon}$ and $\bar{\sigma}$ are replaced by scalars ϵ and σ, Ampere's law reduces to

$$\mathbf{k}\times\mathbf{H} = -\omega\epsilon_o n^2\mathbf{E}$$

where

$$n^2 \equiv \sqrt{\epsilon_r - j\frac{\sigma}{\omega\epsilon_o}}.$$

As shown in the next example it is possible to express Ampere's law in an anisotropic medium in a similar format, like in

$$\mathbf{k}\times\mathbf{H} = -\omega\epsilon_o\sum_{m=-1}^{1} n_m^2\mathbf{E}_m.$$

Example 1: In some non-conducting crystal it is known that

$$\begin{pmatrix} D_x \\ D_y \\ D_z \end{pmatrix} = \begin{pmatrix} \epsilon_o & 0 & 0 \\ 0 & 4\epsilon_o & 0 \\ 0 & 0 & 9\epsilon_o \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix}.$$

Express plane wave form of Ampere's law explicitly in terms of E_x, E_y, E_z.

Solution: Since the medium is non-conducting $\mathbf{J} = 0$. Thus Ampere's law is

$$-j\mathbf{k} \times \mathbf{H} = j\omega\mathbf{D}$$

with $\mathbf{D} = (D_x, D_y, D_z) = (\epsilon_o E_x, 4\epsilon_o E_y, 9\epsilon_o E_z) = \epsilon_o(E_x\hat{x} + 4E_y\hat{y} + 9E_z\hat{z})$. This leads to

$$\mathbf{k} \times \mathbf{H} = -\omega\epsilon_o(E_x\hat{x} + 4E_y\hat{y} + 9E_z\hat{z})$$

as the requested expression.

To proceed we will assume that in some coordinate system described by orthonormal unit vectors $\hat{u}_m$, $m \in [-1, 0, 1]$, the matrix

$$\frac{\bar{\epsilon}}{\epsilon_o} - j\frac{\bar{\sigma}}{\omega\epsilon_o}$$

in Ampere's law takes a diagonal form with diagonal elements n_m^2. In that case Ampere's law can be expressed as

$$\mathbf{k} \times \mathbf{H} = -\omega\epsilon_o \sum_{m=-1}^{1} n_m^2 E_m \hat{u}_m$$

(as in Example 1), where scalars E_m are the projections of vector $\mathbf{E}$ along unit vectors $\hat{u}_m$, i.e.,

$$\mathbf{E} = \sum_{m=-1}^{1} E_m \hat{u}_m$$

and $E_m = \hat{u}_m^* \cdot \mathbf{E}$. Combining Ampere's law with Faraday's law

$$\mathbf{k} \times \mathbf{E} = \sum_{m=-1}^{1} E_m \mathbf{k} \times \hat{u}_m = \omega\mu_o\mathbf{H},$$

we obtain

$$\sum_{m=-1}^{1} E_m \underbrace{\mathbf{k} \times \mathbf{k} \times \hat{u}_m}_{(\hat{u}_m \cdot \mathbf{k})\mathbf{k} - k^2\hat{u}_m} = -\underbrace{\omega^2\mu_o\epsilon_o}_{\frac{\omega^2}{c^2}} \sum_{m=-1}^{1} n_m^2 E_m \hat{u}_m,$$

which implies that

$$\mathbf{k}(\mathbf{k} \cdot \sum_{m=-1}^{1} E_m \hat{u}_m) = \sum_{m=-1}^{-1} (k^2 - \frac{\omega^2}{c^2} n_m^2) E_m \hat{u}_m,$$

or, equivalently,

$$\mathbf{k}(\mathbf{k}\cdot\mathbf{E}) = \sum_{m=-1}^{-1}(k^2 - \frac{\omega^2}{c^2}n_m^2)E_m\hat{u}_m.$$

The equation above is the counterpart of

$$\mathbf{k}(\mathbf{k}\cdot\mathbf{E}) = (k^2 - \frac{\omega^2}{c^2}n^2)\mathbf{E}$$

obtained for isotropic media in Chapter 2 with $n^2 = \epsilon_r - j\frac{\sigma}{j\omega}$. The implication of the equation was that plane TEM waves with $\mathbf{k}\cdot\mathbf{E} = 0$ satisfy the dispersion relation

$$k^2 - \frac{\omega^2}{c^2}n^2 = 0,$$

while plane space-charge waves for which $\mathbf{k}\cdot\mathbf{E} \neq 0$ satisfy $k^2(\mathbf{k}\cdot\mathbf{E}) = (k^2 - \frac{\omega^2}{c^2}n^2)(\mathbf{k}\cdot\mathbf{E})$, i.e., $n^2 = 0$.

Likewise, in an anisotropic medium where

$$\mathbf{k}(\mathbf{k}\cdot\mathbf{E}) = \sum_{m=-1}^{-1}(k^2 - \frac{\omega^2}{c^2}n_m^2)E_m\hat{u}_m,$$

plane TEM waves with $\mathbf{k}\cdot\mathbf{E} = 0$ must satisfy

$$\sum_{m=-1}^{-1}(k^2 - \frac{\omega^2}{c^2}n_m^2)E_m\hat{u}_m = 0$$

whereas the dispersion relation for non-TEM waves can be obtained as follows: $\hat{u}_m\cdot\hat{u}_m^* = 1$ and $\hat{u}_m\cdot\hat{u}_{k\neq m}^* = 0$ imply that

$$\mathbf{k}\cdot\hat{u}_m^*(\mathbf{k}\cdot\mathbf{E}) = (k^2 - \frac{\omega^2}{c^2}n_m^2)E_m,$$

or, equivalently,

$$E_m = \frac{\mathbf{k}\cdot\hat{u}_m^*(\mathbf{k}\cdot\mathbf{E})}{k^2 - \frac{\omega^2}{c^2}n_m^2}.$$

Thus,

$$\mathbf{k}\cdot\mathbf{E} = \mathbf{k}\cdot\sum_{m=-1}^{1}\hat{u}_mE_m = \mathbf{k}\cdot\sum_{m=-1}^{1}\hat{u}_m\frac{\mathbf{k}\cdot\hat{u}_m^*(\mathbf{k}\cdot\mathbf{E})}{k^2 - \frac{\omega^2}{c^2}n_m^2},$$

which implies, for $\mathbf{k}\cdot\mathbf{E} \neq 0$, the dispersion relation

$$1 = \sum_{m=-1}^{1}\frac{|\mathbf{k}\cdot\hat{u}_m|^2}{k^2 - \frac{\omega^2}{c^2}n_m^2}$$

for non-TEM waves. The same expression can also be written as

$$1 = \sum_{m=-1}^{1} \frac{n^2}{n^2 - n_m^2} |\hat{k} \cdot \hat{u}_m|^2,$$

where $n \equiv \frac{c}{\omega/k}$ denotes the refractive index of plane waves as usual. The constraint on TEM waves obtained earlier on can also be written in terms of n as

$$\sum_{m=-1}^{-1} (n^2 - n_m^2) E_m \hat{u}_m = 0.$$

Finally, the *polarization* of non-TEM waves are determined by

$$E_m = \frac{n^2}{n^2 - n_m^2} \hat{k} \cdot \hat{u}_m^* (\hat{k} \cdot \mathbf{E})$$

ratios for $m = 0, \pm 1$.

The next sequence of examples illustrate the use of these results.

Example 2: Determine the plane TEM waves and their propagation velocities in the medium introduced in Example 1 if the propagation direction is $\hat{k} = \hat{z}$.

Solution: Since

$$\begin{pmatrix} D_x \\ D_y \\ D_z \end{pmatrix} = \begin{pmatrix} \epsilon_o & 0 & 0 \\ 0 & 4\epsilon_o & 0 \\ 0 & 0 & 9\epsilon_o \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix},$$

we are free to choose $\hat{u}_1 = \hat{x}$, $\hat{u}_{-1} = \hat{y}$, and $\hat{u}_0 = \hat{z}$ so that $n_1 = 1$, $n_{-1} = 2$, and $n_0 = 3$. For $\hat{k} = \hat{z} = \hat{u}_0$ the only TEM possibilities are $\mathbf{E} = E_1 \hat{x}$ with $n = n_1$ and $\mathbf{E} = \mathbf{E}_{-1} \hat{y}$ with $n = n_{-1}$. The x-polarized TEM wave propagates with a phase velocity $v_p = \frac{c}{n} = c$ while the y-polarized one with $v_p = \frac{c}{n} = \frac{c}{2}$.

Example 3: What is the propagation velocity of of a wave in the medium considered in Example 2 if the propagation direction is $\hat{z}$ and the polarization direction is $\hat{x} + \hat{y}$?

Solution: As we found out in Example 2 only $\hat{x}$ and $\hat{y}$ polarized waves are allowed to propagate in the medium in $\hat{z}$ direction. Hence, a $\hat{x} + \hat{y}$ polarized wave is not a possibility — $\hat{x}$ and $\hat{y}$ components of the wave field will propagate independently with different wave speeds meaning that the polarization will vary with propagation distance z.

Example 4: Determine the refractive index of a plane wave solutions propagating in the direction $\hat{k} = \frac{1}{2}\hat{x} + \frac{\sqrt{3}}{2}\hat{y}$ in the medium considered in Example 2.

Solution: Once again $\hat{u}_1 = \hat{x}$, $\hat{u}_{-1} = \hat{y}$, and $\hat{u}_0 = \hat{z}$ so that $n_1 = 1$, $n_{-1} = 2$, and $n_0 = 3$. Clearly, one mode of propagation in the medium in direction $\hat{k} = \frac{1}{2}\hat{x} + \frac{\sqrt{3}}{2}\hat{y}$ is the $\hat{u}_0 = \hat{z}$-polarized TEM wave with the refractive index $n = n_0 = 3$. To determine the refractive index of the second mode allowed we evaluate

$$1 = \sum_{m=-1}^{1} \frac{n^2}{n^2 - n_m^2} |\hat{k} \cdot \hat{u}_m|^2$$

with $\hat{k}\cdot\hat{u}_1=\frac{1}{2}$, $\hat{k}\cdot\hat{u}_{-1}=\frac{\sqrt{3}}{2}$, and $\hat{k}\cdot\hat{u}_0=0$; the result

$$\frac{n^2}{4(n^2-1)}+\frac{3n^2}{4(n^2-4)}=1 \;\;\Rightarrow\;\; n^2(n^2-4)+3n^2(n^2-1)=4(n^2-1)(n^2-4)$$

has a solution $n=\sqrt{\frac{16}{13}}\approx 1.11$. An alternate solution $n=-\sqrt{\frac{16}{13}}$ describes the same mode propagating in $-\hat{k}$ direction.

Example 5: Determine the polarization of the second propagation mode identified in Example 4.

Solution: From Example 4, $\hat{k}=\frac{1}{2}\hat{x}+\frac{\sqrt{3}}{2}\hat{y}$, $\hat{u}_1=\hat{x}$, $\hat{u}_{-1}=\hat{y}$, $\hat{u}_0=\hat{z}$, $n_1=1$, $n_{-1}=2$, $n_0=3$, and $n=\sqrt{\frac{16}{13}}$ for the allowed mode. Thus the polarization relation

$$E_m=\frac{n^2}{n^2-n_m^2}\hat{k}\cdot\hat{u}_m^*(\hat{k}\cdot\mathbf{E})$$

indicates that $\frac{E_1}{E_0}=\frac{E_x}{E_z}=\infty$ and $\frac{E_{-1}}{E_0}=\frac{E_y}{E_z}=\infty$ since $\hat{k}\cdot\hat{u}_0^*=\hat{k}\cdot\hat{z}=0$. Therefore $E_z=0$. On the other hand, the ratio

$$\frac{E_y}{E_x}=\frac{E_{-1}}{E_1}=\frac{\hat{k}\cdot\hat{u}_{-1}^*}{\hat{k}\cdot\hat{u}_1^*}\frac{n^2-n_1^2}{n^2-n_{-1}^2}=\frac{\sqrt{3}/2}{1/2}\frac{\frac{16}{13}-1}{\frac{16}{13}-4}=\sqrt{3}\frac{16-13}{16-52}\approx -0.144.$$

Hence, $\mathbf{E}=\hat{x}E_x+\hat{y}E_y=E_x(\hat{x}-0.144\hat{y})$, and the mode is $\hat{p}$-polarized where

$$\hat{p}=\frac{\hat{x}-0.144\hat{y}}{\sqrt{1+0.144^2}}\approx 0.99\hat{x}-0.1425\hat{y}.$$

Notice that $\hat{p}$ is not orthogonal to the wave normal $\hat{k}=\frac{1}{2}\hat{x}+\frac{\sqrt{3}}{2}\hat{y}$.

Example 6: In some propagation medium $\mathbf{D}=\epsilon_o\mathbf{E}$, but

$$\begin{pmatrix} J_x\\ J_y\\ J_z\end{pmatrix}=\sigma_o\begin{pmatrix}2&0&0\\0&2&0\\0&0&3\end{pmatrix}\begin{pmatrix}E_x\\E_y\\E_z\end{pmatrix},$$

where $\sigma_o=-j\omega\epsilon_o X$. Given that $\hat{u}_1=\hat{x}$, $\hat{u}_2=\hat{y}$, and $\hat{u}_0=\hat{z}$, determine $n_{\pm1}^2$ and n_0^2.

Solution: Clearly, $\epsilon_r=1$, $\sigma_{\pm1}=2\sigma_o$, and $\sigma_0=3\sigma_o$. Thus it follows that

$$n_m^2=\epsilon_r-j\frac{\sigma_m}{\omega\epsilon_o}=1-j\frac{\sigma_m}{\omega\epsilon_o},$$

which yields $n_{\pm1}^2=1-2X$ and $n_0^2=1-3X$.

6.2 Magnetoionic theory

In a magnetoplasma we are allowed to take μ and ϵ as μ_o and ϵ_o as long as $\mathbf{J}$ — and ρ, if needed — takes into account all the charge carriers in the medium. Thus, Ampere's law can be written as

$$-j\mathbf{k}\times\mathbf{H}=\mathbf{J}+j\omega\mathbf{D}=\sum_{m=-1}^{1}\sigma_m E_m\hat{u}_m+j\omega\epsilon_o\mathbf{E}=j\omega\epsilon_o\sum_{m=-1}^{1}n_m^2E_m\hat{u}_m,$$

where

$$n_m^2 = 1 + \frac{\sigma_m}{j\omega\epsilon_o}$$

and σ_m denote diagonal elements of a diagonal conductivity matrix $\bar{\sigma}$ realized in some coordinate system with orthonormal unit vectors $\hat{u}_m$. Consequently, the results of previous section can be utilized directly once σ_m and the associated $\hat{u}_m$ are identified. As we will see below we can interpret σ_m as the *eigenvalues* of $\bar{\sigma}$ and $\hat{u}_m$ as the corresponding *eigenvectors*.

To identify $\bar{\sigma}$ for a magnetoplasma permeated by a DC magnetic field $\mathbf{B}_o$ we start with Newton's 2nd law

$$m\frac{d\mathbf{v}}{dt} = -e(\mathbf{E} + \mathbf{v} \times \mathbf{B}_o)$$

for a free electron where on the right we see the Lorentz force expression. In phasor form

$$mj\omega\mathbf{v} = -e(\mathbf{E} + \mathbf{v} \times \mathbf{B}_o),$$

or

$$\mathbf{v} + \frac{e}{mj\omega}\mathbf{v} \times \mathbf{B}_o = -\frac{e}{mj\omega}\mathbf{E}.$$

Next, multiplying both sides by $-eN$,

$$\mathbf{J} + \frac{e}{mj\omega}\mathbf{J} \times \mathbf{B}_o = \frac{Ne^2}{mj\omega}\mathbf{E} = -j\omega\epsilon_o\frac{\omega_p^2}{\omega^2}\mathbf{E},$$

where ω_p is the plasma frequency and $\mathbf{J} = -eN\mathbf{v}$ as usual. Assuming

$$\mathbf{B}_o = B_o\hat{z},$$

the same result can also be written as

$$\mathbf{J} - jY\mathbf{J} \times \hat{z} = -j\omega\epsilon_o X\mathbf{E},$$

where

$$X \equiv \frac{\omega_p^2}{\omega^2}$$

and

$$Y \equiv \frac{\Omega}{\omega},$$

with

$$\Omega = \frac{e|\mathbf{B}_o|}{m} = \frac{eB_o}{m}.$$

Postponing a detailed discussion of Ω — which is the radian the frequency of circular motions of free electrons about a DC field $\mathbf{B}_o$ — we note that a magnetoplasma cannot be described by a scalar conductivity parameter because of the second term in

$$\mathbf{J} - jY\mathbf{J} \times \hat{z} = -j\omega\epsilon_o X\mathbf{E}$$

on the left.

However, the second term is zero if $\mathbf{J}$ is in $\hat{z}$ direction, in which case

$$\mathbf{J} = \bar{\sigma} \cdot \mathbf{E} = -j\omega\epsilon_o X\mathbf{E},$$

and thus $\hat{u}_0 \equiv \hat{z}$ is one of the eigenvalues of conductivity matrix $\bar{\sigma}$ associated with the eigenvalue $\sigma_0 \equiv -j\omega\epsilon_o X$. To identify the remaining eigenvalues/eigenvectors of we $\bar{\sigma}$ replace $\mathbf{J} = \bar{\sigma} \cdot \mathbf{E}$ and $\mathbf{E}$ in

$$\mathbf{J} - jY\mathbf{J} \times \hat{z} = -j\omega\epsilon_o X\mathbf{E} = \sigma_0\mathbf{E}$$

with $\sigma_m\hat{u}_m$ and $\hat{u}_m$, respectively, to obtain a constraint

$$\sigma_m(\hat{u}_m - jY\hat{u}_m \times \hat{z}) = \sigma_0\hat{u}_m.$$

The constraint is satisfied by

$$\hat{u}_m = \frac{\hat{x} - mj\hat{y}}{\sqrt{2}}$$

and

$$\sigma_m = \frac{\sigma_0}{1 - mY}$$

for $m = \pm 1$, since, as it can be easily verified,

$$-j\hat{u}_m \times \hat{z} = \frac{-j\hat{x} - m\hat{y}}{\sqrt{2}} \times \hat{z} = \frac{j\hat{y} - m\hat{x}}{\sqrt{2}} = -m\frac{-mj\hat{y} + \hat{x}}{\sqrt{2}} = -m\hat{u}_m.$$

In summary, then, the complete set of orthonormal eigenvectors of the conductivity matrix $\bar{\sigma}$ of a magnetized plasma with $\mathbf{B}_o = B_o\hat{z}$ is

$$\hat{u}_0 = \hat{z} \quad \text{and} \quad \hat{u}_m = \frac{\hat{x} - mj\hat{y}}{\sqrt{2}}, \quad m = \pm 1,$$

with the corresponding eigenvalues

$$\sigma_m \equiv -j\omega\epsilon_o\frac{X}{1 - mY},$$

where $X = \omega_p^2/\omega^2$ and $Y = \Omega/\omega$. Thus, for the magnetized plasma,

$$n_m^2 = 1 - j\frac{\sigma_m}{\omega\epsilon_o} = 1 - \frac{X}{1 - mY},$$

and, in specific,

$$n_0 = \sqrt{1 - X}, \quad n_1 = \sqrt{1 - \frac{X}{1 - Y}}, \quad n_{-1} = \sqrt{1 - \frac{X}{1 + Y}}.$$

The parameters n_0, n_1, and n_{-1} obtained above happen to be the refractive index n of the allowed set of TEM waves in a magnetoplasma satisfying the constraint

$$\sum_{m=-1}^{-1} (n^2 - n_m^2)E_m\hat{u}_m = 0;$$

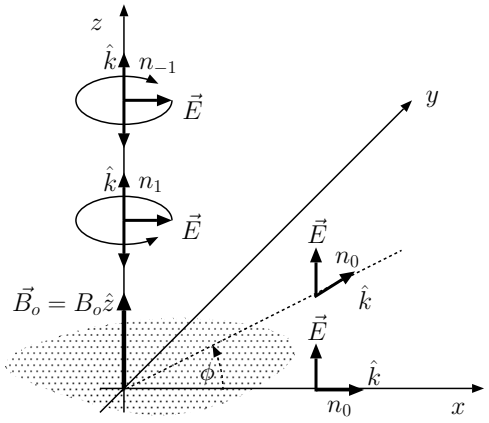


Figure 6.1: A sketch showing the three allowed modes of TEM propagation in a plasma with a DC magnetic field $\mathbf{B}_o = \hat{z}B_o$, $B_o > 0$.

these characteristic refractive index parameters also enable the computation of the refractive index n and polarization of non-TEM modes via relations

$$1 = \sum_{m=-1}^{1} \frac{n^2}{n^2 - n_m^2} |\hat{k} \cdot \hat{u}_m|^2,$$

and

$$E_m = \frac{n^2}{n^2 - n_m^2} \hat{k} \cdot \hat{u}_m^* (\hat{k} \cdot \mathbf{E})$$

as we will see later in this section.

6.2.1 TEM waves in a magnetoplasma

We write the condition for TEM waves explicitly as

$$(n^2 - n_0^2)E_0\hat{u}_0 + (n^2 - n_1^2)E_1\hat{u}_1 + (n^2 - n_{-1}^2)E_{-1}\hat{u}_{-1} = 0.$$

This condition can be satisfied by

$$\mathbf{E} = E_0\hat{u}_0 = E_0\hat{z}$$

if

$$n^2 = n_0^2 = 1 - X.$$

Other solutions satisfying the same constraint are

$$\mathbf{E} = E_{\pm 1}\hat{u}_{\pm 1} = E_{\pm 1}\frac{\hat{x} \mp j\hat{y}}{\sqrt{2}}$$

with

$$n^2 = n_{\pm 1}^2 = 1 - \frac{X}{1 \mp Y}.$$

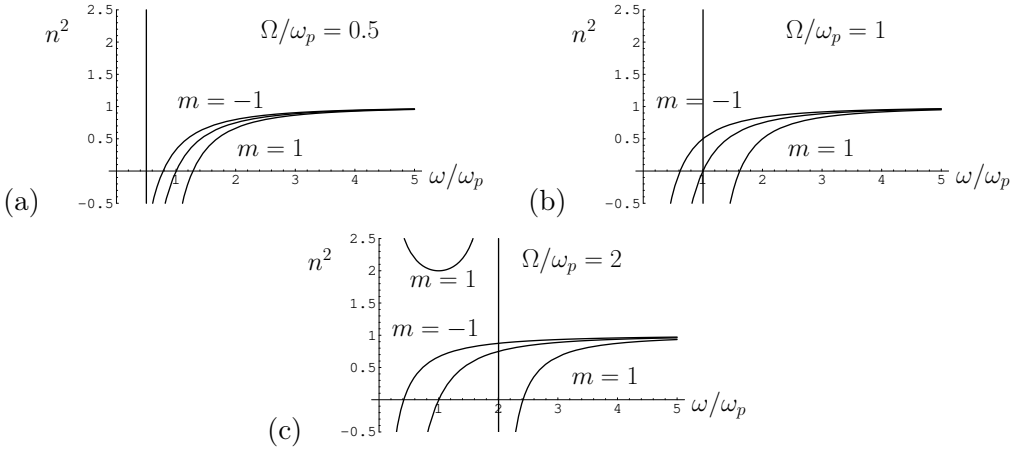


Figure 6.2: Plots of n_m^2 versus $\frac{\omega}{\omega_p}$ for (a) $\frac{\Omega}{\omega_p} = 0.5$, (b) $\frac{\Omega}{\omega_p} = 1$, and (c) $\frac{\Omega}{\omega_p} = 2$. Propagation is only possible for $\frac{\omega}{\omega_p}$ with $n^2 > 0$. The vertical line at $\omega = \Omega$ is known as *gyro-resonance* and corresponds to the pole of n_1^2.

Note that the latter modes can only be TEM, i.e., transverse to the propagation direction $\hat{k}$ if $\hat{k} = \pm\hat{z}$. Thus the circular polarized modes associated with $m = \pm 1$ correspond to TEM waves propagating in the direction of $\mathbf{B}_o$ or in the opposite direction. Hence these circular polarized modes are known as *longitudinal modes.*

Of course the remaining and linear polarized TEM wave $\mathbf{E} = E_0\hat{z}$ is only allowed in a direction perpendicular to $\mathbf{B}_o||\hat{z}$ (since the wave is TEM and thus wavenormal $\hat{k} \perp \mathbf{B}_o$). This wave is therefore a *transverse mode* of propagation. It is known as as *ordinary wave* of a magnetoplasma because its refractive index $n = n_0 = \sqrt{1-X} = \sqrt{1-\frac{\omega_p^2}{\omega^2}}$ is the same as the refractive index in a non-magnetized plasma. This result makes sense since electrons activated by $\mathbf{E}||\mathbf{B}_o$ will move parallel to $\mathbf{B}_o$ and will not feel the effect of $\mathbf{B}_o$.

Figure 6.1 is a summary plot of TEM modes — polarizations as well as propagation directions with respect to the DC magnetic field $\mathbf{B}_o$ — identified above. The variation of $n^2$for the same modes — ordinary wave ($m = 0$), RC ($m = 1$), LC ($m = -1$) — as a function of $\frac{\omega}{\omega_p} = X^{-\frac{1}{2}}$ are shown in Figures 6.2a-c for the cases of $\frac{\Omega}{\omega_p} = 0.5$(weak magnetic field), 1, and 2 (strong magnetic field). Since a mode m propagates only when $n^2 = n_m^2 > 0$ the zero-crossing of each n_m^2 curve in Figure 6.2 signifies a *cutoff* for the corresponding mode. The cutoff condition $n^2 = 0$ implies that at cutoff

$$\frac{X}{1-mY} = 1 \;\Rightarrow\; X = 1 - mY \;\Rightarrow\; \frac{\omega_p^2}{\omega^2} = 1 - m\frac{\Omega}{\omega}.$$

Hence, for ordinary wave the cutoff condition is $\frac{\omega_p}{\omega} = 1$ whereas for RC and LC waves $\frac{\omega_p}{\omega} < 1$ and $\frac{\omega_p}{\omega} > 1$, respectively. Vertical lines in Figure 6.2 correspond to $\omega = \Omega$ at which $n_1^2 = 1 - \frac{X}{1-Y} \to \pm\infty$; $\omega = \Omega$ can be regarded as a *pole* of n_1^2 and is referred to as *gyro-resonance* condition. Notice that in Figure 6.2c describing the case $\frac{\Omega}{\omega_p} = 2$, $n_1^2 > 0$ for $\omega < \Omega$; in fact $n_1^2 > 0$ for $\omega < \Omega$ for all non-zero cases of $\frac{\Omega}{\omega_p}$ but in Figures 6.2a and

b n_1^2 curve of $\omega < \Omega$ band is above the upper plot limit of 2.5. The significance of RC wave propagation in $\omega < \Omega$ band will be examined in Section 4.4. The following example explains the significance of gyro-resonance $\omega = \Omega$ and the gyro-frequency Ω.

6.2.2 Gyro frequency Ω and gyro-resonance

Linear systems exhibit resonant response when an external forcing is applied at a frequency that coincides with the frequency of some unforced internal dynamics. We will next show that in a magnetized plasma gyro-frequency Ω describes unforced motions of free electrons in circular orbits around the magnetic field lines:

Consider once again Newton's 2nd law for a free electron in a magnetoplasma, that is,

$$m\frac{d\mathbf{v}}{dt} = -e(\mathbf{E} + \mathbf{v} \times B_o\hat{z}).$$

Clearly, "zero-input" electron response in the absence of external forcing $\mathbf{E}$ is governed by the homogeneous equation

$$m\frac{d\mathbf{v}}{dt} + e\mathbf{v} \times B_o\hat{z} = 0 \;\; \Rightarrow \;\; \frac{d\mathbf{v}}{dt} + \Omega\mathbf{v} \times \hat{z} = 0.$$

If this equation describes an undamped zero-input response (like oscillations in an LC circuit with a resonance frequency $\omega_o = \frac{1}{\sqrt{LC}}$), then the homogeneous equation above must have a sinusoidal steady-state solution with a phasor constraint

$$j\omega\mathbf{v} + \Omega\mathbf{v} \times \hat{z} = 0 \;\; \Rightarrow \;\; \begin{array}{c} j\omega v_x + \Omega v_y = 0 \\ j\omega v_y - \Omega v_x = 0 \\ j\omega v_z = 0 \end{array}.$$

The constraint is satisfied with $v_z = 0$ and $v_y = -jv_x$ for $\omega = \Omega$ (the only other solution is the trivial one $v_x = v_y = v_z = 0$). Hence, resonant unforced electron oscillations in a magnetized plasma $\mathbf{B}_o = B_o\hat{z}$ are described by

$$\text{Phasor velocity} \;\; \mathbf{v} = V(\hat{x} - j\hat{y}) \;\; \text{and} \;\; \text{Oscillation frequency } \Omega,$$

where V is an arbitrary constant. Velocity phasor $\mathbf{v} = V(\hat{x} - j\hat{y})$ in turn describes a periodic orbit $\mathbf{r}(t)$ with an orbital velocity $\mathbf{v}(t) = \frac{d}{dt}\mathbf{r}(t)$ and a phasor $\mathbf{r}$ such that $j\Omega\mathbf{r} = V(\hat{x} - j\hat{y})$ implying

$$\mathbf{r} = R(\hat{x} - j\hat{y}) \;\; \text{with} \;\; R = -j\frac{V}{\Omega}.$$

In summary, then, free electrons in a magnetized plasma can exhibit unforced revolutions around magnetic field lines described by the tip of (right-circular) vector $\mathbf{r}(t) = \text{Re}\{R(\hat{x} - j\hat{y})e^{j\Omega t}\}$. The revolution frequency is the gyro-frequency $\Omega = \frac{eB}{m}$ and the revolution period is $T_g = \frac{2\pi}{\Omega}$. Notice that V is the orbital speed of the electron and $|R| = \frac{V}{\Omega} = \frac{mV}{eB_o}$ is the orbital radius also known as *gyro-radius*. The mean value of gyro-radius is obtained by replacing V with the electron thermal speed $C_e = \sqrt{\frac{KT_e}{m}}$.

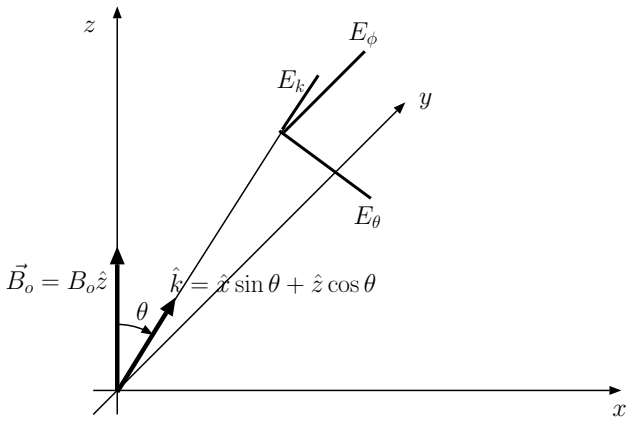


Figure 6.3: Propagation angle θ is measured from DC magnetic field direction $\hat{z}$ towards the x-axis on xz-plane containing the propagation direction $\hat{k} = \hat{x}\sin\theta + \hat{z}\cos\theta$. The field has three mutually orthogonal components E_k, E_θ, and E_ϕ in directions $\hat{k}$, $\hat{\theta}$, and $\hat{\phi}$, respectively. The wave field $\mathbf{E} = \hat{k}E_k + \hat{\theta}E_\theta + \hat{\phi}E_\phi$ is non-TEM.

As for gyro-resonance — a feature of RC TEM wave in a magnetoplasma — it is a simple consequence of the fact that a gyrating electron would see the rotating electric field of a RC TEM wave ....

Finally, $\mathbf{E}_o \times \mathbf{B}_o$ drift ....

6.2.3 O- and X-modes for an arbitrary θ

Plane wave solutions of Maxwell's equations are not limited by the plane TEM waves — ordinary, RC, and LC — examined above. Propagation in an arbitrary direction $\hat{k} = \cos\theta\hat{z} + \sin\theta\hat{x}$ with respect to $\mathbf{B}_o = B_o\hat{z}$ — see Figure 6.3 — is possible in terms of non-TEM modes to be examined next. For these modes the relevant constraint

$$1 = \sum_{m=-1}^{1} \frac{n^2}{n^2 - n_m^2} |\hat{k} \cdot \hat{u}_m|^2.$$

can be written as

$$1 = \frac{n^2}{n^2 - n_0^2}\cos^2\theta + \frac{n^2}{n^2 - n_1^2}\frac{\sin^2\theta}{2} + \frac{n^2}{n^2 - n_{-1}^2}\frac{\sin^2\theta}{2},$$

or, equivalently,

$$\frac{n_0^2}{n^2 - n_0^2}\cos^2\theta + \frac{n_1^2}{n^2 - n_1^2}\frac{\sin^2\theta}{2} + \frac{n_{-1}^2}{n^2 - n_{-1}^2}\frac{\sin^2\theta}{2} = 0,$$

which is known as *Astrom's equation*. Multiplying it with $(n^2-n_0^2)(n^2-n_1^2)(n^2-n_{-1}^2)$ we obtain *Booker's quartic*

$$n_0^2(n^2-n_1^2)(n^2-n_{-1}^2)\cos^2\theta+(n^2-n_0^2)(n^2\frac{n_1^2+n_{-1}^2}{2}-n_{-1}^2n_1^2)\sin^2\theta=0,$$

which is a quadratic equation of n^2. Hence for each direction θ there are two distinct solutions for n^2 describing propagation in $\pm\hat{k}$ directions such that $\hat{k}\cdot\hat{z}=\cos\theta$.

For example, the modes for $\theta=0$ are $n^2=n_{\pm1}^2$ corresponding to RC and LC waves of longitudinal propagation that we have already seen. Also, for $\theta=90^o$ one of the allowed modes is $n^2=n_0^2$, the familiar ordinary wave of transverse propagation. The second transverse propagation mode has

$$n^2=\frac{2n_{-1}^2n_1^2}{n_1^2+n_{-1}^2}\equiv n_x^2,$$

and is known as *extraordinary wave* since, unlike the ordinary wave, this wave is not TEM.

General solutions of Booker's quartic for an arbitrary θ can be expressed as

$$n^2=1-\frac{X}{1-F},$$

where

$$F\equiv\frac{Y_T^2\pm\sqrt{Y_T^4+4Y_L^2(1-X)^2}}{2(1-X)},\ \ Y_T\equiv Y\sin\theta,\ \text{ and }\ Y_L\equiv Y\cos\theta. \tag{6.1}$$

This result[1] is known as *Appleton-Hartree equation*, and its use within polarization equation

$$E_m=\frac{n^2}{n^2-n_m^2}\hat{k}\cdot\hat{u}_m^*(\hat{k}\cdot\mathbf{E})$$

leads to so-called polarization relations

$$\frac{E_{\pm1}}{E_0}=\frac{\sin\theta}{\sqrt{2}\cos\theta}\frac{F(1\mp Y)}{F\mp Y},\ \text{ and }\ \frac{E_{-1}}{E_1}=\frac{(1+Y)(F-Y)}{(1-Y)(F+Y)} \tag{6.2}$$

describing the allowed modes. Note that $\pm$ sign within F results in two distinct values for n^2 in (6.1) as well as distinct ratios in (6.2). In each propagation direction θ these results indicate the availability of two distinct propagation modes known as O- and X-modes.

It can be easily seen that in limiting cases of $\theta\to\pm90^o$ and $\theta\to0$ the results (6.1) and (6.2) are consistent with transverse and longitudinal TEM waves. For instance, for the case of $\theta=90^o$ corresponding to transverse propagation, $Y_L=0$, $Y_T=Y$, and as a consequence one possible solution for F is zero. In that case the Appleton-Hartree

[1]The easiest way to confirm this result is to substitute $n^2=1-\frac{X}{1-F}$ in Astrom's equation and solve for F. Appleton was awarded Nobel Prize in physics for his work on magneto-ionic theory including the derivation of the Appleton-Hartree equation.

equation (6.1) implies $n^2 = 1 - X = n_0^2$, which is the ordinary wave refractive index. Likewise, for the longitudinal case of $\theta = 0$, $F = \pm Y$, implying RC and LC wave refractive indices $n^2 = n_{\pm 1}^2$.

For propagation in an arbitrary θ direction

$$\mathbf{E} = E_1 \frac{\hat{x} - j\hat{y}}{\sqrt{2}} + E_{-1} \frac{\hat{x} + j\hat{y}}{\sqrt{2}} + E_0 \hat{z} = E_0 \left(\frac{F(1-Y)}{F-Y} \frac{(\hat{x} - j\hat{y}) \sin\theta}{2\cos\theta} + \frac{F(1+Y)}{F+Y} \frac{(\hat{x} + j\hat{y}) \sin\theta}{2\cos\theta} + \hat{z} \right);$$

components of vector $\mathbf{E}$ along $\hat{k} = \hat{x}\sin\theta + \hat{z}\cos\theta$, $\hat{\phi} = \hat{y}$, and $\hat{\theta} = \hat{\phi} \times \hat{k} = \hat{x}\cos\theta - \hat{z}\sin\theta$ (see Figure 6.3) are

$$\begin{aligned}
E_k &= \hat{k} \cdot \mathbf{E} = E_0 F \frac{\sin^2\theta}{2\cos\theta} \left(\frac{1-Y}{F-Y} + \frac{1+Y}{F+Y} \right) + E_0 \cos\theta = E_0 \frac{\sin^2\theta}{\cos\theta} \frac{F(F-Y^2)}{F^2-Y^2} + E_0 \cos\theta \\
E_\phi &= \hat{\phi} \cdot \mathbf{E} = -jFE_0 \frac{\sin\theta}{2\cos\theta} \left(\frac{1-Y}{F-Y} - \frac{1+Y}{F+Y} \right) = -jE_0 \frac{\sin\theta}{\cos\theta} \frac{FY(1-F)}{F^2-Y^2} \\
E_\theta &= \hat{\theta} \cdot \mathbf{E} = E_0 F \frac{\sin\theta}{2} \left(\frac{1-Y}{F-Y} + \frac{1+Y}{F+Y} \right) - E_0 \sin\theta = E_0 \sin\theta \frac{Y^2(1-F)}{F^2-Y^2}.
\end{aligned}$$

Thus, O- and X-mode fields for an arbitrary θ can be expressed as

$$\mathbf{E} = (\hat{k}E_k + \hat{\theta}E_\theta + \hat{\phi}E_\phi) e^{-jk_o n(\sin\theta x + \cos\theta z)} \text{ and } \mathbf{H} = \frac{\hat{k} \times \mathbf{E}}{\eta_o / n}, \tag{6.3}$$

where n is obtained from the Appleton-Hartree equation and

$$\frac{E_k}{E_\theta} = \frac{F^2 - Y_L^2 - FY_T^2}{Y_T Y_L (1-F)}, \quad \frac{E_k}{E_\phi} = j \frac{F^2 - Y_L^2 - FY_T^2}{Y_T F (1-F)}, \text{ and } \frac{E_\phi}{E_\theta} = -j \frac{F}{Y_L}. \tag{6.4}$$

In polarization relations (6.4) pertinent to (6.3) the first two ratios $\frac{E_k}{E_\theta}$ and $\frac{E_k}{E_\phi}$ provide a measure of the deviation of O- and X-mode fields from a TEM character—when either one of these ratios has a small magnitude the field is similar to a TEM wave field. The last ratio in (6.4) relates the transverse components of O- and X-mode fields and determines whether the transverse field vector rotates like a circular polarized field or oscillates like a linear polarized field or has an elliptical variation depending on the value of $-jF/Y_L$.

Cutoff conditions $n^2 = 0$:

The magnetoionic modes described by Appleton-Hartree equation exhibit propagation and evanescence characteristics like ordinary waves in unmagnetized plasmas. The transition between propagation and evanescence is described by the cutoff condition $n^2 = 0$ which is realized when

$$\frac{X}{1-F} = 1 \Rightarrow 1 - X = F = \frac{Y_T^2 \pm \sqrt{Y_T^4 + 4Y_L^2(1-X)^2}}{2(1-X)} \Rightarrow (2(1-X)^2 - Y_T^2)^2 = Y_T^4 + 4Y_L^2(1-X)^2$$

$$\Rightarrow 4(1-X)^4 - 4(1-X)^2 Y_T^2 + Y_T^4 = Y_T^4 + 4Y_L^2(1-X)^2 \Rightarrow (X-1)^4 = Y^2(X-1)^2.$$

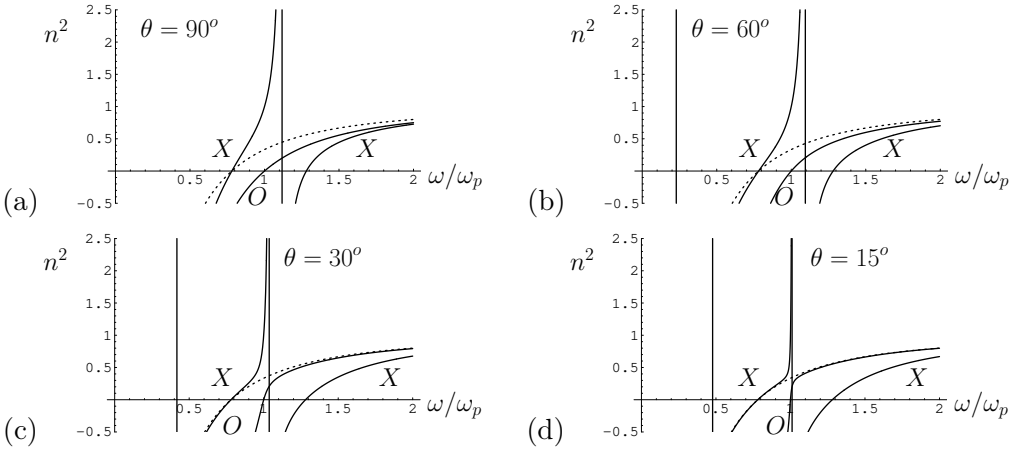


Figure 6.4: Plots of the solutions of Appleton-Hartree equation for $\frac{\Omega}{\omega_p} = 0.5$ and (a) $\theta = 90^o$, (b) $\theta = 60^o$, (c) $\theta = 30^o$, and (d) $\theta = 15^o$. The dashed curve in each plot is n^2_{-1} representing the left-circular longitudinal mode (a valid mode for $\theta = 0^o$). Note that the lower branch of X-mode curve and the O-mode curve are approaching one another near $\frac{\omega}{\omega_p} = 1$ as $\theta \to 0^o$. Also note that for $\theta < 90^o$ the plots exhibit two distinct resonance frequencies where $n^2 \to \pm\infty$. Finally, note that zero-crossings of all n^2 curves at the same values of $\frac{\omega}{\omega_p}$ for all θ.

Therefore one cutoff possibility is $X = 1$, or, equivalently, $\frac{\omega_p^2}{\omega^2} = 1$—the magneto-ionic mode with the cutoff condition $\frac{\omega_p^2}{\omega^2} = 1$ is known as *O-mode*. In $\theta \to 90^o$ limit O-mode turns into the transverse *ordinary wave* examined in Section 9.2. The frequency variation of O-mode refractive index is illustrated in Figures 6.4 and 6.5 for different $\frac{\Omega}{\omega_p}$ and θ conditions.

The second magneto-ionic mode, the *X-mode* (the second solution in Appleton-Hartree equation), exhibits the remaining cutoffs implied by the equation above, namely

$$(X-1)^4 = Y^2(X-1)^2,\ X \neq 1 \ \Rightarrow\ (X-1)^2 = Y^2 \ \Rightarrow\ X = 1 \pm Y,$$

or, equivalently, $\frac{\omega_p^2}{\omega^2} = 1 \pm \frac{\Omega}{\omega}$. To see why the X-mode has two different cutoff conditions note that X-mode traces shown in Figures 6.4 and 6.5 display two distinct frequency bands where $n^2 > 0$—each band has its own cutoff condition where $n^2 \to 0$. These cutoff conditions of low- and high-frequency branches of X-mode, namely $\frac{\omega_p^2}{\omega^2} = 1 + \frac{\Omega}{\omega}$ and $\frac{\omega_p^2}{\omega^2} = 1 - \frac{\Omega}{\omega}$, coincide with the cutoff conditions for *left-* and *right-circular* longitudinal TEM modes discussed in Section 9.2.[2]

[2]Also, as $\theta \to 0^o$ low-frequency branch of X-mode and O-mode merge together to form the longitudinal left-circular mode indicated by the dashed n^2 curves in Figures 6.4 and 6.5. Hence O-mode and its $\frac{\omega}{\omega_p} = 1$ cutoff disappear in $\theta \to 0^o$ limit to be replaced by the left-circular longitudinal TEM mode.

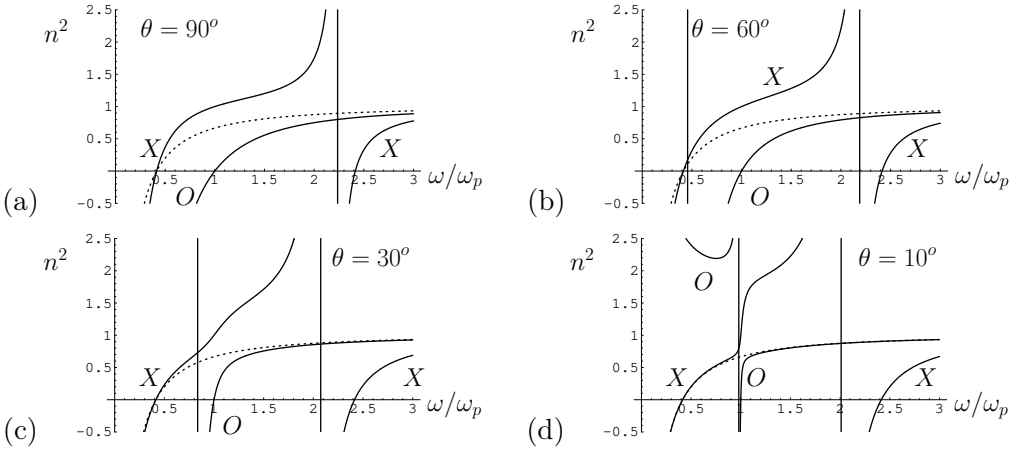


Figure 6.5: Same as Figure 6.4, but for $\frac{\Omega}{\omega_p} = 2$ and for (a) $\theta = 90^o$, (b) $\theta = 60^o$, (c) $\theta = 30^o$, and (d) $\theta = 10^o$.

Orthogonality of magneto-ionic modes:

As we have seen above a magnetized plasma allows only two modes of propagation in each θ direction (with respect to $\mathbf{B}_o$). These normal modes satisfy an important *orthogonality* condition which allows the representation any propagating wave field within the medium as their weighted superpositions.

For instance, right- and left-circular modes of longitudinal propagation ($\theta = 0^o$) are orthogonal since for $\mathbf{E}_r \propto \hat{x} - j\hat{y}$ and $\mathbf{E}_l \propto \hat{x} + j\hat{y}$, $\mathbf{E}_r \cdot \mathbf{E}_l^* = 0$. As a consequence their weighted superposition

$$\mathbf{E}_s = R(\hat{x} - j\hat{y})e^{\pm jk_o n_1 z} + L(\hat{x} + j\hat{y})e^{\pm jk_o n_{-1} z}$$

has an arbitrary polarization—i.e., depending on R and L values the vector $\mathbf{E}_s$ at any z will vary as linear, circular or elliptical polarized way and no known wave polarization is excluded.

Likewise, for an arbitrary $\theta \neq 0^o$, the weighted superposition of transverse components of O- and X-modes,

$$\mathbf{E}_s = A_O(\hat{\theta} - j\hat{\phi}\frac{F_O}{Y_L})e^{-jk_o n_o(\sin\theta x + \cos\theta z)} + A_X(\hat{\theta} - j\hat{\phi}\frac{F_X}{Y_L})e^{-jk_o n_x(\sin\theta x + \cos\theta z)},$$

describes all known polarizations because the vectors $\hat{\theta} - j\hat{\phi}\frac{F_O}{Y_L}$ and $\hat{\theta} - j\hat{\phi}\frac{F_X}{Y_L}$ are also orthogonal:

$$(\hat{\theta} - j\hat{\phi}\frac{F_O}{Y_L}) \cdot (\hat{\theta} - j\hat{\phi}\frac{F_X}{Y_L})^* = 1 + \frac{F_O F_X}{Y_L^2} = 1 + \frac{Y_T^4 - (Y_T^4 + 4Y_L^2(1-X)^2)}{Y_L^2 4(1-X)^2} = 0;$$

In the same limit high-frequency branch of X-mode becomes the right-circular longitudinal TEM mode. Finally, in the same limit low-frequency branch of O-mode—which is only visible in Figure 6.5b—turns into the low-frequency branch of longitudinal right-circular mode shown, for instance, in Figure 6.2c.

to obtain this result we calculated $F_O F_X$ as the product of two versions of F (with $\pm$ difference) corresponding to F_O and F_X.

How does the orthogonality of the components of $\mathbf{E}_s$ guarantee that $\mathbf{E}_s$ can represent an arbitrary polarization? To answer the question let us consider the vector $\mathbf{E}_s$ at $(x, z) = (0, 0)$, i.e.,

$$\mathbf{E}_s = A_O(\hat{\theta} - j\hat{\phi}\frac{F_O}{Y_L}) + A_X(\hat{\theta} - j\hat{\phi}\frac{F_X}{Y_L}),$$

and ask ourselves whether we can find a pair of unique constants A_O and A_X so that $\mathbf{E}_s = T\hat{\theta} + P\hat{\phi}$ where T and P are arbitrary (possibly complex) constants. If the answer is *yes* then $\mathbf{E}_s$ can represent the arbitrary polarized $T\hat{\theta} + P\hat{\phi}$. But the answer is yes because the required constants A_O and A_X are the solutions of the second order linear system

$$A_O + A_X = T \quad \text{and} \quad F_O A_O + F_X A_X = jY_L P,$$

namely

$$A_O = \frac{F_X T - jPY_L}{F_X - F_O} \quad \text{and} \quad A_X = \frac{F_O T - jPY_L}{F_O - F_X};$$

since $F_X \neq F_O$ (if they were equal than the modes would not be orthogonal) we have unique solutions for A_O and A_X.

The implication of above results for propagation in a magnetized plasma is as follows: The only allowed propagation modes O- and X- (or right- and left circular, if $\theta = 0^o$) in the medium are necessary and sufficient to describe the propagation of arbitrary polarized waves injected into the medium. For instance, if an ionosonde transmits a linearly polarized pulse into a magnetized ionosphere, the pulse will travel within the ionosphere as a superposition of a pair of O- and X-polarized pulses (with their distinct phase and group velocities).

Vertical incidence sounding of a magnetized ionosphere:

As a vertical incidence pulse of some frequency ω enters a magnetized ionosphere (like our own ionosphere) with an increasing plasma frequency ω_p it will eventually encounter *two* different reflection heights corresponding to its O- and X-polarized components (unless the pulse happens to be a pure magneto-ionic mode). The reflections will occur as $\frac{\omega_p}{\omega} \to 1$ for O-mode component and as $\frac{\omega_p^2}{\omega^2} \to 1 - \frac{\Omega}{\omega}$ for X-mode—the only exception being the $\theta = 0^o$ case pertinent at geomagnetic poles. As a consequence ionograms of a magnetized ionosphere will typically exhibit two distinct virtual height traces: An O-trace representing the virtual height of $\omega_p = \omega$ and an X-trace representing the virtual height of $\omega_p = \sqrt{\omega(\omega - \Omega)}$. The O-trace, which is the same virtual height curve discussed in Chapter 2, can be distinguished from X-trace as follows: Since at the X-mode reflection height $\frac{\omega}{\omega_p} > 1$ (see Figures 6.4 and 6.5), X-mode reflections continue occurring even after the ionosonde transmission frequency $f = \frac{\omega}{2\pi}$ exceeds the critical frequency of the ionosphere $f_c = \frac{\omega_c}{2\pi}$. Therefore the O-mode trace on an ionogram is the virtual height curve with the narrower frequency extent (see Figure 3.6 once again). Of the two critical

frequencies on an ionogram typically labeled as $Fof2$ and $Fxf2$, only $Fof2$—associated with the O-trace—provides the peak plasma frequency $f_p(z)_{\max} = f_c$ of the ionosphere.

Occasionally, and in particular at high latitudes, ionograms exhibit a third curve known as Z-trace. This occurs when θ is sufficiently small so that the low-frequency X-mode branch and the O-mode nearly touch one another as in Figures 6.4d and 6.5d. When that happens O-mode signal excites some low-frequency branch X-mode field near the $\omega_p = \omega$ height (where the O-mode reflects), which then propagates deeper into the plasma before being reflected from a height where $\frac{\omega_p^2}{\omega^2} \to 1 + \frac{\Omega}{\omega}$; the reflected X-mode signal gets excites an O-mode field near $\omega_p = \omega$ height (via the converse coupling process) which then returns to the ionosonde antenna as a extra-delayed O-mode echo. Since the process requires small values of θ Z-traces are most commonly observed at high latitudes where vertical $\hat{k}$ of ionosonde transmissions and the nearly vertical geomagnetic field vector $\mathbf{B}_o$ are nearly parallel.

Resonances in a magnetized plasma:

The vertical lines in Figures 6.4 and 6.5 correspond to resonances of a magnetized plasma at which $n^2 \to \pm\infty$. These poles of Appleton-Hartree equation occur when

$$F = \frac{Y_T^2 \pm \sqrt{Y_T^4 + 4Y_L^2(1-X)^2}}{2(1-X)} = 1, \;\; X \neq 0 \text{ and } X \neq 1$$

$$\Rightarrow \; (2(1-X)-Y_T^2)^2 = Y_T^4+4Y_L^2(1-X)^2 \; \Rightarrow \; 1-X = Y^2-Y_L^2X \; \Rightarrow \; 1-\frac{\omega_p^2}{\omega^2} = \frac{\Omega^2}{\omega^2}(1-\frac{\omega_p^2}{\omega^2}\cos^2\theta)$$

$$\Rightarrow \; \omega^4 - (\omega_p^2+\Omega^2)\omega^2 + \Omega^2\omega_p^2 = 0 \; \Rightarrow \; \omega^2 = \frac{(\omega_p^2+\Omega^2) \pm \sqrt{(\omega_p^2+\Omega^2)^2 - 4\Omega^2\omega_p^2\cos^2\theta}}{2}.$$

This result indicates two resonant plasma modes (unforced plasma oscillations) for each direction θ with respect to the DC magnetic field $\mathbf{B}_o$. Longitudinal resonances corresponding to $\theta \to 0^o$ are $\omega = \omega_p$ (plasma resonance[3]) and $\omega = \Omega$ (gyro-resonance) while transverse resonances corresponding to $\theta \to 90^o$ are $\omega = \sqrt{\omega_p^2 + \Omega^2}$ (upper-hybrid resonance) and $\omega = 0$ (DC). Resonant frequencies of a magnetized plasma separate the low-and-high frequency branches of permitted magneto-ionic modes. The "zero-input" response of a plasma to an initial perturbation (e.g., caused by a satellite or Space Shuttle action) can also be expressed in terms of a superposition of resonance modes.

[3] As a result of a "pole-zero cancellation" the refractive index n remains finite at plasma resonance $\omega = \omega_p$ (or $X = 1$). This reflects the fact that longitudinal right- and left-circular modes with transverse EM fields cannot couple to longitudinal plasma oscillations with a longitudinal fields in $\mathbf{B}_o$ direction.

Poynting vector:

In a magneto-ionic medium with plane waves

$$\mathbf{E} = (\hat{k}E_k + \hat{\theta}E_\theta + \hat{\phi}E_\phi)e^{-jk_o n(\sin\theta x + \cos\theta z)} \text{ and } \mathbf{H} = \frac{\hat{k}\times\mathbf{E}}{\eta_o/n},$$

the time-averaged Poynting vector (for the propagating case of $n^2 > 0$) is

$$\begin{aligned}
\mathbf{S} = \frac{1}{2}\mathrm{Re}\{\mathbf{E}\times\mathbf{H}^*\} &= \frac{n}{2\eta_o}\mathrm{Re}\{(\hat{k}E_k + \hat{\theta}E_\theta + \hat{\phi}E_\phi)\times(\hat{\phi}E_\theta^* - \hat{\theta}E_\phi^*)\} \\
&= \frac{n}{2\eta_o}\mathrm{Re}\{\hat{k}(|E_\theta|^2 + |E_\phi|^2) - \hat{\theta}E_k E_\theta^* - \hat{\phi}E_k E_\phi^*\} \\
&= \hat{k}\frac{n|\mathbf{E}_\perp|^2}{2\eta_o} - \frac{n}{2\eta_o}\mathrm{Re}\{\frac{F^2 - Y_L^2 - FY_T^2}{Y_T Y_L(1-F)}|E_\theta|^2\hat{\theta} + j\frac{F^2 - Y_L^2 - FY_T^2}{Y_T F(1-F)}|E_\phi|^2\hat{\phi}\} \\
&= \hat{k}\frac{n|\mathbf{E}_\perp|^2}{2\eta_o} - \hat{\theta}\frac{n|E_\theta|^2}{2\eta_o}\frac{F^2 - Y_L^2 - FY_T^2}{Y_T Y_L(1-F)},
\end{aligned}$$

where $\mathbf{E}_\perp \equiv (\hat{\theta}E_\theta + \hat{\phi}E_\phi)e^{-jk_o n(\sin\theta x + \cos\theta z)}$ is the transverse component of the field $\mathbf{E}$, and $|\mathbf{E}_\perp|^2 = \mathbf{E}_\perp \cdot \mathbf{E}_\perp^* = |E_\theta|^2 + |E_\phi|^2$. Hence, in a magneto-ionic medium the direction of time-averaged Poynting vector, i.e., the direction of energy transport, deviates from the wave-normal direction $\hat{k}$ in the plane defined by $\hat{k}$ and $\mathbf{B}_o$. This effect causes the O- and X-components of radiowave pulses to follow slightly different paths in the ionosphere and is referred to as *magneto-ionic splitting*.

6.3 Faraday rotation

Most of the magneto-ionic complications in radiowave propagation disappear at sufficiently large frequencies ω compared to the resonance frequencies ω_p and Ω of the ionosphere. Graphically this is evident from Figures 6.4 and 6.5 which show that O- and X-mode refractive index curves start resembling one another as ω increases to the right of all resonances. In this section we will examine this high-frequency regime where magneto-ionic effects are relatively subdued (to the extent that in some applications they may even be neglected—see Section 9.4.3).

6.3.1 Quasi-longitudinal propagation

For $\omega \gg \omega_p$ and $\omega \gg \Omega$ condition $Y_T^2 \ll 2Y_L(1-X)$ is valid over nearly all propagation angles θ, and, as a consequence,

$$F = \frac{Y_T^2 \pm \sqrt{Y_T^4 + 4Y_L^2(1-X)^2}}{2(1-X)} \approx \pm Y_L,\ n^2 = 1 - \frac{X}{1-F} \approx 1 - \frac{X}{1-mY_L} \equiv n_m^2, \text{ and } \frac{E_\phi}{E_\theta} \approx -mj,$$

$m = \pm 1$. Thus inequality

$$Y_T^2 \ll 2Y_L(1-X) \tag{6.5}$$

assures magneto-ionic modes which are "quasi-circular" in their transverse polarizations (e.g., $\mathbf{E}_\perp \approx E_\theta(\hat{\theta} \mp j\hat{\phi})$) with refractive indices

$$n_m^2 = 1 - \frac{X}{1 - mY\cos\theta} \tag{6.6}$$

that resemble $n_m^2 = 1 - \frac{X}{1-mY}$ of right- and left-circular longitudinal modes. Inequality (6.5) is known as *quasi-longitudinal condition* and magneto-ionic modes with refractive indices (6.6) are known as quasi-circular modes. Quasi-circular modes can also be shown to be quasi-TEM in the sense that $|E_k|^2 \ll |\mathbf{E}_\perp|^2$.

For relatively small ω the quasi-longitudinal condition

$$Y_T^2 \ll 2Y_L(1-X) \;\Rightarrow\; \sin^2\theta \ll 2\cos\theta\frac{1-X}{Y} \;\Rightarrow\; \tan\theta\sin\theta \ll 2\frac{1-X}{Y} = 2\frac{\omega^2-\omega_p^2}{\omega\Omega}$$

holds over a very narrow band of θ centered about $\theta = 0^o$. However, for $\omega \gg \omega_p$ and $\omega \gg \Omega$ the same condition

$$\tan\theta\sin\theta \ll 2\frac{\omega^2-\omega_p^2}{\omega\Omega} \approx \frac{2\omega}{\Omega}$$

has a very broad range of validity in θ. Note that since $\sin\theta \le 1$, for the above condition to be valid it is sufficient that $\tan\theta \ll \frac{2\omega}{\Omega}$, or, equivalently,

$$\theta \ll \tan^{-1}\frac{2\omega}{\Omega}.$$

With $f = 10\Omega$, the condition is satisfied if $\theta \ll \tan^{-1} 40\pi \approx 89.54^o$. Furthermore, very rapid variation of $\tan\theta$ near $\theta = 90^o$ ensures that the condition is well satisfied even at angles as close as $\theta = 86^o$ to $\theta = 89.54^o$. To see that, note that for $\theta = 86^o$, $\sin\theta\tan\theta \approx 14.27 \ll 40\pi \approx 125.7$. Since ionospheric gyro-frequency Ω is typically less than $10^7\,\mathrm{s}^{-1}$, this result is pertinent for $f \ge 100$ MHz in the ionosphere—at VHF or higher frequencies used in trans-ionospheric links nearly all propagation angles θ satisfy the quasi-longitudinal condition.

In summary then, in ionospheric propagation in VHF or higher frequency bands the normal modes of propagation can be described by

$$\mathbf{E} \approx E_o(\hat{\theta} \pm j\hat{\phi})e^{\pm jk_o n_m(x\sin\theta + z\cos\theta)},\;\; \mathbf{H} = \frac{n_m\hat{k}\times\mathbf{E}}{\eta_o},\;\; \text{and}\;\; n_m^2 = 1 - \frac{X}{1-mY\cos\theta},$$

as long as the angle θ between $\mathbf{B}_o = B_o\hat{z}$ and wave-normal $\hat{k} = \hat{x}\sin\theta + \hat{z}\cos\theta$ is not 90^o. We use $\pm$ above to refer to propagation along $(-)$ and against $(+)$ direction $\hat{k} \equiv \hat{x}\sin\theta + \hat{z}\cos\theta$, $0^o \le \theta \le 90^o$, in order to avoid confusion between n_1 (right-circular) and n_{-1} (left-circular) formulas that arises when $\theta > 90^o$ is allowed.

6.3.2 Faraday rotation and applications

An arbitrary quasi-longitudinal plane wave propagating in $\pm\theta$ direction (except $\theta \approx 90^o$) can be expressed as a weighted superposition

$$\mathbf{E} = R(\hat{\theta}-j\hat{\phi})e^{\pm jk_o n_1(x\sin\theta+z\cos\theta)}+L(\hat{\theta}+j\hat{\phi})e^{\pm jk_o n_{-1}(x\sin\theta+z\cos\theta)},\ n_m^2 = 1-\frac{X}{1-mY\cos\theta}$$

of quasi-circular modes. For $L = 0$ the field $\mathbf{E}$ is right-circular, for $R = 0$ left-circular, and otherwise linear or elliptical polarized.

In trans-ionospheric *communication* links (Earth-to-satellite or satellite-to-satellite) the best strategy is to use either a right-circular $\mathbf{E}$ (i.e., $L = 0$) or a left-circular ($R = 0$). Otherwise, unequal propagation (phase and group) speeds of the two normal mode components of $\mathbf{E}$ generate unnecessary complications. Also circular transmissions can be detected by antennas of arbitrary linear polarization with no risk of full mismatch, which, as we will see next, is a serious concern for trans-ionospheric links relying on linear polarized transmissions.

However, linear polarized transmissions through the ionosphere play an important function for remote sensing purposes. The reason is, the rotation of polarization-plane of a linear polarized field traversing the ionosphere can be used to infer the amount of free-electrons in the ionosphere along the propagation path. This rotation is known as *Faraday effect* (or Faraday rotation) and is described next.

Faraday rotation:

Consider a homogeneous magnetized plasma with a DC magnetic field $\mathbf{B}_o = \hat{z}B_o$ and a wave propagating in $\hat{k} = \hat{z}$ direction with the field

$$\mathbf{E} = R(\hat{\theta} - j\hat{\phi})e^{-jk_o n_1 z} + L(\hat{\theta} + j\hat{\phi})e^{-jk_o n_{-1} z},\ n_m^2 = 1 - \frac{X}{1-mY\cos\theta}.$$

At $z = 0$ the field is

$$\mathbf{E}(0) = R(\hat{\theta} - j\hat{\phi}) + L(\hat{\theta} + j\hat{\phi}),$$

and assuming $R = L = 1$ it is linear polarized in $\hat{\theta} = \hat{x}$ direction. But with $R = L = 1$ the field at an arbitrary z,

$$\mathbf{E}(z) = (\hat{\theta} - j\hat{\phi})e^{-jk_o n_1 z} + (\hat{\theta} + j\hat{\phi})e^{-jk_o n_{-1} z}$$

is not $\hat{x}$-polarized because $n_{-1} \neq n_1$. What is polarization of $\mathbf{E}(z)$? To find out we will first define

$$\bar{n} \equiv \frac{n_{-1}+n_1}{2} \text{ and } \delta n \equiv n_{-1} - n_1 \ \text{ so that } \ n_{-1} = \bar{n} + \frac{\delta n}{2} \text{ and } n_1 = \bar{n} - \frac{\delta n}{2};$$

then,

$$\begin{aligned}\mathbf{E}(z) &= e^{-jk_o\bar{n}z}[(\hat{\theta} - j\hat{\phi})e^{jk_o\delta n z/2} + (\hat{\theta} + j\hat{\phi})e^{-jk_o\delta n z/2}] \\ &= 2e^{-jk_o\bar{n}z}[\hat{\theta}\cos(k_o\delta n z/2) + \hat{\phi}\sin(k_o\delta n z/2)].\end{aligned}$$

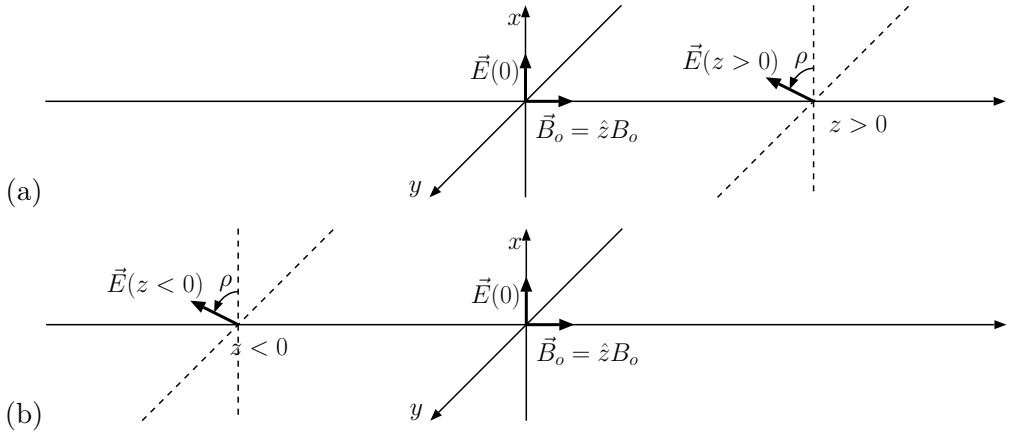


Figure 6.6: (a) An $\hat{x}$-polarized field propagating in z-direction parallel to a DC magnetic field $\mathbf{B}_o = \hat{z}B_o$, $B_o > 0$, rotates by an angle ρ over a propagation distance $z > 0$—rotation direction of ρ is right-handed with the thumb pointing in $\mathbf{B}_o$ direction. (b) The polarization plane also rotates in right-handed sense when the field propagates in the opposite direction to $\mathbf{B}_o = \hat{z}B_o$, $B_o > 0$.

Clearly, $\mathbf{E}(z)$ is linear polarized at an angle

$$\rho = \frac{k_o \delta n}{2} z \,\text{rad}$$

with respect to $\hat{\theta} = \hat{x}$ as shown in Figure 6.6a. Angle ρ, which is a linear function of travel distance z, is known as *Faraday rotation angle*. Given that $n_m^2 = 1 - \frac{X}{1-mY}$, and assuming small X and Y (justified in quasi-longitudinal regime)

$$\begin{aligned} \frac{\rho}{zk_o} &= \frac{\sqrt{1-\frac{X}{1+Y}} - \sqrt{1-\frac{X}{1-Y}}}{2} \approx \frac{-\frac{X}{1+Y} + \frac{X}{1-Y}}{4} \\ &= \frac{X}{4}\left(\frac{1}{1-Y} - \frac{1}{1+Y}\right) \approx \frac{XY}{2}, \end{aligned}$$

or

$$\rho = \frac{k_o XY}{2} z = \frac{\omega}{2c}\frac{\omega_p^2}{\omega^2}\frac{\Omega}{\omega} z = \frac{\Omega}{2c}\frac{f_p^2}{f^2} z \approx \frac{40.3\Omega N}{cf^2} z \,\text{rad},$$

where N is plasma density in m^{-3} units[4]. Notice that Faraday angle ρ is positive, and therefore polarization rotation is clockwise as shown in Figure 6.6a when the field vector is viewed looking in the direction of $\mathbf{B}_o = \hat{z}B_o$.

Now, here is a tricky detail: What if the wave propagated from $z = 0$ to some $z < 0$, i.e., in the direction opposite to $\mathbf{B}_o = \hat{z}B_o$, starting with $\hat{x}$-polarization at $z = 0$ as before? What would be the Faraday angle ρ?

[4]We used the formula $f_p^2 \approx 80.6N$.

Here is the answer: In that case the answer above would still be valid when both k_o and z were entered as negative quantities—negative k_o because the propagation direction is opposite to the originally assumed one, and negative z because we are inquiring the polarization state of the field at negative z. The upshot is, two negatives make a positive, and Faraday rotation is still clockwise as shown in Figure 6.6b (when the field vector is viewed looking in the direction of $\mathbf{B}_o = \hat{z}B_o$).

In summary, for Faraday rotation it does not matter whether propagation is along or against $\mathbf{B}_o$; rotation is always clockwise—in the sense illustrated in Figures 6.6a and b, and only the propagation distance, that is $|z|$, matters. Thus, with $r \equiv |z|$ denoting propagation path length,

$$\rho \approx \frac{40.3\Omega N}{cf^2} r \,\text{rad}.$$

Example 1

Faraday angle formula above was obtained assuming longitudinal propagation. But the only difference in refractive index expressions of longitudinal and quasi-longitudinal modes is $Y = \frac{\Omega}{\omega}$ vs $Y\cos\theta = \frac{\Omega\cos\theta}{\omega}$. Hence inserting a $\cos\theta$ term next to Ω in the equation for ρ, we obtain

$$\rho \approx \frac{40.3\Omega N\cos\theta}{cf^2} r$$

to cover the quasi-longitudinal case. The sense of Faraday rotation in quasi-longitudinal case are illustrated in Figures 6.6a and b.

Finally, above results assumed a homogeneous magneto-plasma. If the plasma is inhomogeneous, so that $N = N(r)$ and also possibly $\Omega = \Omega(r)$ and $\theta = \theta(r)$ because of a slow variation of DC field $\mathbf{B}_o$, then over an infinitesimal section dr of the propagation path

$$d\rho \approx \frac{40.3\Omega(r)N(r)\cos\theta(r)}{cf^2} dr,$$

and, therefore,

$$\rho(r) \approx \frac{40.3}{cf^2}\int_0^r dr'\Omega(r')\cos\theta(r')N(r')$$

over a path r. This result neglects ray bending due to refraction, which is reasonable within the quasi-longitudinal regime.

Example 2: Assume a ground-satellite link traverses a "slab-ionosphere" of a constant plasma density of $N = 3\times10^{12}\,\text{m}^{-3}$. If the propagation path through the ionosphere is 300 km, $\Omega(r') \approx 5\times10^6\,\text{s}^{-1}$, and $\theta(r') = 60^o$, determine ρ for $f = 150$ MHz.

Solution:

$$\rho(r) \approx \frac{40.3}{(3\times10^8)(150\times10^6)^2}(5\times10^6)(0.5)(3\times10^{12})(300\times10^3) \approx 13.4\,\text{rad} \approx 769.7^o.$$

Note that $\rho > 2\times360^o = 720^o$, indicating more than two-full rotation of the plane of polarization.

Estimating the TEC:

Suppose over a trans-ionospheric path $\Omega(r') \approx \Omega_i$ and $\theta(r') \approx \theta_i$ within ionospheric heights; then the Faraday angle can be written as

$$\rho(r) \approx \frac{40.3\Omega_i \cos\theta_i}{cf^2} N_T(r), \quad \text{where} \quad N_T(r) \equiv \int_0^r dr' N(r'),$$

denotes a total electron content TEC. While the Faraday angle $\rho(r) \propto$ TEC, an unambiguous estimation of TEC from Faraday measurements can only be possible if the number of rotations of the polarization vector through the ionosphere can be determined. The most straightforward way of doing that is as follows:

Let $\rho_1(r)$ and $\rho_2(r)$ denote the polarization angles of signals received from a single source transmitting at two different frequencies f_1 and $f_2 > f_1$; assuming identical linear polarizations for the two signals at the source and using the Faraday angle formula above,

$$\Delta\rho(r) \equiv \rho(r)_{f_1} - \rho(r)_{f_2} \approx \frac{40.3\Omega_i \cos\theta_i}{c}\left(\frac{1}{f_1^2} - \frac{1}{f_2^2}\right) N_T(r) = \frac{40.3\Omega_i \cos\theta_i}{cf_1f_2} \frac{(f_2 - f_1)(f_2 + f_1)}{f_1 f_2} N_T(r).$$

For sufficiently close f_1 and $f_2 > f_1$, the TEC $N_T(r)$ can be obtained from measured $\Delta\rho(\mathbf{r})$ with no ambiguity.

Example 3: Given that $f_1 = 149$ MHz and $f_2 = 151$ MHz, calculate $\Delta\rho(r)$ for the same trans-ionospheric path described in Example 2 above.

Solution:

$$\Delta\rho(r) \approx \frac{40.3(5 \times 10^6)(0.5)}{(3 \times 10^8)(149 \times 10^6)(151 \times 10^6)} \frac{(2)(300)}{(149)(151)} (3{\times}10^{12})(300{\times}10^3) \approx 0.36\,\text{rad} \approx 20.63^o.$$

Thus, with $f_1 \equiv f - \frac{\Delta f}{2}$ and $f_2 \equiv f + \frac{\Delta f}{2}$, and a sufficiently small $\Delta f \ll f$,

$$N_T(r) \approx \frac{cf^3}{80.5\Omega_i \cos\theta_i} \frac{\Delta\rho(r)}{\Delta f}.$$

Differential rotation method for measuring ionospheric electron density N:

Recall the relation

$$d\rho \approx \frac{40.3\Omega(r)N(r)\cos\theta(r)}{cf^2} dr, \;\; \Rightarrow \;\; N(r) = \frac{cf^2}{40.3\Omega(r)\cos\theta(r)} \frac{d\rho}{dr};$$

this suggests that if Faraday rotation angle $\rho(r)$ is measured as a function of $0 \le r \le R$, then the corresponding derivative $\frac{d\rho}{dr}$ can be used to estimate $N(r)$, $0 \le r \le R$, assuming $\Omega(r)$ and $\theta(r)$ are known. This "differential rotation" method for ionospheric electron density measurements can be implemented in several different ways:

Rocket measurements: Let $r(t)$ denote the instantaneous distance of a sounding rocket[5] launched at $t = 0$ from a ground based receiving antenna where the angle ρ of a linear polarized transmission from the rocket is measured as a function of time t. In that case

$$\frac{d\rho}{dt} = \frac{d\rho}{dr}\frac{dr}{dt} = v(t)\frac{d\rho}{dr}$$

at the ground based receiver, where $v(t) = \frac{dr}{dt}$ is the rocket speed relative to the receiving antenna. Replacing $\frac{d\rho}{dr}$ with $\frac{1}{v(t)}\frac{d\rho}{dt}$ in the expression above for $N(r)$, we obtain

$$N(r(t)) = \frac{cf^2}{40.3\Omega(r(t))\cos\theta(r(t))v(t)}\frac{d\rho}{dt}. \quad \text{Sounding rocket equation} \tag{6.7}$$

This formula can be used to obtain ionospheric density profile $N(z)$ if the rocket trajectory $(x(t),y(t),z(t))$ with respect to the ground based antenna is known. Note that $r(t) = \sqrt{x^2(t) + y^2(t) + z^2(t)}$, and $\Omega(r(t))$ and $\theta(r(t))$ can be either obtained from a *magnetometer* carried by the rocket or from global geomagnetic field models such as IGRF (see Appendix IV).

Backscatter radar measurements: Backscatter research radars designed to study the upper atmosphere detect echoes of their transmitted signal (typically a pulse modulated carrier) scattered from atmospheric refractive index irregularities (see Chapter 5). For powerful research radars such as Jicamarca and Arecibo the echoes arrive continuously from nearly all radar ranges $r = \frac{ct}{2}$ where t denotes the time after pulse transmission. A radar echo returned from range r at time t has traveled twice over the same ionospheric path (on the way up and on the way down) and therefore the Faraday rotation angle $\rho(t)$ measured by the radar (assuming the radar transmits linear polarized pulses) is twice the amount corresponding to a one way travel over the path $r = \frac{ct}{2}$. Therefore, the electron density formula for backscatter radars is one-half the rocket formula (which accounts for only one way rotation) with the rocket speed $v(t)$ replaced by $\frac{dr}{dt} = \frac{c}{2}$, the "speed of radar echo height"; the upshot is,

$$N(\frac{ct}{2}) = \frac{f^2}{40.3\Omega(\frac{ct}{2})\cos\theta(\frac{ct}{2})}\frac{d\rho}{dt}. \quad \text{Backscatter radar equation} \tag{6.8}$$

In the application of this formula Ω and θ are obtained by evaluating the IGRF model along the radar beam coordinates.

Example 4: At Jicamarca radar in Peru $f = 50$ MHz and $\theta \approx 87^o$. Assuming that the variation of $\rho(t)$ are sampled at $\Delta t = 0.1$ ms intervals, estimate $\Delta\rho$ (over one sampling interval) for $\Omega = 10^6\,\mathrm{s}^{-1}$ and $N = 10^{12}\,\mathrm{m}^{-3}$.

[5]NASA and European Space Agency (ESA) routinely conduct *sounding rocket* experiments for the exploration of upper atmosphere and ionosphere. Sounding rockets carry various electronics designed for the measurement of *in situ* properties of the atmosphere (e.g., density, temperature, atmospheric electric fields) as well as radio beacons for propagation experiments in conjunction with ground based receivers. Sounding rocket measurements are crucial for the *in situ* probing of the 30 km-200 km region of the atmosphere (too high for instrumented balloons and airplanes and too low for satellites). NASA rockets are typically surplus missiles from the U.S. military and include models that can reach apogees of up to 1000 km.

Solution: Using (6.8),

$$\frac{d\rho}{dt} = \frac{40.3\Omega\cos\theta}{f^2}N = \frac{40.3(10^6)\cos 87^o}{(50\times 10^6)^2}10^{12} \approx 843.6\,\frac{\text{rad}}{\text{s}};$$

therefore,

$$\Delta\rho \approx \frac{d\rho}{dt}\Delta t \approx 843.6\,\frac{\text{rad}}{\text{s}} \times 10^{-4}\,\text{s} = 0.084\,\text{rad} \approx 4.81^o.$$

Example 5: Repeat Example 4 for Arecibo with $f = 430$ MHz and $\theta \approx 45^o$.

Solution: Since $\rho \propto \frac{\cos\theta}{f^2}$, for Arecibo

$$\Delta\rho \approx 4.81^o\frac{50^2}{430^2}\frac{\cos 45^o}{\cos 87^o} \approx 0.898^o.$$

A comparison of Examples 4 and 5 shows that 50 MHz Jicamarca radar has a better sensitivity for Faraday density measurements than the 430 MHz Arecibo radar.

Other applications of Faraday rotation:

Faraday effect is not unique to magnetized plasmas and also arises in other materials such as glass, water, salt crystals (an even in air, to a very weak extent) in the presence of strong DC magnetic fields. There are many applications of Faraday effect in optical systems. For instance, Faraday effect can be used to modulate the polarization direction of light coming out of a polarizing filter as a means of information transmission. Also the operation principle of LCDs depend on Faraday rotation of polarized light through a "liquid crystal" (liquids composed of elongated molecules that maintain a directional order between certain temperature limits) medium sandwiched between a pair of linear orthogonal polarized filters.

6.4 Whistlers

The low-frequency branch of O-mode as well as longitudinal right-circular TEM mode exhibit positive n^2 in $\omega \to 0$ regime that we have largely ignored in our discussions so far. These propagation modes known as *whistler*s have interesting properties and are useful for remote sensing the magnetosphere, a buffer region between the ionosphere and the solar wind[6]. We will briefly examine here the case of longitudinal whistlers, namely low-frequency right-circular modes propagating along the geomagnetic field lines.

Atmospheric lightning is one example of a variety of natural processes that can generate whistler mode radiation. Since lightning is a short duration current pulse, the

[6]Via Lorentz force the geomagnetic field deflects the streaming charged particles of solar wind and thereby isolates the ionosphere from the solar wind plasma by generating a magnetospheric cavity. The interaction process between the geomagnetic field and the solar wind compresses the sunward side of the geomagnetic field structure causing dayside/nightside an asymmetry in the cavity. Time variations in this interaction due to solar flares cause many magnetospheric and ionospheric effects, generally referred to as *space weather*. See

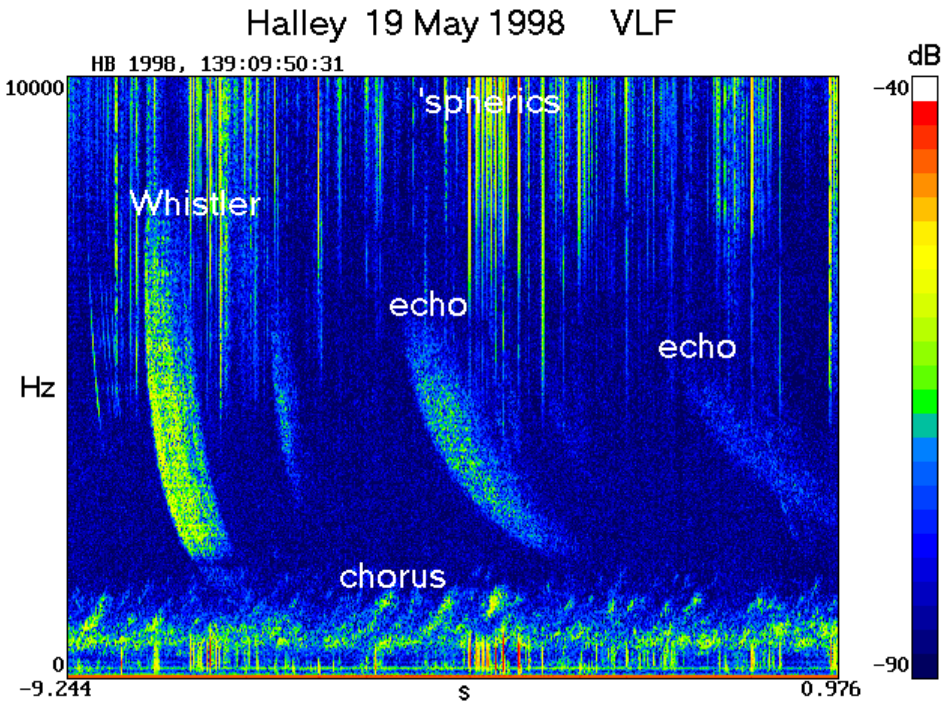


Figure 6.7: A spectrogram (f vs τ plot of signal intensity) showing a whistler and its echoes from the conjugate hemisphere. Downloaded from the web site http://www.nerc-bas.ac.uk/public/uasd/programs/wave/

radiation caused by the lightning has a broad frequency spectrum including the whistler regime. Such a pulse due to, say, a northern hemisphere lightning can penetrate the ionosphere and be ducted[7] along the Earth's dipolar geomagnetic field lines into the southern hemisphere and even penetrate the ionosphere all the way to the ground. Since the broadband pulse—a short "click", as it sounds on a radio receiver placed close to the lightning event—undergoes a severe dispersion along its inter-hemispheric journey, by the time it reaches the southern hemisphere the pulse envelope does not resemble its original form at all. The "click" has effectively turned into a long "whistle", as the transformed pulse will sound on a southern hemisphere radio receiver. The frequency content of the whistle tells something about the magnetospheric regions the pulse has traveled through in its journey. Frequency-time plots of two whistler events are shown in Figure 6.7.

Whistlers were first discovered during World War I during the attempts of Germans and the Allies to eavesdrop on one another's telephone communications. Both parties initially suspected that frequently encountered whistlers—sounding almost like "poiu"—may be some coded form of enemy communications. The generation of whistlers by a combined effect of atmospheric lightnings and magneto-ionic propagation was eventually understood after World War II. Here is a simplified version of the explanation:

For right-circular longitudinal modes, the refractive index

$$n^2 = n_1^2 = 1 - \frac{X}{1-Y} = 1 - \frac{\omega_p^2}{\omega(\omega-\Omega)} = 1 + \frac{\omega_p^2}{\omega(\Omega-\omega)};$$

now, for $\omega_p^2 \gg \omega\Omega$ and $\omega_p \gg \omega$—relevant for the pow-frequency whistler band—the second term of n^2 is dominant, and therefore

$$n \approx \frac{\omega_p}{\sqrt{\omega(\Omega-\omega)}} \;\Rightarrow\; n_g = \frac{d}{d\omega}(\omega n) \approx \frac{d}{d\omega}\frac{\omega^{1/2}\omega_p}{(\Omega-\omega)^{1/2}} = \frac{\omega_p\Omega}{2\omega^{1/2}(\Omega-\omega)^{3/2}}.$$

The upshot is, the whistler band group velocity for right-circular waves is

$$v_g = \frac{c}{n_g} = \frac{2c\sqrt{\omega(\Omega-\omega)^3}}{\omega_p\Omega}.$$

The group velocity maximizes (as confirmed in homework) for $\omega = \frac{\Omega}{4}$ at a value of

$$v_{g,\max} = \frac{3\sqrt{3}c}{8}\frac{\Omega}{\omega_p}.$$

Hence the group delay over a distance r in the whistler band is given as

$$\tau = \frac{r}{v_g} = \frac{r}{2c}\frac{\omega_p\Omega}{\sqrt{\omega(\Omega-\omega)^3}} = \frac{r}{2c\sqrt{\frac{\omega}{\Omega}(1-\frac{\omega}{\Omega})^3}}\frac{\omega_p}{\Omega} \;\text{ with }\; \tau_{\min} = \frac{8r}{3\sqrt{3}c}\frac{\omega_p}{\Omega}.$$

Figure 6.8 shows a plot of $\frac{\tau}{r/c}$ vs $\frac{\omega}{\Omega}$ under the assumption $\frac{\omega_p}{\Omega} = 10$; note the resemblance of the group delay curve to the whistlers shown in Figure 6.7.

[7] Ducting requires the existence magnetic-field aligned plasma density inhomogeneities that act as effective waveguides; such irregularities exist naturally within the magnetosphere.

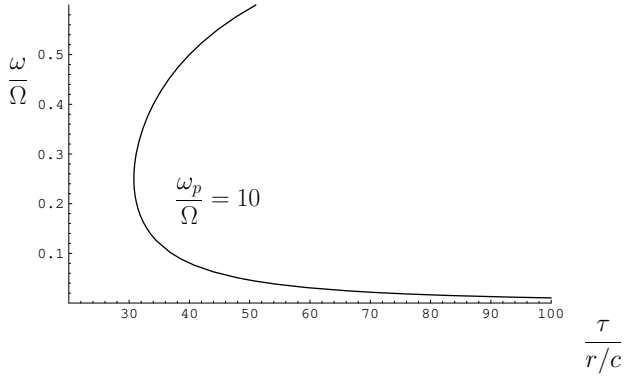


Figure 6.8: Normalized group delay for the electron whistler mode as a function of normalized frequency $\frac{\omega}{\Omega}$ ($\frac{\omega_p}{\Omega} = 10$ case). Note the "nose frequency" $\omega = \frac{\Omega}{4}$ corresponding to the shortest group delay (fastest group velocity).

At very low frequencies $\omega \ll \Omega$ pertinent to the whistler mode propagation, positive-ion dynamics which has been neglected in the conductivity expressions derived in this chapter starts becoming important. Therefore, a careful investigation of whistlers require the inclusion of positive-ion effects in magneto-ionic equations—see "Radiowaves in the ionosphere" by Budden.

7 Appendix I: Chapman theory of ionospheric formation

The ionosphere is a consequence of the ionizing effect of solar X-ray and EUV radiation incident upon the neutral atmosphere. Here we will very briefly outline the process of the formation of the ionosphere.

Assume a hypothetical atmosphere composed of a single species of electrically neutral particles (e.g., hydrogen atoms) with an absorption cross-section σ of photons of a certain frequency ν. If the volumetric number density of neutral particles is N, and if $I(\nu)$ stands for the intensity (power density) of photons of frequency ν arriving with an incidence angle χ (see Figure 7.1), then each atmospheric layer of depth dz will absorb an amount of power $I\sigma N \sec\chi dz$ per unit atmospheric area. As a result of this absorption solar flux I will decrease with decreasing atmospheric height as described by

$$\frac{dI}{dz} = I\sigma N \sec\chi \;\;\Rightarrow\;\; I(z) = I_\infty e^{-\tau(z)}, \;\text{ with }\; \tau(z) \equiv \sigma \sec\chi \int_z^\infty dz'\; N(z'); \qquad (7.1)$$

above, I_∞ denotes the power density of incoming solar radiation prior to entering the atmosphere and $\tau(z)$ is called *optical depth* for I.

For an atmosphere in hydrostatic equilibrium (see Appendix II) it can be shown that (as confirmed in a homework problem)

$$\int_z^\infty dz'\; N(z') = N(z)H(z) \;\text{ with }\; H(z) = \frac{KT(z)}{Mg}$$

where T is atmospheric temperature, M the effective mass of air molecules (28.97 gm/mole=4.809×10^{-26} kg), $g = 9.8$ m/s^2 the gravitational acceleration, and $K = 1.381 \times 10^{-23}$ J/K the Boltzmann's constant. Thus, the optical depth $\tau(z)$ in (7.1) is explicitly

$$\tau(z) = \sigma \sec\chi N(z)H(z).$$

We will next assume that the number of electron-ion pairs produced via photo ionization per unit absorbed energy is $\eta\,\frac{\text{pairs}}{\text{J}}$. The *ionization rate* $q(z)$ per unit volume is then $\eta\frac{dI}{dz}$, or

$$q(z) = \eta\sigma N(z) I_\infty e^{-\tau(z)} = \eta\sigma N(z) I_\infty e^{-\sigma\sec\chi N(z)H(z)}\,\frac{\text{pairs}}{\text{m}^3\text{s}}.$$

In an isothermal atmosphere where $H(z) = H_o = \frac{KT_o}{Mg}$ is constant, the ionization rate $q(z)$ can be simplified and expressed in a normalized form (as confirmed in a homework problem) as

$$\frac{q(z)}{q_o} = e^{1-Z-e^{-Z}\sec\chi} \;\text{ where }\; Z = \frac{z-z_o}{H_o} \;\text{ and }\; q_o = \eta\sigma N(z_o) I_\infty e^{-\tau(z_o)};$$

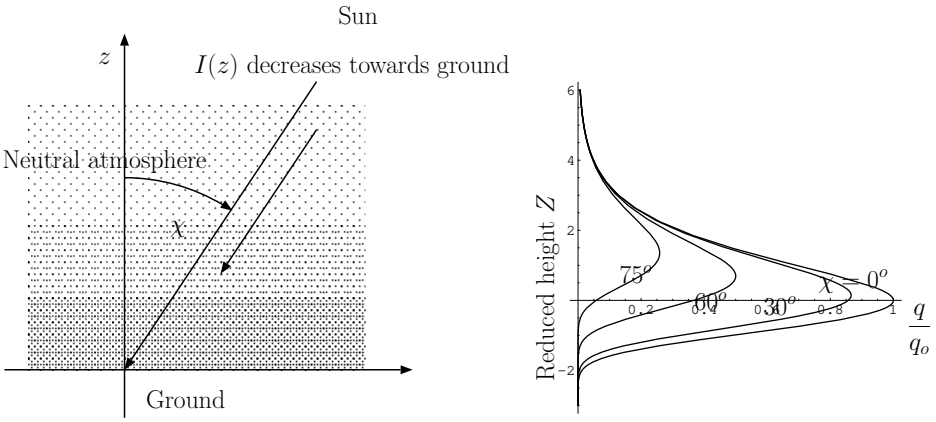


Figure 7.1: Solar flux intensity $I(z)$ decreases towards the ground as photons are absorbed by neutral atmospheric particles to produce electron-ion pairs. The Chapman production function $q(z)$ peaks at some atmospheric height z_o when the solar zenith angle $\chi = 0$, and it varies as a function of reduced height $Z \equiv \frac{z-z_o}{H_o}$ as shown on the right. As χ increases solar flux per unit area drops and results in the reduction of maximum of $\frac{q}{q_o}$. Also as χ increases the total path length of solar photons through the atmosphere increases which causes an upward shift of the peak of the production rate curve. The single humped shape of production rate curves is caused by the reduction of ionizing photons with decreasing height and the reduction of ionizable neutral particle density with increasing height. Chapman functions are parabolic near their peaks, grow as double exponentials below the peak and decay exponentially above the peak.

Z is known as "reduced height" and parameter z_o, satisfying the constraint

$$\tau(z_o) = \sigma N(z_o) H_o = 1,$$

is the height of maximum $q(z)$ when the solar zenith angle $\chi = 0^o$ (i.e., the condition at noon).

Plots of *Chapman production function* $\frac{q(z)}{q_o} = e^{1-Z-e^{-Z}\sec\chi}$ for various solar zenith angles χ are shown in Figure 7.1 on the right (see the caption for a discussion of $q(z)$ variation with height and solar zenith angle).

Electron production rate in the Earth's upper atmosphere is obtained by superposing the Chapman production functions $q(z)$ for multiple particle species responding to multiple photon frequencies, including the effects of a height dependent atmospheric temperature distribution.[1]

The number density of free electrons in the ionosphere represents an equilibrium between the electron production rate $q(z)$ and electron loss rate $l(z)$ due electron-ion recombination processes. For example, the loss rate in *dissociative recombination*

[1]See Chapter 3 in "Introduction to ionospheric physics" by H. Rishbeth and O. K. Garriot.

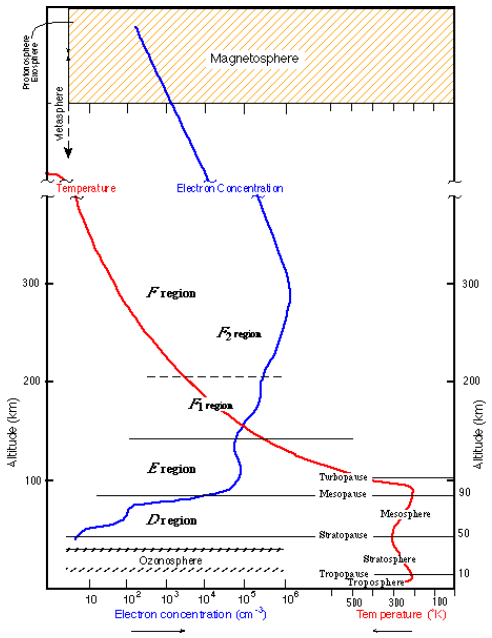


Figure 7.2: Sketches of typical profiles ionospheric electron density $N_e(z)$ and temperature $T(z)$ (from http://ion.le.ac.uk/images/ionosphere/); note that $N_e(z)$ is labeled in cm^{-3} units instead of m^{-3} units.

$(XY^+ + e \to X + Y)$ between ionized molecules XY^+ and electrons e is of the form

$$l = \alpha N_e N_+,$$

where α is the *recombination rate coefficient* and $N_{e,+}$ are electron and positive ion number densities. Equating l with q, and taking $N_+ = N_e$ (because the ionosphere is charge neutral on the average), we obtain

$$N_e(z) = \sqrt{q(z)/\alpha}$$

for an ionosphere dominated by dissociative recombination.

Since there are may different types of recombination processes in the ionosphere the relationship between electron density $N_e(z)$ and production rate $q(z)$ is considerably more involved than the simple formula above and exhibits a height variation that reflects the dominance of different recombination processes at different height regions. Nevertheless in very crude terms the localization of the ionospheric peak density at some finite height can be viewed as a consequence of height localization of the production rate function. Day-night differences in ionospheric densities also reflect the fact that production $q(z)$ ceases after sunset (except for minor processes such as corpuscular ionization). The maintenance of some regions of the ionosphere (F-region, in particular—see Figure 7.2) after the sunset is a result of very small rates of certain types of recombination processes (atomic ions of the F-region, in particular, recombine very slowly because energy and

momentum conservation requires the presence of a third particle in any electron/atomic-ion encounter that leads to a successful recombination).

Figure 7.2 shows plots describing the general features of ionospheric electron density $N_e(z)$. Note the nomenclature used to refer to different regions of the ionosphere. The region of the peak electron density is referred to as F2 region. The lowest altitude part of the ionosphere is D-region. In between lies the E-region which was the first ionospheric region detected by remote sensing techniques. The main effect of D-region on radiowaves is absorption. E-region and F-region are important in ionospheric reflections of HF signals used in sky-wave links. The most significant ionospheric effects on high-frequency trans-ionospheric links take place in the high-density F1 and F2 regions.

8 Appendix II: Tropospheric refractive index model

Within the lowest ∼15 km of the Earth's atmosphere air temperature decreases with increasing height at an average rate of about 6.5 K/km. This region of the atmosphere is known as *troposphere*. Tropospheric temperature profile is a consequence of solar heating of the Earth's surface — the bottom boundary of the troposphere — and infrared cooling of air. The temperature profile starts rising above the troposphere in a region known as *stratosphere* because of solar UV absorption by atmospheric ozone. However, situated above 15 km height, the stratosphere does not play a significant role in space-wave communications — only the lowest few kilometers of the troposphere really matters.

Refractive index $n(z)$ drops slowly and almost linearly with z within the first few kms of the troposphere. This variation mainly reflects the thinning of the atmosphere with increasing height. Since $\mu = \mu_o$ and $n^2 = \epsilon_r = 1 + \chi$, the dependence of electric susceptibility $\chi \propto N$ on the number density N of polarizable air molecules is what dictates how refractive index n varies with height within the atmosphere. According to ideal gas law

$$p = NKT, \tag{8.1}$$

where $K = 1.381 \times 10^{-23}$ J/K is Boltzmann's constant, $N \propto \frac{p}{T}$; therefore

$$n^2 = 1 + \chi = 1 + A\frac{p}{T}$$

where K is a constant that depends on the polarizability of air molecules and p and T are atmospheric pressure and temperature. For dry air—with a composition of $\sim 78\%$ N_2, $\sim 20\%$ O_2, and 2% minor constituents—laboratory measurements show that $A = 2 \times 77.6 \times 10^{-6}\ \frac{\mathrm{K}}{\mathrm{mB}}$ and thus for $A\frac{p}{T} \ll 1$,

$$n = 1 + 77.6 \times 10^{-6}\frac{p}{T}. \tag{8.2}$$

Measurements with moist air, on the other hand, show that

$$n = 1 + 77.6 \times 10^{-6}(\frac{p}{T} + 4810\frac{e}{T^2}), \tag{8.3}$$

where e is partial pressure of water vapor in mB units.

To infer the tropospheric refractive index profile $n(z)$ from these general results it is necessary to know how the ratios $\frac{p}{T}$ and $\frac{e}{T^2}$ change with z within the troposphere. When the atmosphere is in hydrostatic equilibrium—as in the standard troposphere—downward gravitational force $ANmgdz$ on any air parcel of thickness dz and surface

area A is balanced by an upward buoyancy force of $A[p(z) - p(z + dz)]$ due to pressure differential. This implies that under hydrostatic equilibrium

$$\frac{dp}{dz} = -NMg, \tag{8.4}$$

where M denotes the effective mass of air molecules (28.97 gm/mole=4.809×10^{-26} kg) and g is gravitational acceleration, 9.8 m/s^2. Combining the ideal gas law (8.1) with *hydrostatic equation* (8.4) it can be shown that

$$\frac{dp}{p} = -\frac{dz}{H(z)}, \tag{8.5}$$

where

$$H(z) \equiv \frac{KT(z)}{Mg} \tag{8.6}$$

is known as *atmospheric scale height*. Integrating (8.5) between heights 0 and z we find

$$p(z) = p_o e^{-\int_0^z dz/H(z')},$$

where $p_o = 950$ mB=95000 Pascales (P) is the mean surface pressure. Since tropospheric temperature varies as $T(z) = T_o - \alpha z$, where, on the average, $T_o = 290$ K and $\alpha = 6.5\ \frac{\text{K}}{\text{km}}$, within the first few kilometers of the troposphere the scale height $H(z) \approx H_o \approx 8.5$ km can be taken as constant. Thus for $z \ll H_o$,

$$p(z) = p_o e^{-\int_0^z dz/H(z')} \approx p_o e^{-\frac{z}{H_o}} \approx p_o(1 - \frac{z}{H_o}) = p_o - \frac{p_o Mg}{KT_o} z = p_o - N_o Mgz.$$

and

$$\frac{p(z)}{T(z)} = \frac{p_o - N_o Mgz}{T_o - \alpha z} = \frac{p_o}{T_o} \frac{1 - \frac{z}{H_o}}{1 - \frac{\alpha z}{T_o}} \approx KN_o(1 - (\frac{1}{H_o} - \frac{\alpha}{T_o})z). \tag{8.7}$$

Tropospheric humidity measurements, on the other hand, indicate that water vapor pressure is on the average

$$e(z) = e_o - \beta z,$$

with $e_o = 8$ mB and $\beta = 3\ \frac{\text{mB}}{\text{km}}$ within the first 2 km of the troposphere. Thus,

$$\frac{e(z)}{T^2(z)} = \frac{e_o - \beta z}{(T_o - \alpha z)^2} \approx \frac{e_o}{T_o^2} \frac{1 - \frac{\beta}{e_o} z}{1 - \frac{2\alpha}{T_o} z} \approx \frac{e_o}{T_o^2}(1 - (\frac{\beta}{e_o} - \frac{2\alpha}{T_o^2})z). \tag{8.8}$$

Substituting (8.7) and (8.8) in (8.3), and multiplying out the constant terms it is found that

$$n(z) = 1 + 10^{-6}(289 - 36z), \tag{8.9}$$

where z is measured in kilometer units. The constant term in (8.9) is approximately 1.0003 while the coefficient of z, $36 \times 10^{-6}\ \frac{1}{\text{km}}$, is very close to $\frac{1}{4a}$, where $a \approx 6371.2$ km is the Earth radius. Hence, (8.9) can be approximated as

$$n(z) = n_o - \frac{z}{4a}, \tag{8.10}$$

with $n_o = 1.0003$. This result is known as *standard tropospheric refractive index* model. The model is only valid within the lowest few kilometers of the troposphere and $n(z)$ does not fall below 1 until ionospheric heights are reached ($z \geq 50$ km).

9 Appendix III: Antenna reception

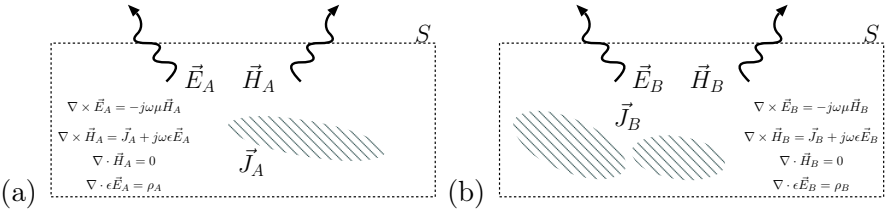


Figure 9.1: (a) and (b) depict the same volume in space under two different electromagnetic states with fields and sources defined with subscripts A and B governed by Maxwell's equations shown in the figures. No currents are assumed to flow outside the surface S even though fields propagate out of the surface in both cases.

A reciprocity relation

Figures 9.1a and b depict two distinct situations describing the same region of space bounded by some surface S. In Figure 9.1a some current distribution $\mathbf{J}_A$ exists within surface S accompanied by fields $\mathbf{E}_A$ and $\mathbf{H}_A$. Although these fields can propagate to regions outside S their source is assumed to be within S, i.e., no currents flow outside S. In the second setting shown in Figure 9.1b we have different currents and fields $\mathbf{J}_B$, $\mathbf{E}_B$, and $\mathbf{H}_B$, but, otherwise, the same rules apply. The rules include Maxwell's equations governing the currents and fields $\mathbf{J}_A$, $\mathbf{E}_A$, and $\mathbf{H}_A$ as well as $\mathbf{J}_B$, $\mathbf{E}_B$, and $\mathbf{H}_B$ as indicated in the figure[1].

Observe that with the set of equations shown in Figures 9.1a and b,

$$\begin{aligned} \mathbf{H}_B \cdot \nabla \times \mathbf{E}_A - \mathbf{H}_A \cdot \nabla \times \mathbf{E}_B &= 0 \\ \mathbf{E}_B \cdot \nabla \times \mathbf{H}_A - \mathbf{E}_A \cdot \nabla \times \mathbf{H}_B &= \mathbf{E}_B \cdot \mathbf{J}_A - \mathbf{E}_A \cdot \mathbf{J}_B, \end{aligned}$$

which can be combined as

$$\nabla \cdot [\mathbf{E}_A \times \mathbf{H}_B - \mathbf{E}_B \times \mathbf{H}_A] = \mathbf{E}_B \cdot \mathbf{J}_A - \mathbf{E}_A \cdot \mathbf{J}_B$$

using a well known vector identity. Integrating both sides over an arbitrarily big spherical volume V centered about surface S and using Gauss' law we obtain

$$\int_{S'} da\ \hat{n} \cdot [\mathbf{E}_A \times \mathbf{H}_B - \mathbf{E}_B \times \mathbf{H}_A] = \int_S dV [\mathbf{E}_B \cdot \mathbf{J}_A - \mathbf{E}_A \cdot \mathbf{J}_B],$$

[1]Since $\mathbf{D}$ and $\mathbf{B}$ are taken as $\epsilon\mathbf{E}$ and $\mu\mathbf{H}$, the discussion here excludes anisotropic media.

where S' is the surface of volume V, $\hat{n}$ is the outward unit vector on surface S', and the right-hand side is a volume integral carried out within surface S. Since $\mathbf{H}_{A,B} = \hat{n} \times \mathbf{E}_{A,B}/\eta$ on S' sufficiently far away from the radiation sources within S, the left-hand side is zero and, consequently,

$$\int_S dV \mathbf{E}_B \cdot \mathbf{J}_A = \int_S dV \mathbf{E}_A \cdot \mathbf{J}_B.$$

Both sides of this *reciprocity relation* denote a volume integral over the region within S.

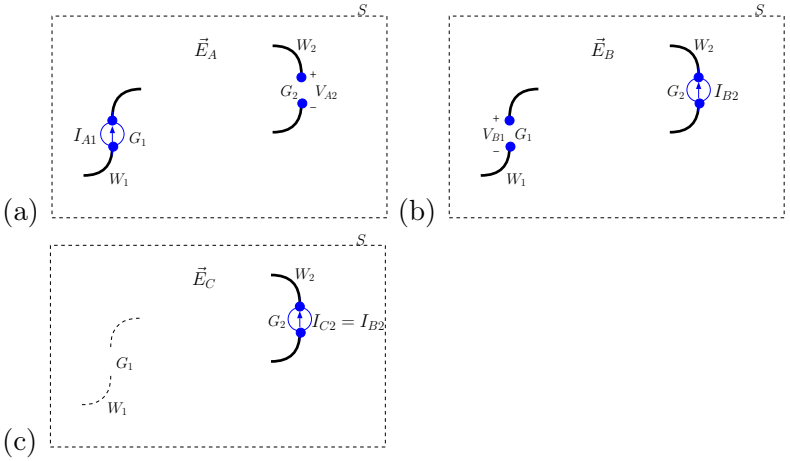


Figure 9.2: Situations (a), (b), and (c) concern the same wire antennas W_1 and W_2. In (a) an independent current source I_{A1} bridging the gap G_1 of antenna 1 drives a field $\mathbf{E}_A$ which also includes a component scattered from W_2. Situation (b) is similar but with antenna 2 driven instead of antenna. Situation (c) is the same as (b) except that antenna 1 has been removed.

Coupling impedance and antenna open circuit voltage

We will next use the reciprocity relation derived above under the assumption that $\mathbf{J}_A$ and $\mathbf{J}_B$ flow only on the surfaces of perfectly conducting and infinitely thin wires contained within S. Under that assumption the volume integrals above reduce to line integrals and we obtain

$$\int_L d\mathbf{l} \cdot \mathbf{E}_B I_A = \int_L d\mathbf{l} \cdot \mathbf{E}_A I_B,$$

where it is implied that line integrals will have to be carried over all the wires located within S. We will use the convention that the current reference direction at each segment agrees with the vector direction of $d\mathbf{l}$.

We will first apply this integral constraint to situations depicted in Figures 9.2a and b. Current[2] I_A needed in the left-hand side of above equation flows on wires labelled as

[2]Needless to say, current I_A is not a constant on wires W_1 and W_2 — it varies with position along

W_1 and W_2 in Figure 9.2a, as well as through the current source I_{A1} bridging the gap G_1 between the two segments of W_1. However, I_A is clearly zero at the "open" gap G_2 between the two segments of W_2. Therefore,

$$\int_L d\mathbf{l}\cdot\mathbf{E}_B I_A = \int_{G_1+W_1+W_2} d\mathbf{l}\cdot\mathbf{E}_B I_A = I_{A1}\int_{G_1} d\mathbf{l}\cdot\mathbf{E}_B + \int_{W_1+W_2} d\mathbf{l}\cdot\mathbf{E}_B I_A = I_{A1}\int_{G_1} d\mathbf{l}\cdot\mathbf{E}_B,$$

since the electric field $\mathbf{E}_B$ will have a zero tangential component along the infinitely conducting wire segments W_1 and W_2. Applying the same analysis next with the current I_B flowing along W_1, W_2, and G_2 in Figure 9.2b, we see that

$$\int_L d\mathbf{l}\cdot\mathbf{E}_A I_B = \int_{W_1+W_2+G_2} d\mathbf{l}\cdot\mathbf{E}_A I_B = \int_{W_1+W_2} d\mathbf{l}\cdot\mathbf{E}_A I_B + I_{B2}\int_{G_2} d\mathbf{l}\cdot\mathbf{E}_A = I_{B2}\int_{G_2} d\mathbf{l}\cdot\mathbf{E}_A,$$

where I_{B2} is the current source bridging gap G_2 shown in Figure 9.2b.

Clearly, according to above results

$$I_{A1}\underbrace{\int_{G_1} d\mathbf{l}\cdot\mathbf{E}_B}_{\equiv -V_{B1}} = I_{B2}\underbrace{\int_{G_2} d\mathbf{l}\cdot\mathbf{E}_A}_{\equiv -V_{A2}},$$

or

$$I_{A1}V_{B1} = I_{B2}V_{A2}$$

in terms of *open circuit voltages* V_{B1} (across "antenna 1" when "antenna 2" is transmitting) and V_{A2} (across antenna 2 with antenna 1 transmitting) corresponding to the line integral of electric fields across the open gaps between antenna terminals. A similar analysis contrasting Figures 9.2a and c leads to

$$\int_{G_1+W_1} d\mathbf{l}\cdot\mathbf{E}_C I_A = I_{C2}\underbrace{\int_{G_2} d\mathbf{l}\cdot\mathbf{E}_A}_{-V_{A2}},$$

which can also be written as

$$-\int_{G_1+W_1} d\mathbf{l}\cdot\mathbf{E}_C I_A = I_{B2}V_{A2}$$

since $I_{C2} = I_{B2}$. Note that the path W_1 contributes to the line integral above on the left because the antenna wire along W_1 has been removed in Figure 9.2c and cannot short out $\mathbf{E}_C$ locally.

The implications of

$$I_{A1}V_{B1} = I_{B2}V_{A2} \quad \text{and} \quad -\int_{G_1+W_1} d\mathbf{l}\cdot\mathbf{E}_C I_A = I_{B2}V_{A2}$$

are as follows:

W_1 and W_2 but in order to keep our notation simple we have avoided writing something like $I_A(\mathbf{r})$. Similar comments apply to all currents defined and used in this section.

1. $$\frac{V_{B1}}{I_{B2}} = \frac{V_{A2}}{I_{A1}} \equiv Z_c,$$

   where Z_c is a *coupling impedance* between antennas W_1 and W_2 which can be used in equivalent circuit models of coupled antennas (see next section).

2. $$V_{B1} = -\int_{G_1+W_1} d\mathbf{l} \cdot \mathbf{E}_C \frac{I_A}{I_{A1}},$$

   an explicit formula for the *open circuit voltage* of antenna 1 in situation B in terms of the incident electric field $\mathbf{E}_C$ that it has been exposed to. Notice the subtlety that since $I_{C2} = I_{B2}$ the field $\mathbf{E}_C$ is also the field generated by antenna 2 in situation B before a scattered field from antenna 1 turns it into the sum $\mathbf{E}_B$. Also notice that $\frac{I_A}{I_{A1}}$ in the open-circuit voltage formula for V_{B1} is the self-normalized current distribution of antenna 1 when it is used as a transmitting antenna. Finally, notice that an explicit expression for the coupling impedance Z_c can be readily obtained by dividing $V_{B1}$with I_{B2} and expressing $\mathbf{E}_C$ in terms of input current I_{B2} (see homework).

We can generalize the *very important* second implication from above as follows:

An antenna that supports a normalized current distribution $\frac{I(\mathbf{r})}{I_o}$ when it is being used for transmission purposes will exhibit an open circuit voltage

$$V_o = -\int_L d\mathbf{l} \cdot \mathbf{E}(\mathbf{r}) \frac{I(\mathbf{r})}{I_o}$$

at its feed point when it is exposed to an incident field $\mathbf{E}(\mathbf{r})$; the integration path L in the formula is "self-defined" by all the locations with non-zero $\frac{I(\mathbf{r})}{I_o}$ and, *by definition*, $\frac{I(\mathbf{r})}{I_o} = 1$ at the antenna feed point.

Although we have only examined thin wire antennas here, similar arguments can be used to generalize our findings for antennas with planar surfaces as well.

Open circuit voltage response to plane waves

Consider next an antenna located at the origin — such as a short dipole as shown in Figure 9.3 — which is exposed to an incident plane TEM wave field

$$\mathbf{E}(\mathbf{r}) = \mathbf{E}_o e^{j\mathbf{k}\cdot\mathbf{r}}$$

arriving from a direction

$$\hat{k} = \frac{\mathbf{k}}{|\mathbf{k}|} = \frac{\mathbf{k}}{k}.$$

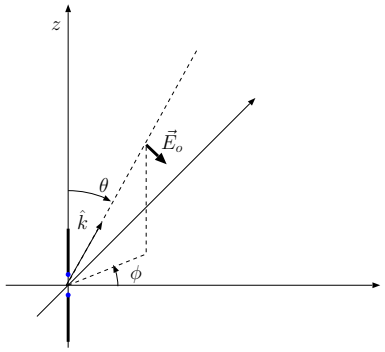


Figure 9.3: A TEM plane wave $\mathbf{E}_o e^{j\mathbf{k}\cdot\mathbf{r}}$ is incident on an antenna (e.g., a short dipole) at the origin from a direction $\hat{k}$. Note that $\mathbf{k}\cdot\mathbf{E}_o = 0$.

The open circuit voltage response of the antenna to the plane wave will be

$$V_o = -\int_L d\mathbf{l}\cdot\mathbf{E}(\mathbf{r})\frac{I(\mathbf{r})}{I_o} = -\mathbf{E}_o\cdot\int_L d\mathbf{l}\frac{I(\mathbf{r}')}{I_o}e^{jk\hat{k}\cdot\mathbf{r}'}.$$

Noting that

$$\hat{k}\times(\hat{k}\times d\mathbf{l}) \quad = \quad (\hat{k}\cdot d\mathbf{l})\hat{k} - \underbrace{(\hat{k}\cdot\hat{k})}_{1} d\mathbf{l} = (\hat{k}\cdot d\mathbf{l})\hat{k} - d\mathbf{l}$$

and, thus,

$$\mathbf{E}_o\cdot\hat{k}\times(\hat{k}\times d\mathbf{l}) \quad = \quad (\hat{k}\cdot d\mathbf{l})\underbrace{\hat{k}\cdot\mathbf{E}_o}_{\text{0 since the field is TEM}} - \mathbf{E}_o\cdot d\mathbf{l} = -\mathbf{E}_o\cdot d\mathbf{l},$$

we can write

$$V_o = -\mathbf{E}_o\cdot\int_L d\mathbf{l}\frac{I(\mathbf{r}')}{I_o}e^{jk\hat{k}\cdot\mathbf{r}'} = \mathbf{E}_o\cdot\hat{k}\times\hat{k}\times\int_L d\mathbf{l}\frac{I(\mathbf{r}')}{I_o}e^{jk\hat{k}\cdot\mathbf{r}'} = \mathbf{E}_o\cdot\mathbf{f}(\hat{k}),$$

where

$$\mathbf{f}(\hat{k}) = \hat{k}\times\hat{k}\times\int_L d\mathbf{l}\frac{I(\mathbf{r}')}{I_o}e^{jk\hat{k}\cdot\mathbf{r}'}$$

is the *effective length* $\mathbf{f}(\hat{r})$ of the receiving antenna evaluated at $\hat{r} = \hat{k}$, the direction of the incident plane wave.

10 Appendix IV: Antenna feed networks and circular polarized antennas

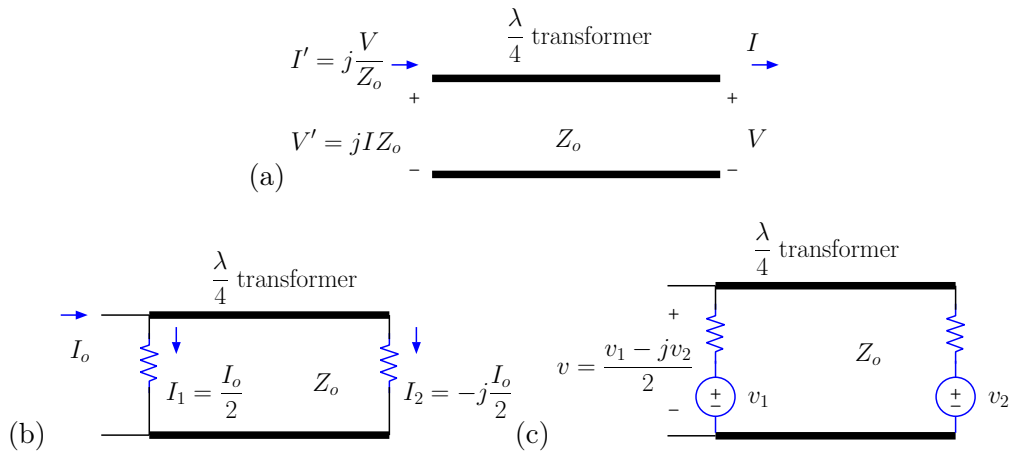


Figure 10.1: (a) A quarter-wave transformer, and the feed network of a 2-element array with (b) radiating elements and (c) receiving elements.

Feed networks

The voltage and current variables indicated in the transmission line circuits shown in Figures 10.1b and c can be easily confirmed using the "quarter-wave transformer" voltage and current transformation rules marked in Figure 10.1a. All resistances (including the characteristic impedance of $\frac{\lambda}{4}$ length transmission lines) in the figures are assumed equal to Z_o.

Suppose that resistors in Figure 10.1b represent radiation resistances of identical z-polarized dipole antennas[1] located at the origin (with current $I_1 = \frac{I_o}{2}$) and at $x = \frac{\lambda}{4}$ (with current $I_2 = -j\frac{I_o}{2}$) forming a 2-element array. You should be able to verify that the described array is "endfire" and forms a beam pointed along the x-axis. Furthermore, the beam has a null in $-x$-direction.

Clearly then, the circuit in Figure 10.1b depicts the "feed network" of a very simple antenna array. The feed network in this case is just a $\frac{\lambda}{4}$-long transmission line which accomplishes two tasks: Splits the input current I_o of the array equally between its

[1] So called "resonant" line antennas, e.g., $\frac{\lambda}{2}$-dipole, have purely resistive impedances that can be matched by lossless transmissions lines as shown in Figures 23b and c.

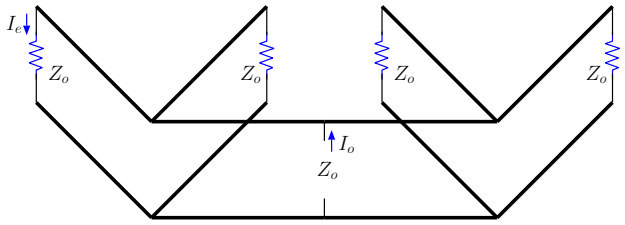


Figure 10.2: *Corporate ladder* feed network. Each transmission line segment (6 of them) is a quarter-wave transformer and all transmission lines have the same characteristic impedance as the load resistors Z_o. The impedance at the feed point of the network (shown at the center) is also Z_o. As shown in the text $|I_e| = \frac{|I_o|}{2}$, where I_e denotes the element currents.

radiating elements, and applies a 90^o phase difference between the element inputs to cause a radiation peak in x-direction. To steer the beam to some other direction than x-axis the dipoles would need to be fed by a different network (e.g., a transmission line section with a different length).

The response of the same array to incoming radiation can be examined with the help of Figure 10.1c. In this figure the current source has been removed from the array terminals, but sources V_1 and V_2 represent the Thevenin voltages of the two dipole elements which will be non-zero when the array is illuminated by some incoming wave. For the particular feed network that combines the two dipoles (same as in Figure 10.1b) the array open-circuit voltage is $V = \frac{V_1 - jV_2}{2}$ as shown in the diagram. With a different feed network V would have a different dependence on V_1 and V_2.

Figure 10.2 shows a popular feed structure known as *corporate ladder*. Resistors Z_o represent identical antenna elements, and the network consists of 6 quarter-wave transformers with Z_o characteristic impedances. Working your way from the antenna elements towards the feed point and using the quarter-wave impedance transformation rule you should be able to confirm that the input impedance of the network is also Z_o. Since the total path length from the input to each element is the same each element carries the same antenna current I_e (equal phase and equal amplitude). The array radiates $4 \times \frac{1}{2}|I_e|^2 Z_o = \frac{1}{2}|I_o|^2 Z_o$ power (assuming a lossless ladder) and therefore $|I_e| = \frac{|I_o|}{2}$. Since element impedances match the line impedances, the length of segments connected to the antennas can be adjusted to vary the phase of element currents without altering their amplitudes. Thus a four element array fed by a corporate ladder can be easily steered by simple length adjustments.

We can also use corporate ladders (5 of them) to feed a 16 element array, etc.

Circular-polarized cross-dipole antennas

Figures 10.3a and b are identical with Figures 10.1b and c except for voltage and current labeling. In Figure 10.3a currents I_x and I_y stand for input currents of identical x-

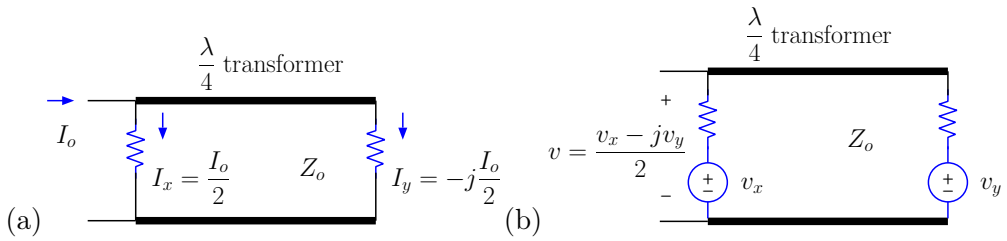


Figure 10.3: The feed network of a *cross-dipole* antenna with (a) radiating elements and (b) receiving elements. The antenna, which consists of a pair of x- and y-polarized dipoles with I_x and I_y current inputs, radiates and receives circular polarized waves in $\pm z$ directions.

and y-polarized dipoles located at the origin. In Figure 10.3b V_x and V_y represent the open-circuit voltage response of the same two dipole elements to incoming radiation. In this case, the feed network does not construct an array (because the elements are co-located) but generates a *cross-dipole* antenna that radiates circular polarized waves in $\pm z$-directions.

Here is the explanation:

With an input $I_x = \frac{I_o}{2}$ the x-polarized dipole radiates a field

$$E_x \propto \frac{I_o}{2}\frac{e^{-jkz}}{z}$$

in z-direction. Likewise, the y-polarized dipole with input $I_y = -j\frac{I_o}{2}$ radiates

$$E_y \propto -j\frac{I_o}{2}\frac{e^{-jkz}}{z}.$$

Therefore, using superposition, the field radiated in z-direction is

$$\mathbf{E} \propto (\hat{x} - j\hat{y})\frac{e^{-jkz}}{z}.$$

This is a circular polarized phasor since the corresponding field vector rotates — instead of growing and shrinking — as a function of time.[2]

How does the same cross-dipole respond to radiation coming from z-direction? To find out let us express the incoming plane wave phasor as

$$\mathbf{E}(z) = (L(\hat{x} - j\hat{y}) + R(\hat{x} + j\hat{y}))e^{jkz} = ((L+R)\hat{x} + j(R-L)\hat{y})e^{jkz},$$

which, depending on scalars L and R, could be linear, circular, or elliptically polarized. Now, as in the previous section, the open-circuit voltage of two dipoles are $V_x \propto \hat{x}\cdot\mathbf{E}(0) =$

[2]The cross-dipole will radiate elliptically polarized waves in other directions. The radiation is circular polarized only in broadside directions $\pm z$.

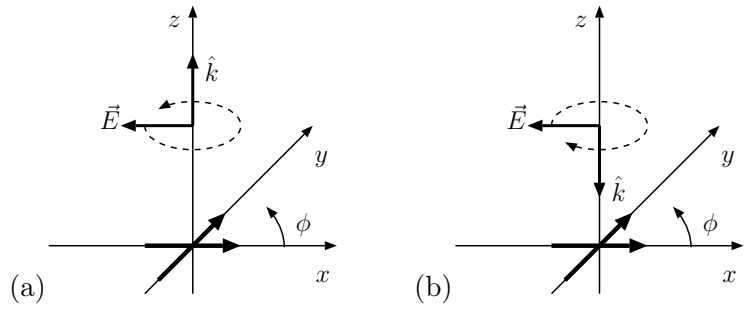


Figure 10.4: (a) Cross-dipole described in Figure 10.3 radiates a field in z-direction that rotates in ϕ-direction, and (b) the same antenna only responds to fields arriving from z-direction that rotate opposite to ϕ-direction. However, both radiated and received fields are right-handed. A left-handed cross-dipole can be constructed by exchanging the positions of x- and y-polarized dipoles in the feed network shown in Figure 10.3.

$(L+R)$ and $V_y \propto \hat{y} \cdot \mathbf{E}(0) = j(R-L)$. But the open-circuit voltage of the cross-dipole taken as a unit is (as indicated in Figure 25b)

$$V = \frac{V_x - jV_y}{2} \propto \frac{(L+R) - j(j(R-L))}{2} = \frac{(L+R)+(R-L)}{2} = R.$$

Evidently, the cross-dipole responds only to $\hat{x} + j\hat{y}$ polarized component of the incoming field with R amplitude—the second component $\propto \hat{x} - j\hat{y}$ appears to be "cross-polarized" since it is rejected (by the feed structure of the antenna).

In conclusion then, the cross-dipole antenna described in Figure 10.3 radiates and receives circular polarized fields $\propto \hat{x} - j\hat{y}$ and $\hat{x} + j\hat{y}$, respectively, which rotate in opposite directions[3] (see Chapter 1 and Figure 10.4).

Circular polarized antennas have some interesting properties. For instance, when they are illuminated by linear polarized fields they only receive half the available power (because of partial polarization mismatch). Somewhat more interesting, circular polarized antennas are "blind" to the reflections of their own radiations. The reason: reflected field rotates in the same direction as the radiated field and therefore it is fully cross-polarized. A radar using a right-hand polarized transmitting antenna needs a left-hand polarized (or linear polarized) receiving antenna to detect the reflected or backscattered returns. Finally, *any* linear polarized antenna will extract half the available power form *any* circular polarized field—thus if the orientation of a linear polarized receiver is unknown the best strategy is to transmit circular, as most FM stations do nowadays.

[3]Even though, *by convention*, they are both right-handed fields when their opposite propagation directions are taken into account.

11 Appendix V: Plane wave expansion of spherical waves — Weyl's identity

The field of a Hertzian dipole located at the origin is proportional to a scalar spherical wave function

$$V(r) = \frac{e^{-jkr}}{r},$$

where $r = \sqrt{x^2+y^2+z^2}$, which is a solution of the wave equation

$$\nabla^2 V(\mathbf{r}) + k^2 V(\mathbf{r}) = 0$$

away from the origin[1]. The same wave equation has also plane wave solutions of the form

$$V(x,y,z) = e^{-j(k_x x + k_y y + k_z z)},$$

such that $k^2 = k_x^2 + k_y^2 + k_z^2$, propagating in all directions indicated by the wavevector $\mathbf{k} = k_x\hat{x} + k_y\hat{y} + k_z\hat{z}$.

It should surely be possible to express the spherical wave solution e^{-jkr}/r as a weighted superposition of all $e^{-j(k_x x + k_y y \pm k_z z)}$ propagating away from $z = 0$ surface above (top sign) and below (bottom sign), i.e., have an identity of the form

$$\frac{e^{-jkr}}{r} = \int_{-\infty}^{\infty} \frac{dk_x}{2\pi} \int_{-\infty}^{\infty} \frac{dk_y}{2\pi} F(k_x,k_y) e^{-j(k_x x + k_y y \pm k_z z)}$$

with an appropriate choice of the weighting function $F(k_x, k_y)$. To determine $F(k_x, k_y)$ we first evaluate the expression above at $z = 0$ to obtain

$$\frac{e^{-jk\rho}}{\rho} = \int_{-\infty}^{\infty} \frac{dk_x}{2\pi} \int_{-\infty}^{\infty} \frac{dk_y}{2\pi} F(k_x,k_y) e^{-j(k_x x + k_y y)}$$

where $\rho \equiv \sqrt{x^2+y^2}$. The integral on the right is the 2D inverse Fourier transform of $F(k_x, k_y)$ and therefore $F(k_x, k_y)$ can be found as the 2D forward Fourier transform

$$F(k_x,k_y) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy \frac{e^{-jk\rho}}{\rho} e^{j(k_x x + k_y y)}.$$

By moving to polar coordinates by replacing $dxdy$ by $\rho d\phi d\rho$ while $x = \rho\cos\phi$ and $y = \rho\sin\phi$, and also using $k_x = q\cos\phi'$ and $k_y = q\sin\phi'$, where $q = \sqrt{k_x^2 + k_y^2}$, we have

$$F(k_x,k_y) = \int_0^{2\pi} d\phi \int_0^{\infty} d\rho e^{-jk\rho} e^{jq\rho(\cos\phi\cos\phi' + \sin\phi\sin\phi')} = \int_0^{2\pi} d\phi \int_0^{\infty} d\rho e^{-j\rho(k - q\cos(\phi-\phi'))},$$

[1] And also satisfy $\nabla^2 V(\mathbf{r}) + k^2 V(\mathbf{r}) = -4\pi\delta(\mathbf{r})$ *everywhere* in 3D space.

and the integral on the right evaluates to

$$\int_0^\infty d\rho e^{-j\rho(k-q\cos(\phi-\phi'))} = \frac{e^{-j\rho(k-q\cos(\phi-\phi'))}}{-j(k-q\cos(\phi-\phi'))}|_{\rho=0}^{\infty} = \frac{1}{j(k-q\cos(\phi-\phi'))},$$

within a region of convergence (ROC), $\mathrm{Im}\{k\}<0$, associated with damped e^{-jkr}/r waves. Consequently we find

$$F(k_x,k_y) = \int_0^{2\pi} d\phi \frac{1}{j(k-q\cos(\phi-\phi'))} = \frac{-2\pi j}{\sqrt{k^2-q^2}} = \frac{-2\pi j}{\sqrt{k^2-k_x^2-k_y^2}} = \frac{-2\pi j}{k_z}$$

after using a well known result

$$\int_0^{2\pi} \frac{dx}{a+b\cos(x)} = \frac{2\pi}{\sqrt{a^2-b^2}}$$

available in most definite integral tables.

Thus, using $F(k_x,k_y) = -2\pi j/k_z$, which is an *analytic function*[2] of complex valued k, we obtain *Weyl's identity*

$$\frac{e^{-jkr}}{r} = \frac{-j}{2\pi}\int_{-\infty}^{\infty} dk_x \int_{-\infty}^{\infty} dk_y \frac{e^{-j(k_x x+k_y y\pm k_z z)}}{k_z}$$

that expresses a damped outgoing spherical wave (rippling away from the origin) as a superposition of damped plane waves $e^{-j(k_x x+k_y y\pm k_z z)}$ propagating away from the $z=0$ surface above (top sign) and below (bottom sign). However, this identity is also valid for undamped waves e^{-jkr}/r with real valued k (in terms of undamped plane waves) permitted by *analytic continuation*[3] of the 2D Fourier transform $F(k_x,k_y) = -2\pi j/k_z$ from the bottom half of the complex k plane to the real k axis — or in simpler terms,

[2] A complex valued function $f(x,y) = u(x,y) + jv(x,y)$ of real variables x and y, where functions $u(x,y)$ and $v(x,y)$ are real valued, is called *analytic* (or *holomorphic* or *regular*) if it can be shown that $f(x,y) = f(x+jy) \equiv f(z)$, $z \equiv x+jy$. Analytic functions $f(x,y) = f(z)$ have uniquely defined derivatives df/dz and consequently they satisfy *Cauchy-Reimann conditions* $\partial u/\partial x = \partial v/\partial y$ and $\partial u/\partial y = -\partial v/\partial x$. They are infinitely differentiable and their path integrals performed on the complex plane are path independent so long as the path does not pass through singularity points. Furthermore, if $f(z)$ is analytic, its imaginary part can be obtained from its real part (using Cauchy-Riemann conditions and up to an additive constant) and vice versa. If an analytic $f(z)$ is known only over some region D of the complex plane, it can be uniquely determined outside this region by a method known as *analytic continuation*.

[3] Just as $\frac{1}{s}$ represents the unique "analytic continuation" of the Laplace transform $\hat{H}(s)$ of $h(t) = u(t)$ outside its ROC $\mathrm{Re}\{s\} > 0$ (see *Kudeki and Munson* (2009), pp 365; https://ccrma.stanford.edu/~jos/filters/Analytic_Continuation.html), $\frac{-2\pi j}{k_z}$ is the unique analytic continuation of the Fourier transform $F(k_x,k_y)$ of $\frac{e^{-jk\rho}}{\rho}$ outside its ROC $\mathrm{Im}\{k\}<0$. Just as $\frac{1}{j\omega}$ describes correctly the "frequency response" of an integrator system with the "transfer function" $\frac{1}{s}$, function $\frac{-2\pi j}{k_z}$ evaluated with real valued k describes correctly the plane wave amplitudes needed to synthesize undamped spherical waves.

In general, the knowledge of an analytic function within *any* portion of the complex plane is sufficient to determine the function everywhere else on the plane uniquely, by performing a sequence

by considering undamped waves to be an idealization of waves with vanishingly weak damping (ignoring a small negative k_i in k).

Replacing j with $-i$ in Weyl's identity, we get

$$\frac{e^{ikr}}{r} = \frac{i}{2\pi}\int_{-\infty}^{\infty} dk_x \int_{-\infty}^{\infty} dk_y \frac{e^{i(k_x x + k_y y \pm k_z z)}}{k_z},$$

which also expresses the same fact, but with the sign convention preferred in physics literature where $e^{-i\omega t}$ time variation is assumed instead of $e^{j\omega t}$ of electric engineering literature — in these notes we will go back and forth between these two conventions as needed by replacing all $-j$ with $+i$, or vice versa.

Let's now transform the k_x-k_y integral in Weyl's identity to a θ-ϕ integral such that $k_x = k\sin\theta\cos\phi$, $k_y = k\sin\theta\sin\phi$, meaning that, using a Jacobian, we have

$$\begin{aligned} dk_x dk_y &= |\det\begin{bmatrix} \partial k_x/\partial\theta & \partial k_x/\partial\phi \\ \partial k_y/\partial\theta & \partial k_y/\partial\phi \end{bmatrix}| \, d\theta d\phi = |\det\begin{bmatrix} k\cos\theta\cos\phi & -k\sin\theta\sin\phi \\ k\cos\theta\sin\phi & k\sin\theta\cos\phi \end{bmatrix}| \, d\theta d\phi \\ &= k_z k \sin\theta \, d\theta d\phi \end{aligned}$$

where $k_z = \sqrt{k^2 - k_x^2 - k_y^2} = k\cos\theta$ — note that $k_z k\sin\theta\, d\theta d\phi$ is $k\sin\theta d\phi$ times the $\cos\theta$ projection of $kd\theta$, the dimensions of an infinitesimal area element on the k_x-k_y surface to replace the area element $dk_x dk_y$. Since k_z varies first from k to 0 and then from 0 to $i\infty$ as k_x,k_y vary from 0 to $\pm\infty$, θ first varies from 0 to $\pi/2$ and then varies as a complex angle $\pi/2 + i\theta_i$ to reach $\pi/2 - i\infty$ as shown in Figure 11.1. We are therefore able to express Weyl's identity as

$$\frac{e^{ikr}}{r} = \frac{ik}{2\pi}\int_0^{2\pi} d\phi \int_0^{\pi/2-i\infty} d\theta \, \sin\theta \, e^{i(k_x x + k_y y + k_z|z|)}$$

in terms of angles ϕ and θ, the latter being complex valued. Angles θ from 0 to $\pi/2$ contribute to the integral propagating uniform plane waves while from $\pi/2$ to $\pi/2 - i\infty$

---

of Taylor series expansions — always possible because analytic functions are infinitely differentiable — of the function near the edge of the current ROC, except for the points where the function is singular — see, e.g., *Byron and Fuller* [1992], for a discussion of this property of analytic functions and *analytic continuation*. Related to this is *identity theorem* (see `https://en.wikipedia.org/wiki/Identity_theorem`) which states that if $f(z)$ and $g(z)$ are analytic in some region D on the complex plane and $f(z) = g(z)$ within some subdomain of D, then it is necessary that $f(z) = g(z)$ everywhere within D — this is true because $h(z) \equiv f(z) - g(z) = 0$ within the subdomain implies that n-th order derivatives $h^{(n)}(z_o) = 0$ for all n with z_o anywhere within the subdomain, which in turn implies that Taylor series extension of $h(z)$ to the rest of D results in $h(z) = 0$ (and thus $f(z) = g(z)$) everywhere in D.

Returning to the derivation of Weyl's identity, there is some $F(k_x, k_y)$ function depending on a real variable k that describes the plane wave expansion of undamped spherical waves in 3D space; this function has a unique analytic continuation into the domain of complex $k = k_r + jk_i$, an extension also called $F(k_x, k_y)$ and used to describe the expansion of damped ($k_i < 0$) and growing ($k_i > 0$) waves; this extended $F(k_x, k_y)$ having the same algebraic dependence on variable k as the original $F(k_x, k_y)$ is readily identifiable with a Fourier transform operation within an ROC of damped waves ($k_i < 0$), but it also works, for obvious reasons, when it is used with $k_i = 0$ and $k_r = k$ to describe undamped spherical waves (and even growing waves if $k_i > 0$ is taken).

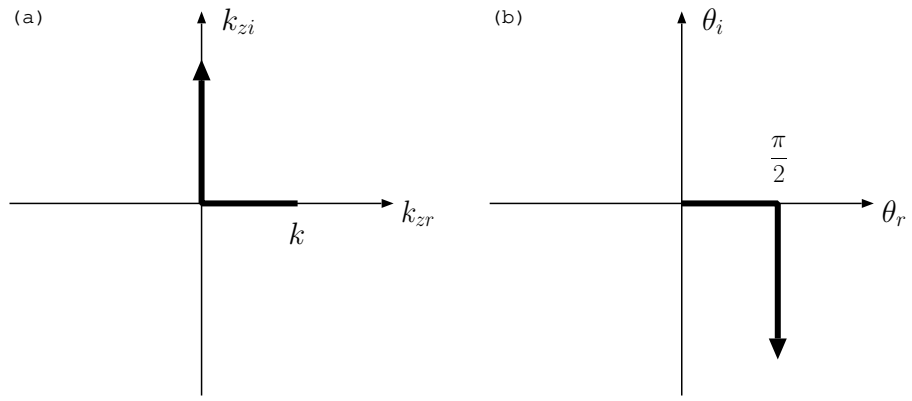


Figure 11.1: (a) Path of $k_z = k_{zr} + ik_{zi}$ on the complex plane as k_x and k_y in Weyl's identity are swept from zero to ∞, (b) integration path on complex $\theta = \theta_r + i\theta_i$ plane used in Weyl's identity as θ is swept from 0 to $\pi/2$ and then from $\pi/2$ to $\pi/2 - i\infty$.

evanescent waves decaying away from the $z = 0$ plane. On the right hand side of Weyl's identity we have a complex plane contour integral over the specified 0 to $\pi/2 - i\infty$ path. Notice that we have removed the pole at $k_z = 0$ by this change of variables at the expense of transforming an ordinary double integral over real $k_x k_y$ to one to be performed as a contour integral on the complex plane — we will illustrate how to handle such integrals in the next section.

12 Appendix VI: Complex functions

Here we provide a quick review of derivatives and path integrals of complex functions of complex variables carried out on the complex plane. This is an important topic covered in MATH 448 rigorously.

- A complex valued function

$$f(x,y) = u(x,y) + jv(x,y)$$

of real variables x and y, where functions $u(x,y)$ and $v(x,y)$ are real valued, is called *analytic* (or *holomorphic* or *regular*) if it can be shown that

$$f(x,y) = f(x+jy) \equiv f(z),\ z \equiv x+jy.$$

  - Functions $f(x,y) = z$ and $f(x,y) = z^2$ are analytic but $f(x,y) = x - jy = z^*$ and $f(x,y) = zz^* = x^2 + y^2$ are not, whereas $f(x,y) = \frac{1}{x+jy-1} = \frac{1}{z-1}$ is analytic across the entire complex plane with the exception of the singularity point $z = 1$ where $\frac{1}{z-1}$ is undefined.

- Analytic functions $f(x,y) = f(z)$ have uniquely defined derivatives df/dz and consequently they satisfy the *Cauchy-Reimann conditions*

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \text{ and } \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$$

following from

$$\frac{df}{dz} = \lim_{\Delta x \to 0} \frac{f(z+\Delta x) - f(z)}{\Delta x} = \lim_{\Delta y \to 0} \frac{f(z+j\Delta y) - f(z)}{j\Delta y}$$

  - It follows that u and v satisfy the 2D *Laplace's equations*

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \text{ and } \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0$$

just like electrostatic potential $\phi(x,y)$ in 2D away from sources (in charge free regions) — hence $\nabla u = u_x\hat{x} + u_y\hat{y}$ as well as $\nabla v = v_x\hat{x} + v_y\hat{y}$ are valid electrostatic fields that could exist in charge free regions.

- Path integrals of analytic $f(z)$ performed on the complex z-plane,

$$\int_A^B f(z)dz = \int_A^B (u+jv)(dx+jdy),$$

are *path independent* — as can be seen by applying 2D Stoke's thm to a closed contour together with Cauchy-Reimann conditions — so long as the path does not pass through singularity points. Hence:

1. $$\int_A^B f(z)dz = \int_A^B dF = F(b) - F(a)$$
   in terms of a complex differentiable anti-derivative function $F(z) = U(x,y) + jV(z,y)$ such that
   $$\frac{dF}{dz} = f(z) \text{ as well as } \frac{\partial U}{\partial x} = \frac{\partial V}{\partial y} \text{ and } \frac{\partial U}{\partial y} = -\frac{\partial V}{\partial x}$$
   since $F(z)$ must also be analytic having a uniquely defined derivative $f(z)$ on the complex plane,
2. the loop integral around any closed loop C,
   $$\oint_C f(z)dz = 0, \quad \textbf{Cauchy integral theorem}$$
   when $f(z)$ is analytic on and inside contour C.
   - **Outline of a proof (to be completed in HW) in analogy with *electrostatics* (covered in ECE 329):**

   a) given an analytic function $f(z) = u(x,y) + jv(x,y)$, having an anti-derivative $F = U + jV$, consider a pair of electric field vectors
   $$\mathbf{E}_1 \equiv \nabla U = u\hat{x} - v\hat{y}$$
   and
   $$\mathbf{E}_2 \equiv \nabla V = v\hat{x} + u\hat{y},$$
   built with u and v of $f(z)$;

   b) using the Cauchy-Reimann conditions for u and v, show that both $\mathbf{E}_1$ and $\mathbf{E}_2$ satisfy $\nabla \times \mathbf{E} = 0$ as well as $\nabla \cdot \mathbf{E} = 0$, conditions for static electric fields in source-free regions, implying that

   c) $\mathbf{E}_1$ and $\mathbf{E}_2$ must be circulation-free, that is $\oint_C \mathbf{E}_1 \cdot d\mathbf{l} = 0$ and $\oint_C \mathbf{E}_2 \cdot d\mathbf{l} = 0$,

   d) which, after an expansion in terms of u and v, can be used to show that the closed path integral
   $$\oint_C f(z)dz = \oint_C (u+jv)(dx+jdy) = \oint_C \underbrace{(udx - vdy)}_{\mathbf{E}_1 \cdot d\mathbf{l}} + j \oint_C \underbrace{(vdx + udy)}_{\mathbf{E}_2 \cdot d\mathbf{l}} = 0 \quad \textbf{QED}$$

- Furthermore, when $f(z)$ is analytic, knowing its real part is sufficient to derive its imaginary part (using Cauchy-Riemann conditions and up to an additive constant) and vice versa.
- If an analytic $f(z)$ is known only over some domain D of complex z-plane, it can be uniquely determined outside this domain by a method known as *analytic continuation*.

  - One use of analytic continuation — a trivial one — is to turn a real valued function $f(x)$ of real variable x into an analytic complex function $f(z)$ by simply replacing x in $f(x)$ by $z = x + jy$, where y is real valued;
    * e.g., $f(z) = e^z = e^{x+jy}$ is analytic continuation of $f(x) = e^x$,
    * but $f(z^*) = e^{z^*} = e^{x-jy}$ is not because $f(z^*)$ is not even analytic;
    * real function $f(x) = \cos(x) = \frac{e^{jx}+e^{-jx}}{2}$ becomes complex $f(\theta) = \cos(\theta) = \frac{e^{j\theta}+e^{-j\theta}}{2}$, having an analytic derivative $\frac{df}{d\theta} = \frac{je^{j\theta}-je^{-j\theta}}{2} = -\frac{e^{j\theta}-e^{-j\theta}}{2j} = -\sin(\theta)$, where $\sin\theta$ is analytic continuation of real $\sin(x) = \frac{e^{jx}-e^{-jx}}{2j}$ across the complex $\theta = x + iy$ plane — complex valued angles θ arise in this context.
  - Another use is to extend analytic $f(z)$ obtained by some transformation or infinite sum within some region of convergence (ROC) to the entire complex plane including the real line.

- If $g(z)$ is analytic on and within a contour C, $f(z) = g(z)/(z - z_o)$ will also be so except at the pole location $z = z_o$, and if z_o is contained within C, then, assuming a **counterclockwise** *tour* around C (by applying Cauchy integral thm with a path excluding z_o),

$$2\pi j g(z_o) = \oint_C \frac{g(z)}{z - z_o} dz \quad \textbf{Cauchy integral formula}$$

$$\dots$$

$$2\pi j \underbrace{\frac{g^{(n)}(z_o)}{n!}} = \oint_C \frac{g(z)}{(z - z_o)^{n+1}} dz$$

residue of $g(z)/(z - z_o)^{n+1}$ **at** z_o, with $g^{(n)}(z_o)$ the n-th derivative.

  - Consequently,

$$\oint_C f(z)dz = 2\pi j \times \sum \text{residues of all poles within } C$$

    * The residue $\frac{g^{(n)}(z_o)}{n!}$ is also the coefficient c_{-1} of the $\frac{1}{z-z_o}$ term in the power series $f(z) = \sum_{p=-\infty}^{\infty} c_p(z - z_o)^p$ expansion of $f(z)$ about the z_o pole of some order n.
    * To apply these contour integration techniques to multi-valued analytic functions such as $f(z) = z^{1/2}$ and $f(z) = \ln(z)$ concepts such as *branch points*, *branch cuts*, and *Riemann sheets* need to be introduced — see, e.g., Mathematical Physics, Kusse and Westwig (2008).

- Since analytic functions have analytic derivatives, they are differentiable an infinite number of times (residue formula is also n-th derivative formula for $g(z_o)$ with arbitrary n), and, accordingly, play an essential role in physical models built with differentiable functions (i.e., solutions of differential equations).

13 Appendix VII: Space wave and surface wave links

The z component of the radiation field of a $\hat{z}$ polarized Hertzian dipole located at $z = z_0$ will be, at a location $\mathbf{r} = (x, y, z)$,

$$E_z(x, y, z) = -\eta_o j I_o k \Delta z \sin^2\theta_0 \frac{e^{-jkR_0}}{4\pi R_0},$$

where $R_0 = \sqrt{x^2 + y^2 + (z - z_0)^2}$ and $\cos\theta_0 = (z - z_0)/R_0$ — see Figure 13.1.

Using Weyl's identity and physics notation, we can express this as

$$E_z = \eta_o i I_o \frac{k\Delta z}{4\pi} \sin^2\theta_0 \frac{e^{ikR_0}}{R_0} = \eta_o i I_o \frac{k\Delta z}{4\pi} \sin^2\theta_0 \frac{ik}{2\pi} \int_0^{2\pi} d\phi \int_0^{\pi/2-i\infty} d\theta \, \sin\theta \, e^{i(k_x x + k_y y + k_z|z-z_0|)}.$$

The open circuit voltage of a $\hat{z}$ polarized Hertzian dipole antenna at $\mathbf{r} = (x, y, z)$ will simply be $E_z \Delta z$ if there are no reflecting surfaces near these dipoles.

However, if a reflecting surface were located on $z = 0$ plane, where downgoing plane wave components of E_z shown explicitly under the $\phi\theta$-integral above would reflect at an angle θ, with a reflection coefficient $\Gamma_v(\theta)$, we would have E_z altered as a "Sommerfeld integral"

$$E_z = \eta_o i I_o \frac{k\Delta z}{4\pi} \sin^2\theta_0 \frac{ik}{2\pi} \int_0^{2\pi} d\phi \int_0^{\pi/2-i\infty} d\theta \, \sin\theta \, \{e^{i(k_x x + k_y y + k_z|z-z_0|)} + \Gamma_v(\theta) e^{i(k_x x + k_y y + k_z|z+z_0|)}\},$$

where the added term of reflected waves with relative amplitudes $\Gamma_v(\theta)$ synthesizes a spherical wave that appears to be spreading out from the image point $z = -z_0$ of the dipole located at $z = z_0$. Hence we can re arrange E_z as

$$E_z = \eta_o i I_o \frac{k\Delta z}{4\pi} \sin^2\theta_0 \{\frac{e^{ikR_0}}{R_0} + \frac{\sin^2\theta_1}{\sin^2\theta_0}\Gamma_v(\theta_1)\frac{e^{ikR_1}}{R_1} + \frac{ik}{2\pi} \int_0^{2\pi} d\phi \int_0^{\pi/2-i\infty} d\theta \, \sin\theta \, (\Gamma_v(\theta) - \frac{\sin^2\theta_1}{\sin^2\theta_0}\Gamma_v(\theta_1)) e^{i(k_x x + k_y y + k_z|z+z_0|)}\}$$

where the angles θ_0 and θ_1 are the zenith angle of the observation point seen from the dipole and its image location, respectively, and $R_1 = \sqrt{x^2 + y^2 + (z + z_0)^2}$ is the distance from image location. The first two terms above, namely,

$$E_{sw} \equiv \eta_o i I_o \frac{k\Delta z}{4\pi} \{\sin^2\theta_0 \frac{e^{ikR_0}}{R_0} + \sin^2\theta_1 \Gamma_v(\theta_1) \frac{e^{ikR_1}}{R_1}\}$$

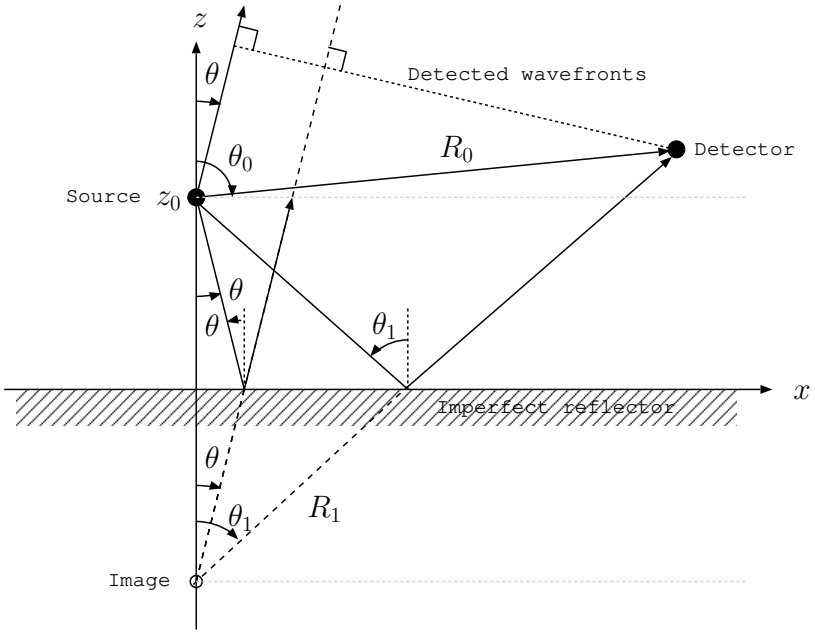


Figure 13.1: A $\hat{z}$-polarized (vertically) dipole source at $(x, z) = (0, z_0)$ and its image at $(x, z) = (0, -z_0)$ given a homogeneous reflector in $z < 0$ half-space. A reflection coefficient $\Gamma_v(\theta)$ relates the $\hat{z}$-components of the electric fields of the incident and reflected plane waves (uniform and non-uniform) from the $z = 0$ surface. Space waves reaching from the source and its image to a detection point at (x, z), $z \geq 0$, will cover distances R_0 and R_1, respectively. The source and its image see the detection point at zenith angles θ_0 and θ_1, respectively. For $R_0 \gg z, z_0$, $\theta_0 \approx \theta_1 \approx 90^\circ$. Note that angle θ concerns the plane waves included in Weyl's identity covering the 0 to $\pi/2$ and $\pi/2$ to $\pi/2 - i\infty$ range — for upgoing and downcoming plane waves included in Weyl's identity (for $z > z_o$ and $z < z_o$, respectively) θ is defined separately as indicated in this diagram.

justifies the space-wave link model discussed in Chapter 2, whereas the third term is a "ground wave" term

$$E_{gw} \equiv \eta_o i I_o \frac{k\Delta z}{4\pi} \sin^2\theta_0 \{ \frac{ik}{2\pi} \int_0^{2\pi} d\phi \int_0^{\pi/2-i\infty} d\theta \, \sin\theta \, (\Gamma_v(\theta) - \frac{\sin^2\theta_1}{\sin^2\theta_0} \Gamma_v(\theta_1)) e^{i(k_x x + k_y y + k_z |z+z_0|)} \}$$

that was "missed" in the simplistic derivation of the space wave model earlier on. It can be shown that the ground wave term is negligible away from the $z = 0$ surface, but for $R_1 \gg z_0, z$ and as $\theta_1 \to 90^\circ$ we have $\Gamma(\theta_1) \to -1$ (for both vertical and horizontal polarizations) in which case E_{sw} vanishes and only E_{gw} remains in the form

$$E_{gw} = \eta_o i I_o \frac{k\Delta z}{4\pi} \{ \frac{ik}{2\pi} \int_0^{2\pi} d\phi \int_0^{\pi/2-i\infty} d\theta \, \sin\theta \, (\Gamma_v(\theta) + \frac{\sin^2\theta_1}{\sin^2\theta_0}) e^{i(k_x x + k_y y + k_z |z+z_0|)} \}$$

for vertical polarized link. Let's examine the integral

$$I \equiv \frac{ik}{2\pi}\int_0^{2\pi} d\phi \int_0^{\pi/2-i\infty} d\theta\, \sin\theta\,(\Gamma_v(\theta)+\frac{\sin^2\theta_1}{\sin^2\theta_0})e^{i(k_x x+k_y y+k_z|z+z_0|)}$$

in some detail taking into account that (see R_v in Chapter 2)

$$\Gamma_v(\theta)=\frac{n\cos\theta-1}{n\cos\theta+1} \;\Rightarrow\; \Gamma_v(\theta)+\frac{\sin^2\theta_1}{\sin^2\theta_0}=\frac{n\cos\theta-1}{n\cos\theta+1}+\frac{\sin^2\theta_1}{\sin^2\theta_0}=1+\frac{\sin^2\theta_1}{\sin^2\theta_0}-\frac{2}{n\cos\theta+1}.$$

Consequently we can write

$$I=(1+\frac{\sin^2\theta_1}{\sin^2\theta_0})\frac{e^{ikR_1}}{R_1}-\frac{ik}{\pi}\int_0^{2\pi} d\phi\int_0^{\pi/2-i\infty} d\theta\,\frac{\sin\theta}{n\cos\theta+1}\,e^{i(k_x x+k_y y+k_z|z+z_0|)}$$

where, we notice, $|z+z_0|=R_1\cos\theta_1$ and hence, also using $x=\rho\cos\phi_1$ and $y=\rho\sin\phi_1$, $\rho=\sqrt{x^2+y^2}=R_1\sin\theta_1$, and $k_z=k\cos\theta$, $k_x=k\sin\theta\cos\phi$, and $k_y=k\sin\theta\sin\phi$,

$$\begin{aligned} I-(1+\frac{\sin^2\theta_1}{\sin^2\theta_0})\frac{e^{ikR_1}}{R_1} &= -\frac{ik}{\pi}\int_0^{\pi/2-i\infty} d\theta\,\frac{\sin\theta}{n\cos\theta+1}\,e^{ikR_1\cos\theta\cos\theta_1}\int_0^{2\pi} d\phi\, e^{ik\rho\sin\theta(\cos\phi\cos\phi_1+\sin\phi\sin\phi_1)} \\ &= -\frac{ik}{\pi}\int_0^{\pi/2-i\infty} d\theta\,\frac{\sin\theta}{n\cos\theta+1}\,e^{ikR_1\cos\theta\cos\theta_1}2\pi J_o(k\rho\sin\theta) \end{aligned}$$

since $\cos\phi\cos\phi_1+\sin\phi\sin\phi_1=\cos(\phi-\phi_1)$ and

$$J_n(z)\equiv\frac{1}{2\pi i^n}\int_0^{2\pi} d\phi e^{iz\cos\phi}e^{in\phi}$$

is the Bessel function of the first kind of order n. Furthermore, using $J_0(u)=\frac{1}{2}(H_0^{(1)}(u)+H_0^{(2)}(u))$, where $H_0^{(2)}(-u)=-H_0^{(1)}(u)$ are Hankel functions with $H_0^{(1)}(u)\approx\sqrt{\frac{2}{\pi u}}e^{i(u-\frac{\pi}{4})}$ for large $|u|$, we have

$$\begin{aligned} I-(1+\frac{\sin^2\theta_1}{\sin^2\theta_0})\frac{e^{ikR_1}}{R_1} &= -ik\int_0^{\pi/2-i\infty} d\theta\,\frac{\sin\theta}{n\cos\theta+1}\,e^{ikR_1\cos\theta\cos\theta_1}(H_0^{(1)}(k\rho\sin\theta)+H_0^{(2)}(k\rho\sin\theta)) \\ &= -ik\int_0^{\pi/2-i\infty} d\theta\,\frac{\sin\theta}{n\cos\theta+1}\,e^{ikR_1\cos\theta\cos\theta_1}(H_0^{(1)}(k\rho\sin\theta)+H_0^{(1)}(-k\rho\sin\theta)) \\ &= -ik\int_{-\pi/2+i\infty}^{\pi/2-i\infty} d\theta\,\frac{\sin\theta}{n\cos\theta+1}\,e^{ikR_1\cos\theta\cos\theta_1}H_0^{(1)}(k\rho\sin\theta) \\ &= -ik\int_{-\pi/2+i\infty}^{\pi/2-i\infty} d\theta\,\frac{\sin\theta}{n\cos\theta+1}\,e^{ikR_1\cos\theta\cos\theta_1}\sqrt{\frac{2}{\pi k\rho\sin\theta}}e^{i(k\rho\sin\theta-\frac{\pi}{4})} \\ &= -ik\int_{-\pi/2+i\infty}^{\pi/2-i\infty} d\theta\,\frac{e^{-i\frac{\pi}{4}}}{n\cos\theta+1}\sqrt{\frac{2\sin\theta}{\pi kR_1\sin\theta_1}}\,e^{ikR_1(\cos\theta\cos\theta_1+\sin\theta\sin\theta_1)}. \end{aligned}$$

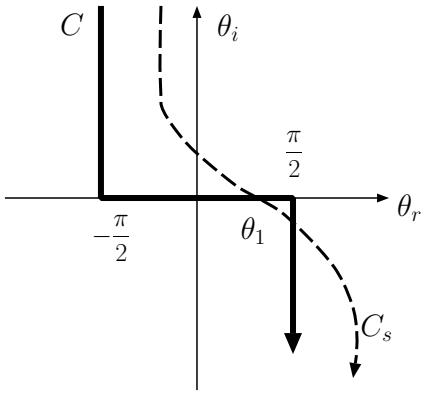


Figure 13.2: Original path of integration C on complex θ-plane and modified integration path C_s used in evaluating the integral I when using the *method of steepest descent*. Notice that the alternate path C_s (permitted for analytic integrand so long as a pole is not crossed) is taken to go through θ_1 with a -45° incline described by $\theta = \theta_1 + (1-i)s$ as the real valued parameter s varies from negative to positive values across $s = 0$.

The most significant contributions to this integral will come from the vicinity of $\theta = \theta_1$ as it becomes evident when it is expressed as

$$I - (1 + \frac{\sin^2\theta_1}{\sin^2\theta_0})\frac{e^{ikR_1}}{R_1} = -i\int_{-\pi/2+i\infty}^{\pi/2-i\infty} d\theta\, \frac{e^{-i\frac{\pi}{4}}}{n\cos\theta + 1}\sqrt{\frac{2k\sin\theta}{\pi R_1 \sin\theta_1}}\, e^{ikR_1\cos(\theta-\theta_1)}.$$

Notice that when kR_1 is large, the exponential $e^{ikR_1\cos(\theta-\theta_1)}$ in the integrand above will vary rapidly as θ departs from θ_1, and contributions from nearby portions of the integration path will be self cancelling. Also, since

$$\cos(\theta-\theta_1) \approx 1 - \frac{1}{2}(\theta-\theta_1)^2$$

for small $\theta - \theta_1$, we would have, on an alternate integration path — see Figure 13.2 — described by

$$\theta - \theta_1 = (1-i)s = \sqrt{2}e^{-i\frac{\pi}{4}}s,$$

where s is a real variable,

$$e^{ikR_1\cos(\theta-\theta_1)} \approx e^{ikR_1}e^{-kR_1 s^2},$$

meaning that the magnitude $e^{-kR_1 s^2}$ of $e^{ikR_1\cos(\theta-\theta_1)}$ would first rise to unity and then descend rapidly[1] as s varies from negative to positive values across $s = 0$. Since path integrals of analytic functions are path independent on the complex plane — just like path integrals of electrostatic potentials in 3D space — a distortion of the integration

[1] We are describing here an application of the *method of steepest descent* — see Berekhovskikh, Chapter 4; Jones (1972)

path above to what amounts to be following the real s axis of a complex variable s, after a change of variables affected by $\theta = \theta_1 + \sqrt{2}e^{-i\frac{\pi}{4}}s$ and $d\theta = \sqrt{2}e^{-i\frac{\pi}{4}}ds$, would lead to

$$\begin{aligned} I - (1 + \frac{\sin^2\theta_1}{\sin^2\theta_0})\frac{e^{ikR_1}}{R_1} &= -2i\int_{-\infty}^{\infty} ds \frac{e^{-i\frac{\pi}{2}}}{n\cos\theta + 1}\sqrt{\frac{k\sin\theta}{\pi R_1 \sin\theta_1}}\, e^{ikR_1}e^{-kR_1 s^2} \\ &= -2\frac{e^{ikR_1}}{R_1}\int_{-\infty}^{\infty} ds \frac{1}{n\cos\theta + 1}\sqrt{\frac{kR_1\sin\theta}{\pi\sin\theta_1}}\, e^{-kR_1 s^2}. \end{aligned}$$

Since only the $s \approx 0$ neighborhood contributes to this integral, we have $\theta \approx \theta_1$ and we may replace $\sin\theta$ by $\sin\theta_1$ above, leading to

$$I = (1 + \frac{\sin^2\theta_1}{\sin^2\theta_0})\frac{e^{ikR_1}}{R_1} - 2\frac{e^{ikR_1}}{R_1}\sqrt{\frac{kR_1}{\pi}}\int_{-\infty}^{\infty} ds \frac{1}{n\cos\theta + 1}\, e^{-kR_1 s^2}.$$

In the denominator

$$\begin{aligned} n\cos\theta + 1 &= n\cos(\theta_1 + \sqrt{2}e^{-i\frac{\pi}{4}}s) + 1 \\ &= n\cos\theta_1\cos(\sqrt{2}e^{-i\frac{\pi}{4}}s) - n\sin\theta_1\sin(\sqrt{2}e^{-i\frac{\pi}{4}}s) + 1 \\ &\approx n\cos\theta_1 + 1 - n\sin\theta_1\sqrt{2}e^{-i\frac{\pi}{4}}s = n\sin\theta_1\sqrt{2}e^{-i\frac{\pi}{4}}(\frac{n\cos\theta_1 + 1}{n\sin\theta_1\sqrt{2}e^{-i\frac{\pi}{4}}} - s) \end{aligned}$$

and therefore

$$I = (1 + \frac{\sin^2\theta_1}{\sin^2\theta_0})\frac{e^{ikR_1}}{R_1} + 2\frac{e^{ikR_1}}{R_1}\sqrt{\frac{kR_1}{2\pi}}\frac{e^{i\frac{\pi}{4}}}{n\sin\theta_1}\int_{-\infty}^{\infty} ds \frac{e^{-kR_1 s^2}}{s - s_p}$$

where

$$s_p \equiv \frac{n\cos\theta_1 + 1}{n\sin\theta_1\sqrt{2}e^{-i\frac{\pi}{4}}}$$

is the pole of the integrand above.

The remaining integral above is tabulated in terms of *complementary error function*

$$\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^{\infty} e^{-t^2}dt$$

(available in Matlab, Python, etc.) as

$$\int_{-\infty}^{\infty} ds \frac{e^{-\xi s^2}}{s - s_p} = \pm i\pi e^{-\xi s_p^2}\mathrm{erfc}(\mp i s_p\sqrt{\xi})$$

(e.g., Jones, 1972, page 309) with the upper (lower) sign when $\mathrm{Im}\{s_p\}>0$ (<0).

In our case $\xi = kR_1$ and $\mathrm{Im}\{s_p\}>0$ and therefore

$$I = (1 + \frac{\sin^2\theta_1}{\sin^2\theta_0})\frac{e^{ikR_1}}{R_1} + 2\frac{e^{ikR_1}}{R_1}\sqrt{\frac{\pi kR_1}{2}}\frac{ie^{i\frac{\pi}{4}}}{n\sin\theta_1}e^{-kR_1 s_p^2}\mathrm{erfc}(-is_p\sqrt{kR_1}).$$

To confirm this, and to acquire the standard notation used in the literature for the Norton surface or ground wave, we note that

$$n = \sqrt{1 + \frac{\sigma}{j\omega\epsilon_o}} = \sqrt{1 + i\frac{\sigma}{\omega\epsilon_o}} \;\Rightarrow\; -in^2 = \frac{\sigma}{\omega\epsilon_o} - i \approx \frac{\sigma}{\omega\epsilon_o}e^{-ib},$$

with $b \equiv \tan^{-1}\frac{\omega\epsilon_o}{\sigma}$, when σ is large. Hence $n^2 \approx i\frac{\sigma}{\omega\epsilon_o}e^{-ib} \equiv i|n|^2e^{-ib}$, and in the limit $z, z_0 \to 0$, $\theta_0, \theta_1 \to 90°$, in which case $\sin\theta_1 \to 1$ and $R, R_1 \to r = x$, we also have

$$s_p = \frac{n\cos\theta_1 + 1}{n\sin\theta_1\sqrt{2}e^{-i\frac{\pi}{4}}} \to \frac{e^{i\frac{\pi}{4}}}{\sqrt{2}n} = \frac{e^{ib/2}}{\sqrt{2}|n|}$$

with a positive imaginary part due to b. Furthermore,

$$kR_1s_p^2 = \frac{kx}{2|n|^2}e^{ib} \equiv pe^{ib},$$

where $p = kx/2|n|^2$is said to be the *numerical distance*. Hence in the same limit,

$$\begin{aligned} I &= 2\frac{e^{ikx}}{x}\{1 + \sqrt{\frac{\pi kx}{2}}\frac{ie^{i\frac{\pi}{4}}}{n}e^{-kR_1s_p^2}\mathrm{erfc}(-is_p\sqrt{kR_1})\} \\ &= 2\frac{e^{ikx}}{x}\{1 + \sqrt{\frac{\pi kx}{2}}\frac{ie^{i\frac{\pi}{4}}}{e^{i\frac{\pi}{4}}|n|e^{-ib/2}}e^{-pe^{ib}}\mathrm{erfc}(-i\sqrt{p}e^{ib/2})\} \\ &= 2\frac{e^{ikx}}{x}\{1 + i\sqrt{\pi p}e^{ib/2}e^{-pe^{ib}}\mathrm{erfc}(-i\sqrt{p}e^{ib/2})\}. \end{aligned}$$

This brings us to our final result, namely,

$$E_{gw} \approx 2E_o(x)A_s,$$

where $E_o(x)$ is the broadside field of the dipole at distance x in vacuum, $2E_o(x)$ the same field when the dipole is located and dipole field detected just above a PEC reflector, and

$$A_s \equiv 1 - j\sqrt{\pi p}e^{-jb/2}e^{-pe^{-jb}}\mathrm{erfc}(j\sqrt{p}e^{-jb/2})$$

is the *ground-wave attenuation factor*[2] when PEC is replaced with a conductor of refractive index n used to compute p and b.

This factor predicts a 40 dB attenuation at a numerical distance of 50.

[2]Also written as $A_s \equiv 1 - j\sqrt{\pi\Omega}e^{-\Omega}\mathrm{erfc}(j\sqrt{\Omega})$ with $\Omega = pe^{-jb}$ — see *Collin* 1985.

14 Appendix VII: Space wave and surface wave links

The z component of the radiation field of a $\hat{z}$ polarized Hertzian dipole located at $z = z_0$ will be, at a location $\mathbf{r} = (x, y, z)$,

$$E_z(x, y, z) = -\eta_o j I_o k \Delta z \sin^2\theta_0 \frac{e^{-jkR_0}}{4\pi R_0},$$

where $R_0 = \sqrt{x^2 + y^2 + (z - z_0)^2}$ and $\cos\theta_0 = (z - z_0)/R_0$ — see Figure 14.1.

Using Weyl's identity and physics notation, we can express this as

$$E_z = \eta_o i I_o \frac{k\Delta z}{4\pi} \sin^2\theta_0 \frac{e^{ikR_0}}{R_0} = \eta_o i I_o \frac{k\Delta z}{4\pi} \sin^2\theta_0 \frac{ik}{2\pi} \int_0^{2\pi} d\phi \int_0^{\pi/2-i\infty} d\theta \, \sin\theta \, e^{i(k_x x + k_y y + k_z|z-z_0|)}.$$

The open circuit voltage of a $\hat{z}$ polarized Hertzian dipole antenna at $\mathbf{r} = (x, y, z)$ will simply be $E_z \Delta z$ if there are no reflecting surfaces near these dipoles.

However, if a reflecting surface were located on $z = 0$ plane, where downgoing plane wave components of E_z shown explicitly under the $\phi\theta$-integral above would reflect at an angle θ, with a reflection coefficient $\Gamma_v(\theta)$, we would have E_z altered as a "Sommerfeld integral"

$$E_z = \eta_o i I_o \frac{k\Delta z}{4\pi} \sin^2\theta_0 \frac{ik}{2\pi} \int_0^{2\pi} d\phi \int_0^{\pi/2-i\infty} d\theta \, \sin\theta \, \{e^{i(k_x x + k_y y + k_z|z-z_0|)} + \Gamma_v(\theta) e^{i(k_x x + k_y y + k_z|z+z_0|)}\},$$

where the added term of reflected waves with relative amplitudes $\Gamma_v(\theta)$ synthesizes a spherical wave that appears to be spreading out from the image point $z = -z_0$ of the dipole located at $z = z_0$. Hence we can re arrange E_z as

$$E_z = \eta_o i I_o \frac{k\Delta z}{4\pi} \sin^2\theta_0 \{\frac{e^{ikR_0}}{R_0} + \frac{\sin^2\theta_1}{\sin^2\theta_0}\Gamma_v(\theta_1)\frac{e^{ikR_1}}{R_1} + \frac{ik}{2\pi} \int_0^{2\pi} d\phi \int_0^{\pi/2-i\infty} d\theta \, \sin\theta \, (\Gamma_v(\theta) - \frac{\sin^2\theta_1}{\sin^2\theta_0}\Gamma_v(\theta_1)) e^{i(k_x x + k_y y + k_z|z+z_0|)}\}$$

where the angles θ_0 and θ_1 are the zenith angle of the observation point seen from the dipole and its image location, respectively, and $R_1 = \sqrt{x^2 + y^2 + (z + z_0)^2}$ is the distance from image location. The first two terms above, namely,

$$E_{sw} \equiv \eta_o i I_o \frac{k\Delta z}{4\pi} \{\sin^2\theta_0 \frac{e^{ikR_0}}{R_0} + \sin^2\theta_1 \Gamma_v(\theta_1) \frac{e^{ikR_1}}{R_1}\}$$

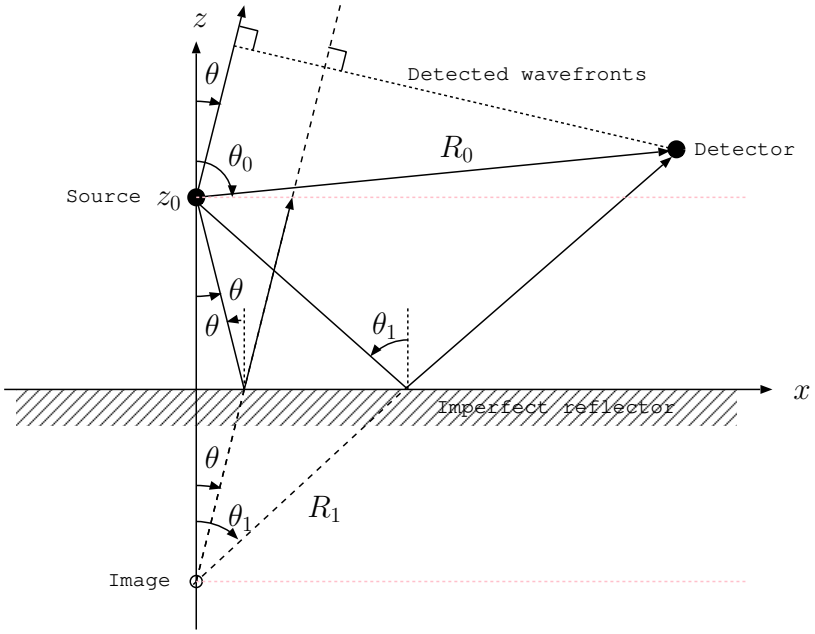


Figure 14.1: A $\hat{z}$-polarized (vertically) dipole source at $(x, z) = (0, z_0)$ and its image at $(x, z) = (0, -z_0)$ given a homogeneous reflector in $z < 0$ half-space. A reflection coefficient $\Gamma_v(\theta)$ relates the $\hat{z}$-components of the electric fields of the incident and reflected plane waves (uniform and non-uniform) from the $z = 0$ surface. Space waves reaching from the source and its image to a detection point at (x, z), $z \geq 0$, will cover distances R_0 and R_1, respectively. The source and its image see the detection point at zenith angles θ_0 and θ_1, respectively. For $R_0 \gg z, z_0$, $\theta_0 \approx \theta_1 \approx 90^\circ$. Note that angle θ concerns the plane waves included in Weyl's identity covering the 0 to $\pi/2$ and $\pi/2$ to $\pi/2 - i\infty$ range — for upgoing and downcoming plane waves included in Weyl's identity (for $z > z_o$ and $z < z_o$, respectively) θ is defined separately as indicated in this diagram.

justifies the space-wave link model discussed in Chapter 2, whereas the third term is a "ground wave" term

$$E_{gw} \equiv \eta_o i I_o \frac{k\Delta z}{4\pi} \sin^2\theta_0 \{\frac{ik}{2\pi}\int_0^{2\pi} d\phi \int_0^{\pi/2-i\infty} d\theta\, \sin\theta\, (\Gamma_v(\theta) - \frac{\sin^2\theta_1}{\sin^2\theta_0}\Gamma_v(\theta_1)) e^{i(k_x x + k_y y + k_z|z+z_0|)}\}$$

that was "missed" in the simplistic derivation of the space wave model earlier on. It can be shown that the ground wave term is negligible away from the $z = 0$ surface, but for $R_1 \gg z_0, z$ and as $\theta_1 \to 90^\circ$ we have $\Gamma(\theta_1) \to -1$ (for both vertical and horizontal polarizations) in which case E_{sw} vanishes and only E_{gw} remains in the form

$$E_{gw} = \eta_o i I_o \frac{k\Delta z}{4\pi} \{\frac{ik}{2\pi}\int_0^{2\pi} d\phi \int_0^{\pi/2-i\infty} d\theta\, \sin\theta\, (\Gamma_v(\theta) + \frac{\sin^2\theta_1}{\sin^2\theta_0}) e^{i(k_x x + k_y y + k_z|z+z_0|)}\}$$

for vertical polarized link. This can be even further simplified as

$$E_{gw} = \eta_o i I_o \frac{k\Delta z}{4\pi}\{\frac{ik}{2\pi}\int_0^{2\pi} d\phi \int_0^{\pi/2-i\infty} d\theta\, \sin\theta\,(\Gamma_v(\theta)+1)e^{ik_x x}\}$$

by taking $\theta_1 = \theta_0 = \pi/2$, $z = z_0 = 0$, and $y = 0$ when the transmitting and receiving dipoles are just above the surface located on the x axis..

Let's examine the integral

$$I \equiv \frac{ik}{2\pi}\int_0^{2\pi} d\phi \int_0^{\pi/2-i\infty} d\theta\, \sin\theta\,(\Gamma_v(\theta)+1)e^{ik_x x}$$

in some detail taking into account that (see R_v in Chapter 2) the reflection coefficient

$$\Gamma_v(\theta) = \frac{n\cos\theta - 1}{n\cos\theta + 1} \;\Rightarrow\; \Gamma_v(\theta)+1 = \frac{n\cos\theta-1}{n\cos\theta+1} + 1 = 2 - \frac{2}{n\cos\theta+1}.$$

Consequently we can write

$$I = 2\frac{e^{ikx}}{x} - \frac{ik}{\pi}\int_0^{2\pi} d\phi \int_0^{\pi/2-i\infty} d\theta\, \frac{\sin\theta}{n\cos\theta+1}\, e^{ik_x x},$$

which, using $k_x = k\sin\theta\cos\phi$, yields

$$\begin{aligned} I - 2\frac{e^{ikx}}{x} &= -\frac{ik}{\pi}\int_0^{\pi/2-i\infty} d\theta\, \frac{\sin\theta}{n\cos\theta+1}\int_0^{2\pi} d\phi\, e^{ikx\sin\theta\cos\phi} \\ &= -\frac{ik}{\pi}\int_0^{\pi/2-i\infty} d\theta\, \frac{\sin\theta}{n\cos\theta+1} 2\pi J_0(kx\sin\theta) \end{aligned}$$

making use of Bessel function of the first kind of order n

$$J_n(z) \equiv \frac{1}{2\pi i^n}\int_0^{2\pi} d\phi e^{iz\cos\phi} e^{in\phi}$$

with $n = 0$. Furthermore, using $J_0(u) = \frac{1}{2}(H_0^{(1)}(u) + H_0^{(2)}(u))$, where $H_0^{(2)}(-u) = -H_0^{(1)}(u)$ are Hankel functions, with $H_0^{(1)}(u) \approx \sqrt{\frac{2}{\pi u}} e^{i(u-\frac{\pi}{4})}$ for large $|u|$, we obtain

$$\begin{aligned} I - 2\frac{e^{ikx}}{x} &= -ik\int_0^{\pi/2-i\infty} d\theta\, \frac{\sin\theta}{n\cos\theta+1}(H_0^{(1)}(kx\sin\theta) + H_0^{(2)}(kx\sin\theta)) \\ &= -ik\int_0^{\pi/2-i\infty} d\theta\, \frac{\sin\theta}{n\cos\theta+1}(H_0^{(1)}(kx\sin\theta) + H_0^{(1)}(-kx\sin\theta)) \\ &= -ik\int_{-\pi/2+i\infty}^{\pi/2-i\infty} d\theta\, \frac{\sin\theta}{n\cos\theta+1} H_0^{(1)}(kx\sin\theta) \\ &\approx -ik\int_{-\pi/2+i\infty}^{\pi/2-i\infty} d\theta\, \frac{\sin\theta}{n\cos\theta+1}\sqrt{\frac{2}{\pi kx\sin\theta}} e^{i(kx\sin\theta-\frac{\pi}{4})} \\ &= -ik\int_{-\pi/2+i\infty}^{\pi/2-i\infty} d\theta\, \frac{e^{-i\frac{\pi}{4}}}{n\cos\theta+1}\sqrt{\frac{2\sin\theta}{\pi kx}}\, e^{ikx\sin\theta}. \end{aligned}$$

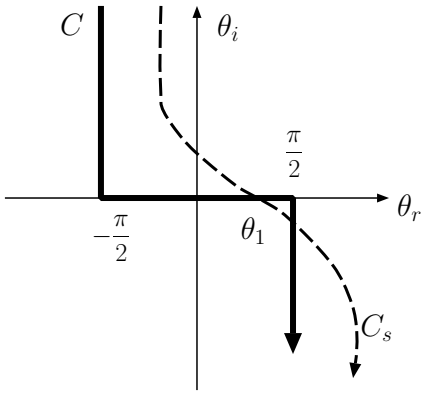


Figure 14.2: Original path of integration C on complex θ-plane and modified integration path C_s used in evaluating the integral I when using the *method of steepest descent*. Notice that the alternate path C_s (permitted for analytic integrand so long as a pole is not crossed) is taken to go through θ_1 with a $-45°$ incline described by $\theta = \theta_1 + (1-i)s$ as the real valued parameter s varies from negative to positive values across $s = 0$. In our implementation $\theta_1 = \pi/2$.

When kx is large the only significant contributions to this path integral will come from complex θ very near $\theta = \pi/2$ if the integration path is distorted as shown in Figure 14.2 to pass through $\theta = \pi/2$ at a 45° angle — this becomes evident when we notice that $\sin\theta = \cos(\theta - \pi/2)$ and therefore

$$I = 2\frac{e^{ikx}}{x} - i\int_{-\pi/2+i\infty}^{\pi/2-i\infty} d\theta\, \frac{e^{-i\frac{\pi}{4}}}{n\cos\theta + 1}\sqrt{\frac{2k\sin\theta}{\pi x}}\, e^{ikx\cos(\theta-\pi/2)},$$

implying that with large kx the exponential $e^{ikx\cos(\theta-\pi/2)}$ in the integrand will vary rapidly as θ departs from $\pi/2$, and contributions from nearby portions of the integration path will be self cancelling except for around $\theta = \pi/2$. Using the modified contour shown in Figure 14.2 to evaluate this integral is known as the *method of steepest descent,* which is in general allowed when the integrand is analytic over the region between the two paths. Since

$$\cos(\theta - \pi/2) \approx 1 - \frac{1}{2}(\theta - \pi/2)^2$$

for small $\theta - \pi/2$, we would have, on the steepest descent path shown in Figure 14.2, that is for the path described by

$$\theta - \frac{\pi}{2} = (1-i)s = \sqrt{2}e^{-i\frac{\pi}{4}}s,$$

where s is a real variable,

$$e^{ikx\cos(\theta-\pi/2)} \approx e^{ikx}e^{-kxs^2},$$

implying that the magnitude e^{-kxs^2} of $e^{ikx\cos(\theta-\pi/2)}$ would first rise to unity and then

descend rapidly[1] as s varies from negative to positive values across $s = 0$. The steepest descent integration path above amounts to following the real s axis of a complex variable s, after a change of variables in integral I affected by $\theta = \frac{\pi}{2} + \sqrt{2}e^{-i\frac{\pi}{4}}s$ and $d\theta = \sqrt{2}e^{-i\frac{\pi}{4}}ds$, leading to

$$I = 2\frac{e^{ikx}}{x} - 2i\int_{-\infty}^{\infty} ds \frac{e^{-i\frac{\pi}{2}}}{n\cos\theta + 1}\sqrt{\frac{k\sin\theta}{\pi x}}\, e^{ikx}e^{-kxs^2}$$
$$= 2\frac{e^{ikx}}{x} - 2\frac{e^{ikx}}{x}\int_{-\infty}^{\infty} ds \frac{1}{n\cos\theta + 1}\sqrt{\frac{kx\sin\theta}{\pi}}\, e^{-kxs^2}.$$

Since only the $s \approx 0$ neighborhood contributes to this integral, we have $\theta \approx \frac{\pi}{2}$ and we can replace $\sin\theta$ by unity, obtaining

$$I = 2\frac{e^{ikx}}{x}\{1 - \sqrt{\frac{kx}{\pi}}\int_{-\infty}^{\infty} ds \frac{e^{-kxs^2}}{n\cos\theta + 1}\}.$$

In the denominator

$$n\cos\theta + 1 = n\cos(\pi/2 + \sqrt{2}e^{-i\frac{\pi}{4}}s) + 1 = -n\sin(\sqrt{2}e^{-i\frac{\pi}{4}}s) + 1$$
$$\approx 1 - n\sqrt{2}e^{-i\frac{\pi}{4}}s = n\sqrt{2}e^{-i\frac{\pi}{4}}(\frac{1}{n\sqrt{2}e^{-i\frac{\pi}{4}}} - s)$$

and therefore

$$I = 2\frac{e^{ikx}}{x}\{1 + \sqrt{\frac{kx}{2\pi}}\frac{e^{i\frac{\pi}{4}}}{n}\int_{-\infty}^{\infty} ds \frac{e^{-kxs^2}}{s - s_p}\}$$

where

$$s_p \equiv \frac{e^{i\frac{\pi}{4}}}{n\sqrt{2}}$$

is the pole of the integrand above.

The remaining integral above is tabulated as

$$\int_{-\infty}^{\infty} ds \frac{e^{-\xi s^2}}{s - s_p} = \pm i\pi e^{-\xi s_p^2}\mathrm{erfc}(\mp i s_p\sqrt{\xi})$$

(e.g., Jones, 1972, page 309) with the upper (lower) sign when $\mathrm{Im}\{s_p\}>0$ (<0), in terms of *complementary error function*

$$\mathrm{erfc}(u) \equiv \frac{2}{\sqrt{\pi}}\int_u^{\infty} e^{-t^2}dt$$

(available in Matlab, Python, etc.).

In our case $\xi = kx$ and $\mathrm{Im}\{s_p\}>0$ and therefore (using the upper sign)

$$I = 2\frac{e^{ikx}}{x}\{1 + \sqrt{\frac{\pi kx}{2}}\frac{ie^{i\frac{\pi}{4}}}{n}e^{-kxs_p^2}\mathrm{erfc}(-is_p\sqrt{kx})\}.$$

[1]For more detailed descriptions of *method of steepest descent* see Berekhovskikh, Chapter 4; Jones (1972)

To confirm this, and to acquire the standard notation used in the literature for the Norton surface or ground wave, we note that

$$n = \sqrt{\epsilon_r + \frac{\sigma}{j\omega\epsilon_o}} = \sqrt{\epsilon_r + i\frac{\sigma}{\omega\epsilon_o}} \;\Rightarrow\; -in^2 = \frac{\sigma}{\omega\epsilon_o} - i\epsilon_r \approx \frac{\sigma}{\omega\epsilon_o}e^{-ib},$$

with $b \equiv \tan^{-1}\frac{\omega\epsilon_o\epsilon_r}{\sigma}$, when σ is large. Hence $n^2 \approx i\frac{\sigma}{\omega\epsilon_o}e^{-ib} \equiv i|n|^2e^{-ib}$ and taking $n = e^{i\frac{\pi}{4}}|n|e^{-ib/2}$

$$s_p = \frac{e^{i\frac{\pi}{4}}}{n\sqrt{2}} = \frac{e^{ib/2}}{\sqrt{2}|n|}$$

with a positive imaginary part due to b. Furthermore

$$kxs_p^2 = \frac{kx}{2|n|^2}e^{ib} \equiv pe^{ib},$$

where

$$p = kx/2|n|^2$$

is said to be the *numerical distance*. Hence,

$$\begin{aligned} I &= 2\frac{e^{ikx}}{x}\{1 + \sqrt{\frac{\pi kx}{2}}\frac{ie^{i\frac{\pi}{4}}}{n}e^{-kxs_p^2}\text{erfc}(-is_p\sqrt{kx})\} \\ &= 2\frac{e^{ikx}}{x}\{1 + \sqrt{\frac{\pi kx}{2}}\frac{ie^{i\frac{\pi}{4}}}{e^{i\frac{\pi}{4}}|n|e^{-ib/2}}e^{-pe^{ib}}\text{erfc}(-i\sqrt{p}e^{ib/2})\} \\ &= 2\frac{e^{ikx}}{x}\{1 + i\sqrt{\pi p}e^{ib/2}e^{-pe^{ib}}\text{erfc}(-i\sqrt{p}e^{ib/2})\}. \end{aligned}$$

This brings us to our final result, namely,

$$E_{gw} \approx 2E_o(x)A_s,$$

where $E_o(x)$ is the broadside field of the dipole at distance x in vacuum, $2E_o(x)$ the same field when the dipole is located and the dipole field is detected just above a PEC reflector, and

$$A_s \equiv 1 - j\sqrt{\pi p}e^{-jb/2}e^{-pe^{-jb}}\text{erfc}(j\sqrt{p}e^{-jb/2})$$

is the *ground-wave attenuation factor*[2] when PEC is replaced with a conductor of refractive index n used to compute p and b.

This factor predicts a 40 dB attenuation at a numerical distance of 50.

[2]Also written as $A_s \equiv 1 - j\sqrt{\pi\Omega}e^{-\Omega}\text{erfc}(j\sqrt{\Omega})$ with $\Omega = pe^{-jb}$ — see *Collin* 1985.

15 Appendix VIII: Aperture antennas and diffraction

Aperture antennas and Huygens-Fresnel Principle

Assume that a wave field $\mathbf{E}(x, y, z)$ due to sources in $z < 0$ half-space is known (measured or calculated) on $z = 0$ plane (see Figure 15.1). In this section we start discussing how $\mathbf{E}(x, y, 0)$ or its scalar components can be propagated from $z = 0$ plane — to be called *aperture plane* — into a homogeneous and isotropic half-space $z > 0$. The procedures to be described, starting with Huygens-Fresnel Principle, are essential for the description of aperture type antennas — slits, horns, reflector type dish antennas, and so on.

Huygens-Fresnel principle predates Maxwell's equations — Huygens observed in 17th century that wave disturbances can be propagated forward by treating every point on a wavefront (of any type of wave motion) as a source of a spherical wave and superposing all the spherical waves reaching to any observation point from all points in the wavefront. In electromagnetic applications, for instance, an infinitesimal area $dx'dy'$ of a wavefront residing on the $z = 0$ surface centered about position $x = x'$ and $y = y'$ with a wave electric field amplitude $E(x', y')$ will generate and propagate a spherical wave described by

$$\frac{jk}{2\pi}E(x', y')dx'dy'\frac{e^{-jkR}}{R},\ R \equiv \sqrt{(x - x')^2 + (y - y')^2 + z},$$

for a location (x, y, z) in $z > 0$ half-space at a distance R from the source at $(x', y', 0)$. Likewise, if the wavefront on $z = 0$ surface only supports fields $E(x')$ independent of y', then a y-independent wave field in $z > 0$ half-space can be calculated by superposing cylindrical waves specified as

$$\sqrt{\frac{jk}{2\pi}}E(x')dx'\frac{e^{-jkR}}{\sqrt{R}},\ R \equiv \sqrt{(x - x')^2 + z}.$$

These recipes of 2-D and 1-D versions of of Huygens-Fresnel principle can be derived from Maxwell's equations as will be shown later in this Appendix. Their sole use is for the computation of field intensities past an aperture plane while the associated vector directions will need to be deduced from Maxwell's equations independently.

We will illustrate the use of the 2-D recipe by computing the fields past circular and square shaped apertures defined on the $z = 0$ plane illuminated by a plane wave $\mathbf{E} = \mathbf{E}_o e^{-jkz}$ propagating in $z < 0$ half-space onto the $z = 0$ plane. The 1-D recipe will later on be used to describe "knife-edge" diffraction effects.

Circular aperture: A uniform plane wave field $\mathbf{E} = \hat{x}E_o e^{-jkz}$ is incident from left onto a circular aperture C_a of a radius a centered about the origin at $z = 0$ plane. Using Huygens-Fresnel principle, the diffracted field amplitude at a point P on the z-axis can be computed as

$$E_P = \frac{jk}{2\pi}\int_{C_a} dx'dy' E_o \frac{e^{-jkR}}{R}$$

where $R = \sqrt{\rho^2 + z^2}$ and $\rho = \sqrt{x'^2 + y'^2}$. Moving to cylindrical coordinates,

$$E_p = \frac{jk}{2\pi}\int_{\phi=0}^{2\pi} d\phi \int_{\rho=0}^{a} d\rho\, \rho\, E_o \frac{e^{-jk\sqrt{\rho^2+z^2}}}{\sqrt{\rho^2+z^2}} = jkE_o \int_{\rho=0}^{a} d\rho\, \rho \frac{e^{-jk\sqrt{\rho^2+z^2}}}{\sqrt{\rho^2+z^2}}.$$

Since $R^2 = \rho^2 + z^2$ implies $RdR = \rho d\rho$, we get

$$\begin{aligned} E_p &= jkE_o \int_{\rho=0}^{a} d\rho\, \rho \frac{e^{-jk\sqrt{\rho^2+z^2}}}{\sqrt{\rho^2+z^2}} = jkE_o \int_{R=z}^{\sqrt{a^2+z^2}} dR e^{-jkR} = jkE_o \frac{e^{-jkR}}{-jk}\Big|_z^{\sqrt{a^2+z^2}} \\ &= E_o\{e^{-jkz} - e^{-jk\sqrt{a^2+z^2}}\}. \end{aligned}$$

Field E_p calculated above has some interesting properties:

- With a purely real k, and a fixed a value, the field E_p will fluctuate between $2E_o e^{-jkz}$ and 0, depending on z, with an average of $E_o e^{-jkz}$.
- $\lim_{a\to\infty} E_p = E_o e^{-jkz}$ in the ROC $\mathrm{Im}\{k\}<0$ for E_p, that is, the plane wave character of the incident wave $E_o e^{-jkz}$ is recovered in $z > 0$ half-space, consistent with what would be expected for an infinitely large aperture (i.e., no blockage of the incident field on $z = 0$ plane) — this is an indirect verification of the Huygens-Fresnel principle[1].

Also, if a circular absorber (occulting disc) of radius a is blocking an incoming plane wave $E_o e^{-jkz}$ centered on $z = 0$ plane about the origin, the field passed the absorber on the z axis would be $E_o e^{-jkz}$ minus E_p above, by using Babinet's principle, leading to a modified

$$E_p = E_o e^{-jkz} - E_o\{e^{-jkz} - e^{-jk\sqrt{a^2+z^2}}\} = E_o e^{-jk\sqrt{a^2+z^2}}$$

that vanishes in $a \to \infty$ limit (as would be expected if the blocking disc is infinitely big) within the region of convergence $\mathrm{Im}\{k\}<0$ corresponding to lossy propagation, however weak in losses. More surprisingly, perhaps, $E_p \neq 0$ just behind ($z = 0$) the disc, however large a may be, as long as it is finite — this "bright spot" just behind the disc can be verified experimentally, it is known as "Arago spot", which can be considered an experimental verification of both the Huygens-Fresnel principle and Babinet's principle.

[1] As well as an excellent example of the validity of the analytic continuation of the result to regions outside the region of convergence (ROC), $\mathrm{Im}\{k\}<0$:

Notice how $E_p = E_o e^{-jkz}$ solution for $a \to \infty$ case (modeling the absence of any screen on $z = 0$ plane) happens to be the correct deduction for $z > 0$ half space whether $E_o e^{-jkz}$ wave is damped ($\mathrm{Im}\{k\}<0$), undamped ($\mathrm{Im}\{k\}=0$), or exponentially growing ($\mathrm{Im}\{k\}>0$), even though it can only be derived from Huygen's principle (as shown in this example) within the ROC $\mathrm{Im}\{k\}<0$.

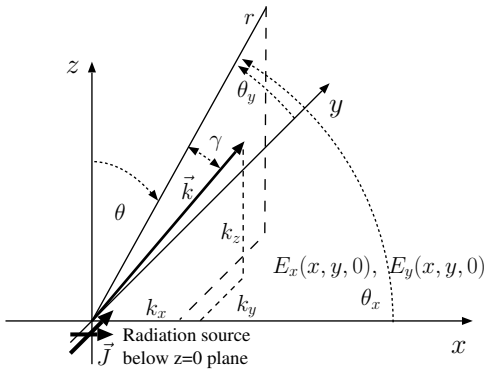


Figure 15.1: A radiation source below $z = 0$ plane generates the $z > 0$ field $\mathbf{E}(\mathbf{r})$ discussed in the text. As shown, $\mathbf{E}(\mathbf{r})$, $z > 0$, can be fully determined by using $E_{x,y}(x, y, 0)$.

Square aperture: A uniform plane wave field $\mathbf{E} = \hat{x}E_o e^{-jkz}$ is incident from left onto a square shaped aperture centered about the origin at $z = 0$ plane. Using Huygens-Fresnel principle, the diffracted field in $z > 0$ half space is

$$E(x, y, z) = \frac{jk}{2\pi} \int_{-L/2}^{L/2} dx' \int_{-L/2}^{L/2} dy' E_o \frac{e^{-jkR}}{R}$$

where $R \equiv \sqrt{(x - x')^2 + (y - y')^2 + z}$. For $R \gg L$, $R \approx r - (x', y', 0) \cdot \mathbf{r}/r$ and

$$E(x, y, z) \approx \frac{jk}{2\pi} E_o \frac{e^{-jkr}}{r} \int_{-L/2}^{L/2} dx' \int_{-L/2}^{L/2} dy' e^{j(x'k\cos\theta_x + y'k\cos\theta_y)}$$

where $\cos\theta_x = \sin\theta\cos\phi$ and $\cos\theta_y = \sin\theta\sin\phi$. This leads to a "far field" radiation solution

$$\begin{aligned} E(x, y, z) &\approx \frac{jk}{2\pi} E_o \frac{e^{-jkr}}{r} \int_{-L/2}^{L/2} dx' e^{jx'k\cos\theta_x} \int_{-L/2}^{L/2} dy' e^{y'k\cos\theta_y} \\ &= \frac{jk}{2\pi} E_o L^2 \mathrm{sinc}(k\cos\theta_x \frac{L}{2}) \mathrm{sinc}(k\cos\theta_y \frac{L}{2}) \frac{e^{-jkr}}{r}. \end{aligned}$$

This is a spherical wave in $z > 0$ half-space having a beam pattern similar to a uniformly illuminated and filled 2D antenna array with a HPBW$=\frac{\lambda}{L}$, beam solid angle $\Omega = \frac{\lambda^2}{L^2}$ and directivity $D = \frac{4\pi}{\Omega} = \frac{4\pi}{\lambda^2} A_{phy}$, where $A_{phy} = L^2$ is the physical area of the square shaped aperture.

This last example brought home the idea of an aperture antenna.

Fourier optics derivation of Huygens-Fresnel Principle

Let

$$F_{x,y}(k_x, k_y) \equiv \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy E_{x,y}(x, y, 0) e^{j(k_x x + k_y y)}$$

denote 2-D spatial Fourier transforms of tangential components $E_{x,y}(x,y,0)$ of $\mathbf{E}(x,y,z)$ at $z=0$. Then via inverse Fourier transform,

$$E_{x,y}(x,y,0) = \int_{-\infty}^{\infty} \frac{dk_x}{2\pi} \int_{-\infty}^{\infty} \frac{dk_y}{2\pi} F_{x,y}(k_x,k_y) e^{-j(k_x x + k_y y)},$$

and, we suggest, by inspection,

$$\mathbf{E}(\mathbf{r}) = \int_{-\infty}^{\infty} \frac{dk_x}{2\pi} \int_{-\infty}^{\infty} \frac{dk_y}{2\pi} [\hat{x}F_x(k_x,k_y) + \hat{y}F_y(k_x,k_y) + \hat{z}F_z(k_x,k_y)] e^{-j(k_x x + k_y y + k_z z)}$$

for $z > 0$, with

$$k_z \equiv \sqrt{k^2 - k_x^2 - k_y^2},$$

and

$$F_z(k_x,k_y) = -\frac{k_x}{k_z} F_x(k_x,k_y) - \frac{k_y}{k_z} F_y(k_x,k_y).$$

Our suggestion turns out to be the unique solution of $z > 0$ field since it is a linear superposition of TEM plane-wave solutions of Maxwell's equations and also matches the required boundary conditions on the aperture plane. The solution can also be written more compactly as

$$\mathbf{E}(\mathbf{r}) = \int_{-\infty}^{\infty} \frac{dk_x}{2\pi} \int_{-\infty}^{\infty} \frac{dk_y}{2\pi} \mathbf{F}(k_x,k_y) e^{-j(k_x x + k_y y + k_z z)},$$

where

$$\mathbf{F}(k_x,k_y) \equiv \hat{x}F_x(k_x,k_y) + \hat{y}F_y(k_x,k_y) - \hat{z}[\frac{k_x}{k_z} F_x(k_x,k_y) + \frac{k_y}{k_z} F_y(k_x,k_y)]$$

is said to be the *plane wave spectrum* of $z > 0$ field.

The expression above represents the $z > 0$ field as a sum of a spectrum of plane waves propagating out of $z = 0$ plane. The same field can also be expressed as a sum of spherical waves radiating out of the same plane in accordance with *Huygens's principle*. To obtain this second representation of $\mathbf{E}(\mathbf{r})$ we first consider the special case with $E_y(x,y,0) = 0$ so that

$$\mathbf{F}(k_x,k_y) \equiv F_x(k_x,k_y)(\hat{x} - \hat{z}\frac{k_x}{k_z}),$$

and

$$\mathbf{E}(\mathbf{r}) = \int_{-\infty}^{\infty} \frac{dk_x}{2\pi} \int_{-\infty}^{\infty} \frac{dk_y}{2\pi} F_x(k_x,k_y)(\hat{x} - \hat{z}\frac{k_x}{k_z}) e^{-j(k_x x + k_y y + k_z z)}.$$

Substituting

$$F_x(k_x,k_y) \equiv \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' E_x(x',y',0) e^{j(k_x x' + k_y y')}$$

into $\mathbf{E}(\mathbf{r})$ we obtain

$$\mathbf{E}(\mathbf{r}) = \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' E_x(x',y',0) \int_{-\infty}^{\infty} \frac{dk_x}{2\pi} \int_{-\infty}^{\infty} \frac{dk_y}{2\pi} (\hat{x} - \hat{z}\frac{k_x}{k_z}) e^{-j(k_x(x-x') + k_y(y-y') + k_z z)},$$

or, equivalently,

$$\mathbf{E}(\mathbf{r}) = \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' E_x(x', y', 0)\mathbf{g}(x - x', y - y', z),$$

where

$$\mathbf{g}(x, y, z) \equiv \int_{-\infty}^{\infty} \frac{dk_x}{2\pi} \int_{-\infty}^{\infty} \frac{dk_y}{2\pi} (1, 0, -\frac{k_x}{k_z}) e^{-j(k_x x + k_y y + k_z z)}.$$

Clearly $\mathbf{g}(x, y, z)$ is the $z > 0$ electric field corresponding to $E_x(x, y, 0) = \delta(x)\delta(y)$, and it can be simplified as

$$\begin{aligned} \mathbf{g}(x, y, z) &= j(\frac{\partial}{\partial z}, 0, -\frac{\partial}{\partial x}) \underbrace{\int_{-\infty}^{\infty} \frac{dk_x}{2\pi} \int_{-\infty}^{\infty} \frac{dk_y}{2\pi} \frac{e^{-j(k_x x + k_y y + k_z z)}}{k_z}}_{= \frac{je^{-jkr}}{2\pi r} \quad \text{"Weyl identity"}} \\ &\rightarrow -(-jk\frac{\partial r}{\partial z}, 0, jk\frac{\partial r}{\partial x})\frac{e^{-jkr}}{2\pi r} = (\frac{z}{r}, 0, -\frac{x}{r})\frac{jk}{2\pi}\frac{e^{-jkr}}{r} \end{aligned}$$

where the last line is valid in $r \gg \lambda$ limit where the neglect of partial derivatives due to r in the denominator can be justified. Notice the important role played by Weyl identity that we learned earlier in the class and also notice that $\mathbf{g}(x, y, z) \propto \frac{jk}{2\pi}\frac{e^{-jkr}}{r}$ is effectively a Maxwell's equations based derivation of the 2-D Huygens-Fresnel source discussed in the previous section.

Now we write, for $r \gg \lambda$,

$$\mathbf{E}(\mathbf{r}) = \frac{jk}{2\pi} \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' E_x(x', y', 0)(\frac{z}{R}, 0, -\frac{x - x'}{R})\frac{e^{-jkR}}{R},$$

where

$$R \equiv \sqrt{(x - x')^2 + (y - y')^2 + z^2}.$$

Note that while the *plane-wave synthesis* of $\mathbf{E}(\mathbf{r})$ makes use of the Fourier transform of $E_x(x, y, 0)$, the *spherical-wave synthesis* just obtained uses $E_x(x, y, 0)$ directly.

Far field approximation for aperture antenna

In the far-field of aperture antennas — e.g., parabolic dishes — with finite dimensions L the variables x' and y' in $\mathbf{E}(\mathbf{r})$ above can be neglected everywhere except for in

$$R = \sqrt{(x - x')^2 + (y - y')^2 + z^2} \approx r - \hat{r} \cdot (x'\hat{x} + y'\hat{y})$$

in the exponent (paraxial approximation) to obtain a far field *approximation*

$$\begin{aligned} \mathbf{E}(\mathbf{r}) &= \frac{jk}{2\pi}(\frac{z}{r}, 0, -\frac{x}{r})\frac{e^{-jkr}}{r} \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' E_x(x', y', 0) e^{jk\hat{r}\cdot(\hat{x}x' + \hat{y}y')} \\ &= \frac{jk}{2\pi} \cos\theta \frac{e^{-jkr}}{r}(1, 0, -\frac{\cos\theta_x}{\cos\theta}) F_x(k\cos\theta_x, k\cos\theta_y), \end{aligned}$$

which is a single spherical wave described in terms of Fourier transform $F_x(k_x, k_y)$.

Likewise, in the far field $r > \frac{2L^2}{\lambda}$ of a y-polarized aperture antenna

$$\mathbf{E}(\mathbf{r}) = \frac{jk}{2\pi}\cos\theta\frac{e^{-jkr}}{r}(0, 1, -\frac{\cos\theta_y}{\cos\theta})F_y(k\cos\theta_x, k\cos\theta_y).$$

These results indicate that the far-field radiation pattern of aperture antenna is determined by the shape of the plane-wave spectrum $\mathbf{F}(k_x, k_y)$.

Fresnel diffraction formula

Next we note that for large z, i.e., for $z \gg \lambda$,

$$e^{-jkR} = e^{-jk\sqrt{(x-x')^2+(y-y')^2+z^2}} = e^{-jkz\sqrt{1+(\frac{x-x'}{z})^2+(\frac{y-y'}{z})^2}}$$

in

$$\mathbf{E}(\mathbf{r}) = \frac{jk}{2\pi}\int_{-\infty}^{\infty}dx'\int_{-\infty}^{\infty}dy'E_x(x', y', 0)(\frac{z}{R}, 0, -\frac{x-x'}{R})\frac{e^{-jkR}}{R}$$

is a highly sensitive function of

$$\frac{x-x'}{z} \quad \text{and} \quad \frac{y-y'}{z}$$

except when

$$|\frac{x-x'}{z}| \ll 1 \quad \text{and} \quad |\frac{y-y'}{z}| \ll 1.$$

Thus for large z the integrand above destructively interferes (i.e., cancels out) with itself everywhere across all of x' and y' except for in the neighborhood of $x' = x$ and $y' = y$ where

$$e^{-jkR} \approx e^{-jkz(1+\frac{(x-x')^2+(y-y')^2}{2z^2})}.$$

For $z \gg \lambda$ we can therefore write

$$E_x(\mathbf{r}) \approx \frac{jk}{2\pi}\frac{e^{-jkz}}{z}\int_{-\infty}^{\infty}dx'\int_{-\infty}^{\infty}dy'E_x(x', y', 0)e^{-\frac{jk}{2z}((x-x')^2+(y-y')^2)}$$

by replacing all R by z — i.e., set $x' = x$ and $y' = y$ within R — except in the phase term where the expansion above is utilized.

This scalar result for $z \gg \lambda$ is known as *Fresnel diffraction formula.* Note that its validity does not depend on z being large compared to an aperture size L. In fact the result is applicable even when $L \to \infty$.

Furthermore, an identical equation is applicable for $E_y(\mathbf{r})$, implying that Fresnel formula can be used with the *scalar* tangential component of the aperture plane field regardless of its polarization. The formula of the scalar component $E(x, y, z)$ can be written in convolution form as

$$E(x, y, z) = \frac{jk}{2\pi}\frac{e^{-jkz}}{z}E(x, y, 0) * *e^{-\frac{jk}{2z}(x^2+y^2)}$$

Aperture-plane fields	$E_{x,y}(x,y,0) \leftrightarrow F_{x,y}(k_x,k_y) \equiv \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy E_{x,y}(x,y,0) e^{j(k_x x + k_y y)}$
Plane-wave spectrum	$\mathbf{F}(k_x,k_y) \equiv \hat{x} F_x(k_x,k_y) + \hat{y} F_y(k_x,k_y) - \hat{z}[\frac{k_x}{k_z} F_x(k_x,k_y) + \frac{k_y}{k_z} F_y(k_x,k_y)]$
Far-field formula (x-pol)	$\mathbf{E}(\mathbf{r}) = \frac{jk}{2\pi} \cos\theta \frac{e^{-jkr}}{r} (1, 0, -\frac{\cos\theta_x}{\cos\theta}) F_x(k\cos\theta_x, k\cos\theta_y)$ for $r > \frac{2L^2}{\lambda}$
Far-field formula y-pol)	$\mathbf{E}(\mathbf{r}) = \frac{jk}{2\pi} \cos\theta \frac{e^{-jkr}}{r} (0, 1, -\frac{\cos\theta_x}{\cos\theta}) F_y(k\cos\theta_x, k\cos\theta_y)$ for $r > \frac{2L^2}{\lambda}$
Fresnel diff. formula (2D)	$E(x,y,z) = \frac{jk}{2\pi} \frac{e^{-jkz}}{z} E(x,y,0) ** e^{-\frac{jk}{2z}(x^2+y^2)}$ for $z \gg \lambda$
Fresnel diff. formula (1D)	$E(x,y,z) = \sqrt{\frac{jk}{2\pi}} \frac{e^{-jkz}}{\sqrt{z}} E_x(x,0,0) * e^{-\frac{jk}{2z}x^2}$ for $z \gg \lambda$
Weyl identity	$\frac{e^{-jkr}}{r} = \frac{-j}{2\pi} \int_{-\infty}^{\infty} dk_x \int_{-\infty}^{\infty} dk_y \frac{e^{-j(k_x x + k_y y + k_z \lvert z \rvert)}}{k_z}$ where $k_z \equiv \sqrt{k^2 - k_x^2 - k_y^2}$

Table 15.1: Formulas and identities relevant to aperture fields and their propagation into $z > 0$ half-space.

where $**$ implies convolution over x and y.

Now, if the aperture field $E_x(x,y,0) = E_x(x,0,0)$ were independent of y we could get a 1D version of Fresnel diffraction formula by observing that

$$\begin{aligned} E_x(\mathbf{r}) &\approx \frac{jk}{2\pi} \frac{e^{-jkz}}{z} \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' E_x(x',0,0) e^{-\frac{jk}{2z}((x-x')^2 + (y-y')^2)} \\ &= \frac{jk}{2\pi} \frac{e^{-jkz}}{z} \int_{-\infty}^{\infty} dx' E_x(x',0,0) e^{-\frac{jk}{2z}(x-x')^2} \underbrace{\int_{-\infty}^{\infty} dy' e^{-\frac{jk}{2z}(y-y')^2}}_{\sqrt{\frac{2\pi z}{jk}}} \\ &= \sqrt{\frac{jk}{2\pi}} \frac{e^{-jkz}}{\sqrt{z}} \int_{-\infty}^{\infty} dx' E_x(x',0,0) e^{-\frac{jk}{2z}(x-x')^2} = \sqrt{\frac{jk}{2\pi}} \frac{e^{-jkz}}{\sqrt{z}} E_x(x,0,0) * e^{-\frac{jk}{2z}x^2} \end{aligned}$$

after using the identity $\int_{-\infty}^{\infty} dx e^{-jpx^2} = \sqrt{\frac{\pi}{p}} e^{-j\frac{\pi}{4}}$ in performing the y' integration.

To summarize, we have introduced in this section a number of ideas including apertures and aperture fields, plane-wave spectrum, as well as far-field and Fresnel diffraction formulas for calculating fields radiated by finite or infinite sized apertures. The following subsections put these tools (see Table 15.1) into useful applications.

Aperture efficiency

Consider an x-polarized aperture antenna with an aperture area A_{phys} on $z = 0$ plane centered about the origin. We will assume a tangential field

$$E_x(y,y) = f(x,y) e^{j\phi}$$

where function $f(x,y)$ real valued and non-negative while phase ϕ is some constant. For $z > 0$ and in the far field we will have

$$\mathbf{E}(\mathbf{r}) = \frac{jk}{2\pi} \cos\theta \frac{e^{-jkr}}{r} (1, 0, -\frac{\cos\theta_x}{\cos\theta}) F_x(k\cos\theta_x, k\cos\theta_y)$$

where

$$F_x(k\cos\theta_x, k\cos\theta_y) = \int dx \int dy f(x,y) e^{j\phi} e^{j(k\cos\theta_x x + k\cos\theta_y y)}.$$

It follows that field magnitude

$$|\mathbf{E}(\mathbf{r})| = \frac{k}{2\pi r}\sqrt{\cos^2\theta + \cos^2\theta_x}\left| \int dx \int dy f(x,y) e^{j\phi} e^{j(k\cos\theta_x x + k\cos\theta_y y)}\right|$$

and maximizes in the direction $\theta = 0$ and $\theta_x = \pi/2$ with a value

$$|\mathbf{E}(\mathbf{r})|_{max} = \frac{k}{2\pi r}\int dx \int dy f(x,y) = \frac{kA_{phys}}{2\pi r}\langle E(x,y)\rangle$$

such that

$$\langle E(x,y)\rangle \equiv \frac{\int dx \int dy f(x,y)}{A_{phys}}.$$

Now, the maximum Poynting flux of the antenna is

$$S_{max} = \frac{|\mathbf{E}|^2_{max}}{2\eta_o} = \frac{k^2 A^2_{phys}\langle E(x,y)\rangle^2}{8\pi^2 r^2 \eta_o} \equiv \frac{P_t}{4\pi r^2} D$$

in terms of directivity D and transmitted power P_t.

Now, if we model P_t as

$$P_t = \int dx \int dy \frac{|E(x,y)|^2}{2\eta_o} \equiv \frac{A_{phys}\langle |E(x,y)|^2\rangle}{2\eta_o},$$

such that

$$\langle |E(x,y)|^2\rangle \equiv \frac{\int dx \int dy f^2(x,y)}{A_{phys}},$$

we find that

$$S_{max} = \frac{|\mathbf{E}|^2_{max}}{2\eta_o} = \frac{k^2 A^2_{phys}\langle E(x,y)\rangle^2}{8\pi^2 r^2 \eta_o} \equiv \frac{A_{phys}\langle |E(x,y)|^2\rangle}{8\pi r^2 \eta_o} D,$$

leading to

$$D = \frac{k^2}{\pi} A_{phys} \frac{\langle E(x,y)\rangle^2}{\langle |E(x,y)|^2\rangle} = \frac{4\pi}{\lambda^2} A_{phys} \frac{\langle E(x,y)\rangle^2}{\langle |E(x,y)|^2\rangle}.$$

This result is usually written as

$$D = \frac{4\pi}{\lambda^2} A_{eff}$$

where

$$A_{eff} = A_{phys}\epsilon_{ap}, \;\; \epsilon_{ap} \equiv \frac{\langle E(x,y)\rangle^2}{\langle |E(x,y)|^2\rangle}$$

is known as *aperture efficiency*. In general $0 \leq \epsilon_{ap} < 1$, with unity efficiency corresponding to uniform aperture illumination, that is $f(x,y) = 1$ over the entire A_{phys}.

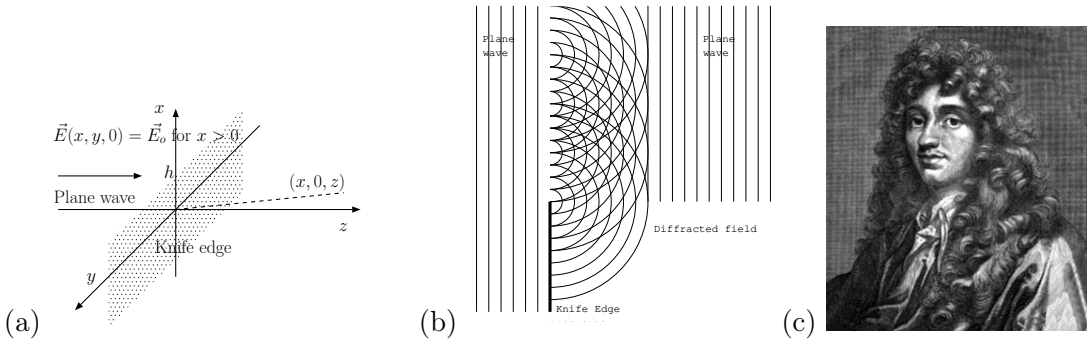


Figure 15.2: (a) Geometry of knife-edge diffraction, the edge extending from $x = -\infty$ to $x = h$, (b) a cartoon showing that superposition of spherical waves from the aperture plane generates a diffracted field in the geometrical optics shadow zone behind the knife edge, and (c) Christiaan Huygens (1629-1695)

Example 1: Consider a rectangular aperture with $2\,\mathrm{m} \times 1\,\mathrm{m}$ dimensions. On the left half of the aperture $E(x, y) = 1$ V/m while it is $E(x, y) = 2$ V/m on the right half.

Consequently

$$\langle E \rangle = \frac{1+2}{2} = 1.5$$

while

$$\langle |E|^2 \rangle = \frac{1+4}{2} = 2.5.$$

Hence

$$\epsilon_{ap} = \frac{\langle E \rangle^2}{\langle |E|^2 \rangle} = \frac{1.5^2}{2.5} = \frac{2.25}{2.5} = 0.9.$$

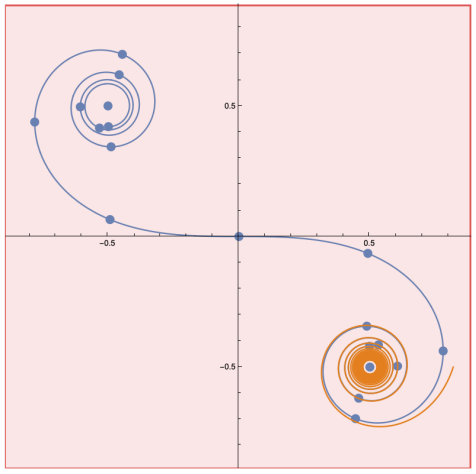
Example 2: Repeat Example 1 if on the left half of the aperture $E(x, y) = 4$ V/m while it is $E(x, y) = 4$ V/m on the right half.

Knife-edge diffraction

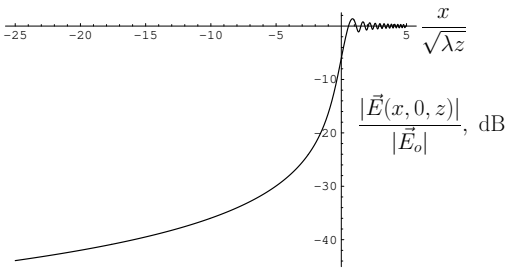
A plane wave incident on a semi-infinite screen — a "knife-edge" — causes an aperture field $\mathbf{E}(x, y, 0) = \hat{x} E_o u(x - h)$ on $z = 0$ plane as shown in Figure 15.2a. For $z \gg \lambda$, 1D Fresnel diffraction formula implies that

$$\begin{aligned} E_x(x, 0, z) &= \sqrt{\frac{jk}{2\pi}} \frac{e^{-jkz}}{\sqrt{z}} \int_{-\infty}^{\infty} dx' E_o u(x' - h) e^{-\frac{jk}{2z}(x-x')^2} \\ &= \sqrt{\frac{jk}{2\pi}} E_o \frac{e^{-jkz}}{\sqrt{z}} \int_{h}^{\infty} dx' e^{-\frac{j\pi}{\lambda z}(x-x')^2}. \end{aligned}$$

```
Fr[x_] := FresnelC[x] - i FresnelS[x]
FrA[x_] := Fr[∞] - 1/(i π x) e^(-i π/2 x^2)
Show[
 ParametricPlot[{FresnelC[x], -FresnelS[x]}, {x, -4, 4}],
 ListPlot[Table[{FresnelC[x], -FresnelS[x]}, {x, -4, 4, .5}]],
 ListPlot[{{-0.5, 0.5}, {0.5, -0.5}}],
 ParametricPlot[{Re[FrA[x]], Im[FrA[x]]}, {x, 1, 12}, PlotStyle → Orange
 PlotRange -> {{-0.8, 0.8}, {-0.8, 0.8}},
 Framed → True, GridLines → True]
```



(a)



(b)

Figure 15.3: (a) Cornu spiral depicting real and imaginary parts of $\mathrm{Fr}(\nu)$ as a function of ν as it varies from $-\infty$ to ∞ where steps 0.5 apart are marked with dots and the orange curve depicts $\mathrm{Fr}(\infty) - \frac{e^{-j\frac{\pi}{2}\nu^2}}{j\pi\nu}$ that approximates $\mathrm{Fr}(\nu)$ accurately for $\nu > 1$. Note that ν is also the arc length along the spiral measured from the origin. (b) Normalized field intensity for $z \gg \lambda$ as a function of normalized distance $\frac{x}{\sqrt{\lambda z}}$ above the knife-edge at $x = 0$.

We continue with a change of variables $p = \sqrt{\frac{2}{\lambda z}}(x' - x)$, so that $dx' = \sqrt{\frac{\lambda z}{2}}dp$, and, therefore

$$E_x(x,0,z) = E_o e^{-jkz}\sqrt{\frac{j}{2}}\int_{\sqrt{\frac{2}{\lambda z}}(h-x)}^{\infty} dp\, e^{-j\frac{\pi}{2}p^2}.$$

Now define the *Fresnel integral function*

$$\mathrm{Fr}(\nu) \equiv \int_0^{\nu} dp\, e^{-j\frac{\pi}{2}p^2}$$

having the properties that $\mathrm{Fr}(0) = 0$, $\mathrm{Fr}(-\nu) =$-$\mathrm{Fr}(\nu)$, and $\mathrm{Fr}(\pm\infty) = \pm\frac{1-j}{2} = \pm\sqrt{\frac{-j}{2}}$. The real and imaginary parts of $\mathrm{Fr}(\nu)$ are coded as the FresnelC[ ] and -FresnelS[ ] functions in Mathematica — see `http://mathworld.wolfram.com/FresnelIntegrals.html` — and a complex plane plot of $\mathrm{Fr}(\nu)$ as a function of real ν from $-\infty$ to ∞ constitutes the *Cornu spiral* shown in Figure 15.3a.

In terms of Fresnel integral function, the knife edge field can be expressed as[2]

$$E_x(x,0,z) = E_o e^{-jkz}\sqrt{\frac{j}{2}}\{\mathrm{Fr}(\infty) - \mathrm{Fr}(\sqrt{\frac{2}{\lambda z}}(h-x))\}$$
$$= E_o e^{-jkz}\{\frac{1}{2} - \sqrt{\frac{j}{2}}\mathrm{Fr}(\sqrt{\frac{2}{\lambda z}}(h-x))\}.$$

Since $\mathrm{Fr}(0) = 0$, we get $E_x(x = h, 0, z) = \frac{1}{2}E_o e^{-jkz}$ at the "shadow edge" from this result. Furthermore the shadow edge is not abrupt and varies as shown in Figure 15.3b, which is a dB plot of $|E_x(x,0,z)|/|E_o|$ as a function of $\frac{x}{\sqrt{\lambda z}}$ for the case of $h = 0$. Note that $|E_x(x,0,z)|/|E_o| \to 0$ as $x \to -\infty$, and $\to 1$ as $x \to \infty$; these limits are consistent with the simple geometric optics viewpoint but the transition near small $\frac{|x|}{\sqrt{\lambda z}}$ show additional diffraction effects. The maximum relative amplitude of ~ 1.2 occurs near $x = \sqrt{\lambda z}$, known as the *Fresnel distance*.

Take $x = 0$ and $h = -\infty$ in the result above — this would correspond to the case of no knife edge at all and yield

$$E_x(x=0,0,z) = E_o e^{-jkz}\{\frac{1}{2} - \sqrt{\frac{j}{2}}\underbrace{\mathrm{Fr}(\sqrt{\frac{2}{\lambda z}}(-\infty - 0))}\} = E_o e^{-jkz}$$
$$-\sqrt{\frac{-j}{2}}$$

as expected.

Let us finally consider two knife edges placed at $x = h_1 < 0$ and $x = h_2 > 0$ such that $x < h_1$ and $x > h_2$ are blocked and an aperture exists only for $h_1 < x < h_2$ on the $z = 0$ aperture plane. Assuming a plane wave $E_o e^{-kz}$ incident from the left, the field in $z > 0$ half space can be obtained by subtracting h_2 dependent knife edge field from the h_1 dependent one (another application of Babinet's principle) as in

$$E_x(x,0,z) = E_o e^{-jkz}\{\frac{1}{2} - \sqrt{\frac{j}{2}}\mathrm{Fr}(\sqrt{\frac{2}{\lambda z}}(h_1 - x))\} - E_o e^{-jkz}\{\frac{1}{2} - \sqrt{\frac{j}{2}}\mathrm{Fr}(\sqrt{\frac{2}{\lambda z}}(h_2 - x))\}$$
$$= E_o e^{-jkz}\frac{\mathrm{Fr}(\sqrt{\frac{2}{\lambda z}}(h_2 - x)) - \mathrm{Fr}(\sqrt{\frac{2}{\lambda z}}(h_1 - x))}{\sqrt{\frac{2}{j}}}$$
$$= E_o e^{-jkz}\frac{\mathrm{Fr}(\sqrt{\frac{2}{\lambda z}}(h_2 - x)) - \mathrm{Fr}(\sqrt{\frac{2}{\lambda z}}(h_1 - x))}{\mathrm{Fr}(\infty) - \mathrm{Fr}(-\infty)}.$$

The normalization used in the last line is convenient because in the field magnitude expression

$$|E_x(x,0,z)| = |E_o|\frac{|\mathrm{Fr}(\sqrt{\frac{2}{\lambda z}}(h_2 - x)) - \mathrm{Fr}(\sqrt{\frac{2}{\lambda z}}(h_1 - x))|}{|\mathrm{Fr}(\infty) - \mathrm{Fr}(-\infty)|}$$

[2] An alternate way of expressing this result is $E_x(x,0,z) = \frac{1}{2}\mathrm{erfc}(-\sqrt{j\pi}\frac{(x-h)}{\sqrt{\lambda z}})E_o e^{-jkz}$ in terms of *complementary error function* that was introduced earlier in the notes.

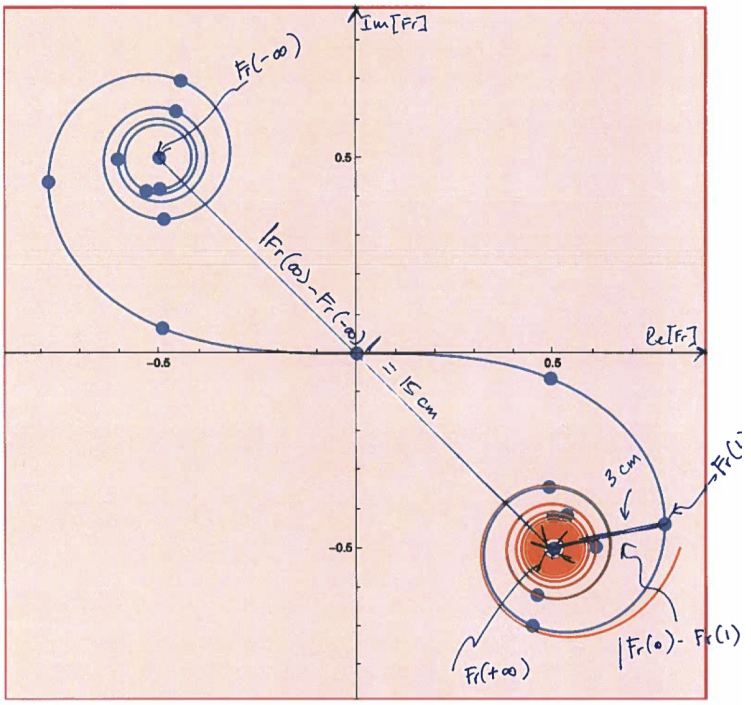


Figure 15.4: Using the Cornu spiral for knife edge gain calculations. Lengths of $\mathrm{Fr}(\infty) - \mathrm{Fr}(-\infty)$ and $\mathrm{Fr}(\infty) - \mathrm{Fr}(1)$ are measured in cm before taking their ratio to obtain the knife edge attenuation factor.

the ratio in the right, the attenuation factor, can be evaluated directly on a Cornu spiral by drawing straight lines and making ruler measurements as shown below.

Example: Let $\lambda = 1$ m, $h_1 = 1$ m, $h_2 = \infty$, $x = 0$, $z = 2$ m. In that case

$$\left|\frac{|\mathrm{Fr}(\sqrt{\frac{2}{\lambda z}}(h_2 - x)) - \mathrm{Fr}(\sqrt{\frac{2}{\lambda z}}(h_1 - x))|}{|\mathrm{Fr}(\infty) - \mathrm{Fr}(-\infty)|}\right| = \frac{|\mathrm{Fr}(\infty) - \mathrm{Fr}(1)|}{|\mathrm{Fr}(\infty) - \mathrm{Fr}(-\infty)|} = \frac{3\,\mathrm{cm}}{15\,\mathrm{cm}} = 0.2 = -13.97\,\mathrm{dB}$$

as illustrated in Figure 15.4.

We see that the Cornu spiral can be used as a computational tool much as the Smith Chart.

Edge waves

Consider a single knife edge with $h = 0$. In that case the knife edge solution is

$$\begin{aligned} E_x(x, 0, z) &= E_o e^{-jkz}\sqrt{\frac{j}{2}}\{\mathrm{Fr}(\infty) - \mathrm{Fr}(\sqrt{\frac{2}{\lambda z}}(0 - x))\} \\ &= E_o e^{-jkz}\sqrt{\frac{j}{2}}\{\mathrm{Fr}(\infty) - \mathrm{Fr}(\sqrt{\frac{2}{\lambda z}} r \sin\theta)\} \end{aligned}$$

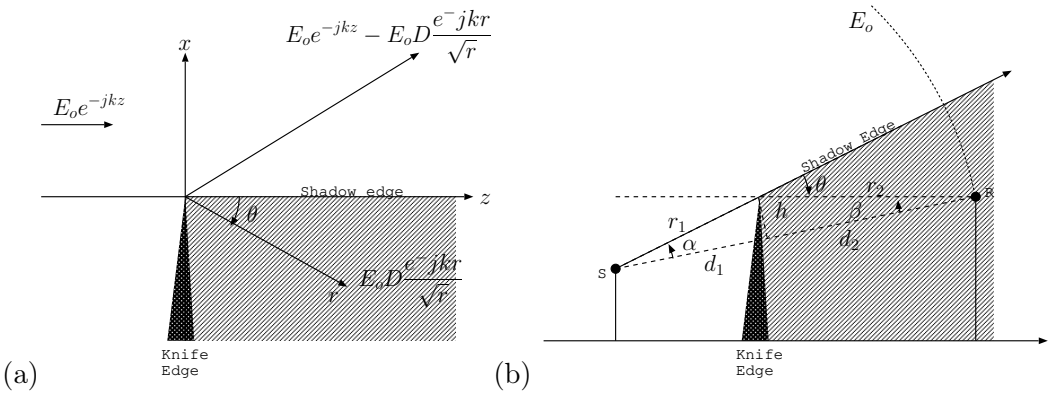


Figure 15.5: (a) Edge wave geometry with a plane wave source (S at infinity), (b) Edge wave with a point source S detected by receiver R in the shadow zone behind a knife edge: $\alpha + \beta = \theta$, $\alpha \approx h/d_1$, $\beta \approx h/d_2$, $\theta \approx h\frac{d_1+d_2}{d_1 d_2}$, $\nu = \theta\sqrt{\frac{2}{\lambda}\frac{d_1 d_2}{d_1+d_2}}$ needed for an excess phase $\frac{\pi}{2}\nu^2 \equiv k(r-d)$, $r = r_1 + r_2$ and $d = d_1 + d_2$.

at $x = -r\sin\theta$, an angle θ below the shadow edge at $x = 0$ — see Figure 15.5a. This result can be approximated using the fact that

$$\mathrm{Fr}(\nu) \equiv \int_0^\nu dp\, e^{-j\frac{\pi}{2}p^2} \approx \mathrm{Fr}(\infty) - \frac{e^{-j\frac{\pi}{2}\nu^2}}{j\pi\nu} \text{ for } \nu \gg 1$$

as shown graphically by the orange curve in Figure 15.3a. With this approximation

$$\begin{aligned} E_x(x,0,z) &= E_o e^{-jkz}\sqrt{\frac{j}{2}}\{\mathrm{Fr}(\infty) - \mathrm{Fr}(\sqrt{\frac{2}{\lambda z}}r\sin\theta)\} \\ &\approx E_o e^{-jkz}\sqrt{\frac{j}{2}}\frac{e^{-j\frac{\pi}{2}\nu^2}}{j\pi\nu}_{|\nu=\sqrt{\frac{2}{\lambda z}}r\sin\theta} = E_o e^{-jkz}\sqrt{\frac{j}{2}}\frac{e^{-jk\frac{x^2}{2z}}}{j\pi\sqrt{\frac{2}{\lambda z}}r\sin\theta} \\ &\approx E_o\frac{e^{-jkr}}{\sqrt{r}}\sqrt{\frac{\lambda z}{jr}}\frac{1}{2\pi\sin\theta} \approx E_o D\frac{e^{-jkr}}{\sqrt{r}} \end{aligned}$$

with

$$D \equiv \sqrt{\frac{\lambda}{j}}\frac{1}{2\pi\sin\theta}.$$

This result

$$E_x(x,0,z) = E_o D\frac{e^{-jkr}}{\sqrt{r}}, \quad x < 0,$$

is known as *edge wave approximation* and is valid deep in the shadow zone ($x < 0$) for

$$\sqrt{\frac{2}{\lambda z}} r \sin\theta > 1 \;\; \Rightarrow \;\; x^2 > \frac{\lambda z}{2},$$

that is outside the parabola $z = \frac{2}{\lambda}x^2$ embracing the shadow edge.

Also, by using Babinet's principle,

$$E_x(x,0,z) = E_o e^{-jkz} - E_o D \frac{e^{-jkr}}{\sqrt{r}}, \;\; x > 0,$$

once again outside the parabola $z = \frac{2}{\lambda}x^2$ embracing the shadow edge.

Edge wave approximation is not valid within the parabolic region where the full knife edge diffraction theory needs to be used.

Going back to the edge wave expression within the shadow zone,

$$E_x(x,0,z) = E_o e^{-jkz} \sqrt{\frac{j}{2}} \frac{e^{-j\frac{\pi}{2}\nu^2}}{j\pi\nu}_{|\nu=\sqrt{\frac{2}{\lambda z}} r \sin\theta},$$

and, therefore,

$$\frac{|E_x(x,0,z)|}{|E_o|} = |\sqrt{\frac{j}{2}} \frac{e^{-j\frac{\pi}{2}\nu^2}}{j\pi\nu}| = \frac{1}{\sqrt{2}\pi\nu} \approx \frac{\sqrt{\lambda/r}}{2\pi\sin\theta},$$

which can be regarded as an *intensity loss* compared to an unblocked incident field.

In the above expressions, the term $\frac{\pi}{2}\nu^2$, with $\nu = \sqrt{\frac{2}{\lambda z}} r \sin\theta$, is an "excess phase" loss suffered by the edge wave compared to the phase loss kz of the plane wave e^{-jkz} in reaching the same field position (in the absence of knife edge). To calculate the intensity loss when the incident field on knife edge is a cylindrical wave, we need to replace the excess phase $\frac{\pi}{2}\nu^2$ by $k(r-d)$, where $d = d_1 + d_2$ is the direct distance of receiver R to the source S and $r = r_1 + r_2$ is the distance to source S via the knife edge — see Figure 15.3b. In this case the excess phase $k(r-d)$ can be calculated as $\frac{\pi}{2}\nu^2$ with

$$\nu = \theta \sqrt{\frac{2}{\lambda} \frac{d_1 d_2}{d_1 + d_2}},$$

and the intensity loss as

$$\frac{|E_x(x,0,z)|}{|E_o|} = \frac{1}{\sqrt{2}\pi\nu} = \frac{1}{\sqrt{2}\pi\theta\sqrt{\frac{2}{\lambda}\frac{d_1 d_2}{d_1+d_2}}} = \frac{\sqrt{\lambda}}{2\pi\theta}\sqrt{\frac{d_1+d_2}{d_1 d_2}}.$$

Here $|E_o|$ would be the unblocked field intensity from the source, while θ is the angle below the shadow edge and $d_1$and $d_2$are the distances of the knife edge to the source and observation points (in the shadow zone, respectively). Notice that as $d_1 \to \infty$, this result goes back to the plane wave result with d_2 in place of r.

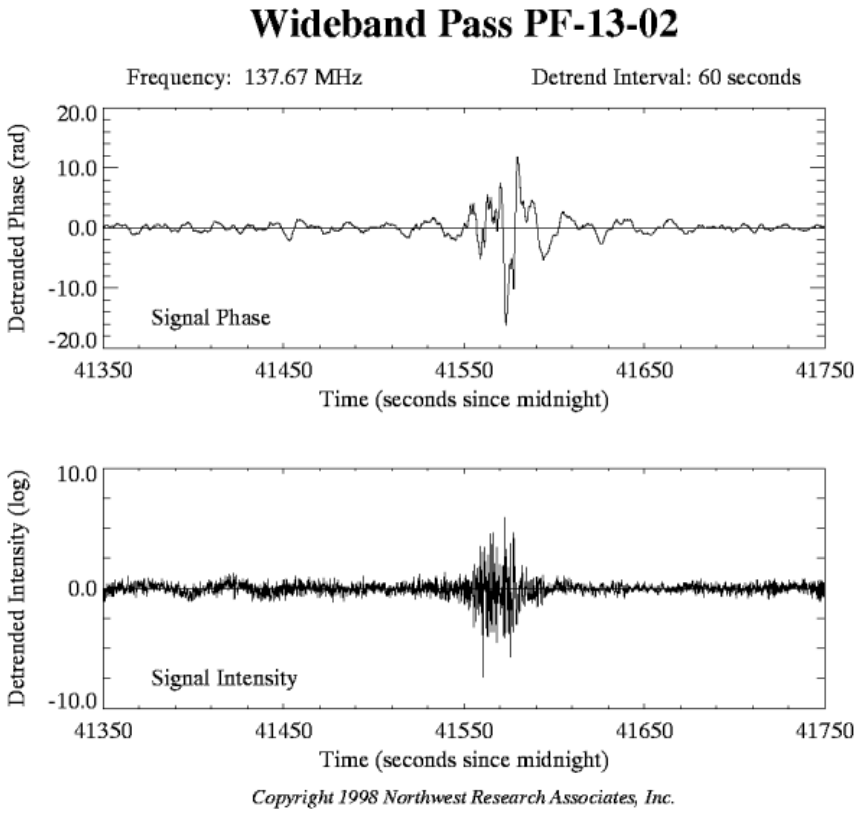


Figure 15.6: Phase and amplitude variations of a satellite signal illustrating the scintillation phenomenon (`https://spawx.nwra.com/ionoscint/sp_tutor.html`).

Radiowave scintillations

Ionospheric electron density irregularities — variations about the mean value of electron density in the ionosphere — cause the scattering and diffraction of radiowaves passing through the ionosphere. We will examine here a diffraction effect which causes the "twinkling" or *scintillation* of EM wave fields passing through the ionosphere.

Consider first a homogeneous slab ionosphere of some thickness w and a constant refractive index

$$n = \sqrt{1 - \frac{\omega_p^2}{\omega^2}} \approx 1 - \frac{\omega_p^2}{2\omega^2} = 1 - \frac{\frac{Ne^2}{m\epsilon_o}}{2\omega^2} = 1 - \frac{\lambda e^2}{4\pi\epsilon_o mc^2 k} N = 1 - \frac{\lambda r_e}{k} N,$$

where N is the ionospheric electron density, $\lambda = \frac{2\pi}{k} = \frac{c}{f}$ denotes the free-space wavelength, and

$$r_e \equiv \frac{e^2}{4\pi\epsilon_o mc^2}$$

is a fundamental constant known as *classical electron radius*. The phase delay of a plane wave propagating through the slab of ionosphere of some thickness w can therefore be

written as

$$\phi = knw = kw - \lambda r_e N w = kw - \lambda r_e N_T,$$

where

$$N_T \equiv wN$$

is the TEC (total electron content) of the slab ionosphere. Thus, a field component of a plane wave emerging out of the ionosphere at, say, $z = 0$ plane can be written as

$$E(x, y, 0) = E_o e^{-j\phi} = E_o e^{-jkw} e^{j\lambda r_e N_T},$$

assuming that E_o is the field amplitude before entering the ionosphere. However, for an inhomogeneous ionosphere with a fluctuating TEC $N_T + \delta N_T(x, y)$, we have

$$\begin{aligned} E(x, y, 0) &= E_o e^{-jkw} e^{j\lambda r_e N_T(x,y)} = E_o e^{-j(kw - \lambda r_e N_T)} e^{j\lambda r_e \delta N_T(x,y)} \\ &= \tilde{E}_o e^{j\lambda r_e \delta N_T(x,y)} \approx \tilde{E}_o (1 + j\lambda r_e \delta N_T(x, y)) \end{aligned}$$

where

$$\tilde{E}_o \equiv E_o e^{-j(kw - \lambda r_e N_T)}$$

is a complex constant and we assumed that $\lambda_o r_e |\delta N_T(x, y)| \ll 1$.

Now, substituting model expression for $E(x, y, 0)$ in Fresnel diffraction formula, we continue as

$$\begin{aligned} E(x, y, z) &= \frac{jk}{2\pi} \frac{e^{-jkz}}{z} \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' \underbrace{\tilde{E}_o (1 + j\lambda r_e \delta N_T(x', y'))}_{E(x, y, 0)} e^{-\frac{jk}{2z}((x-x')^2 + (y-y')^2)} \\ &= \tilde{E}_o e^{-jkz} [1 - \frac{r_e}{z} \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' \delta N_T(x', y') e^{-\frac{jk}{2z}((x-x')^2 + (y-y')^2)}], \end{aligned}$$

where the first part corresponding to the non-fluctuating component of $E(x, y, 0)$ was integrated out to give the first term on the right. Thus,

$$\begin{aligned} \ln E(x, y, z) &= \ln |E_x(x, y, z)| + j \angle E_x(x, y, z) \\ &\approx \ln |\tilde{E}_o| + j(\angle \tilde{E}_o - k_o z) - \frac{r_e}{z} \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' \delta N_T(x', y') e^{-\frac{jk}{2z}((x-x')^2 + (y-y')^2)}, \end{aligned}$$

under the assumption that the term $\propto \frac{r_e}{z}$ above is much less than unity in magnitude — i.e. $\ln(1+a) \approx a$ for small a. Clearly, in this "weak scintillation" case the "log-amplitude" of the wave field is

$$\ln |E(x, y, z)| \approx \ln |\tilde{E}_o| - \frac{r_e}{z} \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' \delta N_T(x', y') \cos(\frac{k}{2z}((x - x')^2 + (y - y')^2),$$

which means that its fluctuating part is (see Eqn 3.16 in Yeh and Liu, 1982)

$$\begin{aligned} &= \frac{r_e}{z} \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' \delta N_T(x', y') \cos(\frac{k}{2z}((x - x')^2 + (y - y')^2) \\ &= \delta N_T(x, y) * * \frac{r_e}{z} \cos(\frac{k}{2z}(x^2 + y^2)), \end{aligned}$$

where $**$ implies a convolution over x and y. Similarly

$$\delta N_T(x,y) * * \frac{r_e}{z} \sin(\frac{k}{2z}(x^2+y^2)) \text{ rad}$$

describes the phase scintillations, that is, the fluctuating part of $j\angle E_x(x,y,z)$.

The upshot is, the fluctuating component $\delta N_T(x,y)$ of ionospheric TEC causes fluctuations in the amplitude and phase of trans-ionospheric plane waves. Log-amplitude fluctuation levels are a filtered version of $\delta N_T(x,y)$ with a filter impulse response (see homework, using Mathematica to implement 2D FT)

$$h(x,y) = \frac{r_e}{z}\cos(\frac{k}{2z}(x^2+y^2)) \leftrightarrow H(k_x,k_y) = r_e\lambda \sin(\frac{z}{2k}(k_x^2+k_y^2)).$$

The frequency response of the filter given on the right peaks at a TEC wavelength of

$$\frac{z}{2k}(k_x^2+k_y^2) = \frac{z}{2k}(\frac{2\pi}{\lambda_T})^2 = \frac{z\lambda}{4\pi}(\frac{2\pi}{\lambda_T})^2 = \frac{\pi}{2} \quad \Rightarrow \quad \lambda_T = \sqrt{2\lambda z}$$

which is the scale in fluctuations $\delta N_T(x,y)$ which is important for amplitude scintillations — $\sqrt{2\lambda z}$ is also the minimal diameter of amplitude fluctuations observed on the ground deduced from doubling the width ρ of $h(x,y)$ obtained by setting $\frac{k\rho^2}{2z}$ to $\pi/2$.

For example, for $f = 300$ MHz $\leftrightarrow$ $\lambda = 1$ m and $z = 300$ km, the critical TEC wavelength and the ground spot size is

$$\lambda_T = \sqrt{2\lambda z} = \sqrt{2\cdot 1\cdot 300\times 10^3} = \sqrt{600\times 1000} \approx 775 \text{ m}.$$

Thus, ∼km scale fluctuations of ionospheric TEC cause a substantial scintillation effect on satellite signals observed on the ground when and where such TEC fluctuations are present. Note that λ_T decreases with increasing f which which is a reason for less severe scintillations at higher frequencies because the fluctuation amplitude of $\delta N_T(x,y)$ decreases with decreasing wavelength.

From above we also see that the phase scintillation filter function is (see homework, using Mathematica to implement 2D FT)

$$h(x,y) = \frac{r_e}{z}\sin(\frac{k}{2z}(x^2+y^2)) \leftrightarrow H(k_x,k_y) = r_e\lambda \cos(\frac{z}{2k}(k_x^2+k_y^2))$$

with the implication that mainly the TEC fluctuation scales $\lambda_T > \sqrt{2\lambda z}$ are responsible for phase scintillations.

Trans-ionospheric communication and navigation systems can be severely influenced by both amplitude and phase scintillations in particular at higher and lower latitudes where structured/irregular/fluctuating ionospheres are quite common — see Ionospheric Scintillations, GNSS and the Ionosphere, GNNS + Ionospheric Scintillations, Space Weather and GPS and Tropospheric effects on GNSS. Scintillating satellite beacon transmissions detected with "spaced antenna" systems (two or more adjacent antennas with independent receivers) can also be used to study the motion of the "scintillation causing screen" after accounting for a "point source effect" (see also) and drift measurement biases caused by random temporal variations in the screen caused by turbulence (Briggs et al., 1950 - Spaced antenna drifts).

16 Appendix IX Plasma conductivity and electrostatic waves

The AC conductivity of a *cold plasma* — a quasi-neutral gas of free charge carriers envisioned to be stationary in the absence of an applied electric field — is known to be

$$\sigma \equiv \frac{J}{E} \approx \frac{Nq^2}{mj\omega}, \tag{16.1}$$

where m and q are electronic mass and charge, ω the oscillation frequency of co-sinusoidal electric field and current density variations $\sim e^{j\omega t}$ with phasors E and J, and N the average number density of free electrons within the plasma. It is known that a cold plasma can support free oscillations at the *plasma frequency*

$$\omega_p \equiv \sqrt{\frac{Ne^2}{m\epsilon_o}} \tag{16.2}$$

satisfying a resonance condition $\sigma + j\omega\epsilon_o = 0$ pertinent to so-called *electrostatic* waves. The conductivity model (16.1) facilities an adequate description of transverse *electromagnetic* (TEM) wave propagation in collisionless and non-magnetized plasmas, but it fails short of providing an accurate description of electrostatic waves which turn out to be sensitive to the velocity distribution of plasma particles (free electrons and neutralizing ions).

In a realistic *hot* plasma where the electrons exhibit random thermal motions described by some velocity distribution function $f(\mathbf{r}, \mathbf{v}, t)$, the conductivity σ depends not only on *zeroth moment*

$$N(\mathbf{r}, t) \equiv \int f(\mathbf{r}, \mathbf{v}, t) d\mathbf{v} \tag{16.3}$$

of $f(\mathbf{r}, \mathbf{v}, t)$, but also on high order moments. While (16.3) describes the variation of electron density in a hot plasma across the 3-D volume that it occupies, the distribution $f(\mathbf{r}, \mathbf{v}, t)$ can be viewed as an electron density function in a 6-D volume consisting of space ($\mathbf{r}$) and velocity ($\mathbf{v}$) coordinates, referred to as *phase space*. In general,

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial \mathbf{r}} \cdot \frac{\partial \mathbf{r}}{\partial t} + \frac{\partial f}{\partial \mathbf{v}} \cdot \frac{\partial \mathbf{v}}{\partial t} = \frac{\partial f}{\partial t} + \mathbf{v} \cdot \frac{\partial f}{\partial \mathbf{r}} + \frac{\mathbf{F}}{m} \cdot \frac{\partial f}{\partial \mathbf{v}} \tag{16.4}$$

is the *total* time derivative of $f(\mathbf{r}, \mathbf{v}, t)$ where $\frac{\mathbf{F}}{m} \equiv \frac{\partial \mathbf{v}}{\partial t}$ denotes particle accelerations expressed in terms of an applied force $\mathbf{F}$. A plasma in which the variations of $f(\mathbf{r}, \mathbf{v}, t)$ is constrained by partial differential equation

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \frac{\partial f}{\partial \mathbf{r}} + \frac{\mathbf{F}}{m} \cdot \frac{\partial f}{\partial \mathbf{v}} = 0 \tag{16.5}$$

is said to be *collisionless*[1], as opposed to a plasma where

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \frac{\partial f}{\partial \mathbf{r}} + \frac{\mathbf{F}}{m} \cdot \frac{\partial f}{\partial \mathbf{v}} = (\frac{\delta f}{\delta t})_{coll}, \tag{16.6}$$

with the right-hand-side indicating the rate of change of $f(\mathbf{r}, \mathbf{v}, t)$ due to particle collisions. Above, (16.6) is known as *Boltzmann equation* whereas (16.5) is referred to as *Vlasov equation*.

We will next calculate the AC conductivity of a hot plasma using the Vlasov equation (16.5). What motivates us for this calculation is the question of what happens to electrostatic waves in hot but collisionless plasmas. The extension of the results to include collision effects will be addressed later on. Now, for (i) particles of charge q, (ii) an applied force $q\mathbf{E}$ — where

$$\mathbf{E} = \mathbf{E}(\mathbf{r}, t) = \hat{z} E_z(t) e^{-jkz} \tag{16.7}$$

is an electrostatic field that points and varies in direction $\hat{z}$ with some (arbitrary but real valued) wavenumber $k \geq 0$ — and (iii) a distribution function that varies as

$$f(\mathbf{r}, \mathbf{v}, t) = N g_o(\mathbf{v}) + f_1(\mathbf{v}, t) e^{-jkz}, \tag{16.8}$$

the Vlasov equation (16.5) reduces to

$$\frac{\partial f_1}{\partial t} - jkv_z f_1 + \frac{Nq}{m} E_z \frac{\partial g_o}{\partial v_z} = 0 \tag{16.9}$$

after neglecting a small nonlinear term $\propto E_z f_1$. This *linearized* Vlasov equation, namely (16.9), describes a linear time-invariant system that can sustain finite amplitude $f_1(\mathbf{v}, t)$ oscillations at a resonance frequency of $\omega_o \equiv kv_z$ even in the absence of system input $E_z(t)$; thus the system is non-dissipative and we choose to analyze it using a Laplace transform technique.

To obtain the *zero-state response* of f_1 to forcing E_z we assume that $f_1 = 0$ for $t < 0$. In that case the Laplace transform of (16.9) is

$$s\hat{f}_1 - jkv_z \hat{f}_1 + \frac{Nq}{m} \hat{E}_z \frac{\partial g_o}{\partial v_z} = 0, \tag{16.10}$$

where

$$\hat{f}_1(\mathbf{v}, s) \equiv \int_{0^-}^{\infty} f_1(\mathbf{v}, t) e^{-st} dt, \tag{16.11}$$

and a similar integral defines $\hat{E}_z(s)$. Above, s denotes a complex frequency of the form $\gamma + j\omega$, and the solution

$$\hat{f}_1 = -\frac{Nq}{m} \frac{\frac{\partial g_o}{\partial v_z}}{s - jkv_z} \hat{E}_z \tag{16.12}$$

[1] We can think of $f d\mathbf{r} d\mathbf{v}$ as the number of plasma particles in an infinitesimal 6-D volume element $d\mathbf{r} d\mathbf{v}$ centered about $(\mathbf{r}, \mathbf{v})$ at time t. An instant dt later the volume has translated itself to a new location $(\mathbf{r} + \mathbf{v}dt, \mathbf{v} + \frac{\mathbf{F}}{m} dt)$ in the 6-D phase space, but the number of particles in it has not changed unless binary collisions knock particles in/or/out of the volume. Thus, in a collisionless plasma $\frac{df}{dt} = 0$, if the applied forces $\mathbf{F}$ preserve the size of $d\mathbf{r} d\mathbf{v}$ as it travels through phase space (like the Lorentz force) — see, e.g., *Gurnett and Bhattacharjee* [2004] for more details and insight.

of (16.10) is valid only in a portion of the complex s-plane where the transform integrals converge. This *region of convergence* (ROC) lies to the right of the singularities of $\hat{f}_1$, including the pole at $s = \gamma + j\omega = jkv_z$ on the imaginary axis of s-plane. Thus the ROC of $\hat{f}_1$ is within the right-half plane (RHP) for complex s, in the region defined by $\gamma > 0$.

Next, we note that for current density

$$J_z \equiv q \int v_z f_1 d\mathbf{v} \tag{16.13}$$

driven by E_z, the Laplace transform is

$$\hat{J}_z = q \int v_z \hat{f}_1 d\mathbf{v} = -\frac{Nq^2}{m} \int \frac{v_z \frac{\partial g_o}{\partial v_z}}{s - jkv_z} d\mathbf{v} \hat{E}_z \tag{16.14}$$

also within the RHP for complex s. This result implies a transfer function

$$\sigma(s) \equiv \frac{\hat{J}_z}{\hat{E}_z} = -\frac{Nq^2}{m} \int \frac{v_z \frac{\partial g_o}{\partial v_z}}{s - jkv_z} d\mathbf{v} \tag{16.15}$$

which yields the AC conductivity of the plasma in the limit $\gamma \to 0^+$. This limit, approached from the RHP, is an *analytic continuation*[2] of transfer function (16.15) and will be computed later in this section. It is useful at this stage to simplify (16.15) as

$$\sigma(s) = -\frac{Nq^2}{m} \int_{-\infty}^{\infty} \frac{v_z G_o'(v_z)}{s - jkv_z} dv_z, \tag{16.16}$$

where

$$G_o(v_z) \equiv \int dv_x \int dv_y g_o(v_x, v_y, v_z) \tag{16.17}$$

is a 1-D velocity distribution, and

$$G_o'(v_z) \equiv \frac{\partial G_o}{\partial v_z}. \tag{16.18}$$

Consider next the Ampere's law,

$$\nabla \times \mathbf{H} = \mathbf{J} + \epsilon_o \frac{\partial \mathbf{E}}{\partial t}, \tag{16.19}$$

in a medium where a conduction current $\mathbf{J}$ is driven by an electric field $\mathbf{E}$. For electrostatic waves $\mathbf{H} = 0$, and, therefore, the Laplace transform of scalar components of (16.19) are

$$0 = \hat{J} + \epsilon_o s \hat{E} - \epsilon_o E(0^-) = (\epsilon_o s + \sigma(s))\hat{E} - \epsilon_o E(0^-), \tag{16.20}$$

[2] The knowledge of an analytic function within *any* portion of the complex plane is sufficient to determine the function everywhere else on the plane except for the points where the function is singular — see, e.g., *Byron and Fuller* [1992], for a discussion of this property of analytic functions (known as *identity theorem*) and related concept of analytic continuation. Thus, there is a unique analytic extension of conductivity function of a plasma from the domain of real ω or imaginary $j\omega$ into the domain of complex $s = \gamma + j\omega$, an extension which is readily identified in the RHP for s using the Laplace transform approach. The limit $\gamma \to 0^+$ then takes us back onto the $j\omega$ axis where the AC conductivity is needed.

and

$$\hat{E}(s) = \frac{1}{s(1+\frac{\sigma(s)}{s\epsilon_o})} E(0^-). \tag{16.21}$$

In (16.21), $E(0^-)$ denotes an initial wave amplitude and the expression itself implies that for $t > 0$ the amplitude will evolve as a superposition of characteristic modes of the system of the form $\propto e^{st}$, where s are the roots of the characteristic equation of the system, namely

$$s\epsilon_r(s) = 0, \tag{16.22}$$

with

$$\epsilon_r(s) \equiv 1 + \frac{\sigma(s)}{s\epsilon_o} \tag{16.23}$$

denoting the *relative permittivity* or *dielectric constant* of the medium (the square of radio refractive index pertinent to electromagnetic waves in $\gamma \to 0^+$ limit). Substituting (16.16) in (16.23), we find that

$$\epsilon_r(s) = 1 - \frac{\omega_p^2}{s} \int_{-\infty}^{\infty} \frac{v_z G_o'(v_z)}{s - jkv_z} dv_z. \tag{16.24}$$

Now, for functions $G_o'(v_z)$ with an odd symmetry, (16.24) can be re-expressed as

$$\epsilon_r(s) = 1 - \frac{\omega_p^2}{jk} \int_{-\infty}^{\infty} \frac{G_o'(v_z)}{s - jkv_z} dv_z, \tag{16.25}$$

while with a *Maxwellian* distribution

$$G_o(v_z) = \frac{e^{-\frac{\frac{1}{2}m(v_z-\bar{v}_z)^2}{KT}}}{\sqrt{\frac{2\pi KT}{m}}} = \frac{e^{-\frac{(v_z-\bar{v}_z)^2}{2C^2}}}{\sqrt{2\pi}C} \tag{16.26}$$

we have

$$\epsilon_r(s) = 1 + \frac{1}{k^2\lambda_d^2}\frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \frac{xe^{-x^2}}{x - \frac{s-jk\bar{v}_z}{jk\sqrt{2}C}} dx \equiv 1 + \frac{1 - j\theta J(\theta)}{k^2\lambda_d^2}, \tag{16.27}$$

where K is *Boltzmann constant*, $\bar{v}_z$ is the mean particle velocity, $C \equiv \sqrt{\frac{KT}{m}}$is thermal speed, $\sqrt{2}C$ is the most probable random speed, $\lambda_d \equiv \frac{C}{\omega_p} = \sqrt{\frac{KT\epsilon_o}{Ne^2}}$ is *Debye length,*

$$\theta \equiv \frac{s - jk\bar{v}_z}{jk\sqrt{2}C}, \tag{16.28}$$

and

$$J(\theta) \equiv \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \frac{j}{x - \theta} e^{-x^2} dx. \tag{16.29}$$

Also, using (16.27) and (16.23) we can write

$$\sigma(s) = \frac{s\epsilon_o}{k^2\lambda_d^2}[1 - j\theta J(\theta)]. \tag{16.30}$$

It is important to remember at this stage that $\sigma(s)$ and $J(\theta)$ were specified for $s = \gamma + j\omega$ in the RHP, or equivalently, for complex $\theta = \frac{\omega - k\bar{v}_z - j\gamma}{k\sqrt{2}C}$ in the bottom-half of θ-plane. But in the bottom-half plane (BHP), that is for $\mathrm{Im}\{\theta\} < 0$, it can be easily verified that

$$\int_0^\infty e^{j(x-\theta)t} dt = \frac{j}{(x-\theta)}. \tag{16.31}$$

Using (16.31) in (16.29) we obtain

$$\begin{aligned} J(\theta) &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \{ \int_0^\infty e^{j(x-\theta)t} dt \} e^{-x^2} dx = \int_0^\infty dt e^{-j\theta t} \int_{-\infty}^{\infty} dx e^{jxt} \frac{e^{-x^2}}{\sqrt{\pi}} \\ &= \int_0^\infty dt e^{-j\theta t} e^{-t^2/4}, \end{aligned} \tag{16.32}$$

a result known as *Gordeyev integral.*

The Gordeyev integral is effectively a Laplace transform of an absolute integrable function $e^{-t^2/4}$ with an ROC that covers the entire complex θ-plane [*Gordeyev,* 1952]. Thus, expressions (16.27) and (16.30) for $\epsilon_r(s)$ and $\sigma(s)$ in terms of the Gordeyev integral are analytic over the entire complex s-plane and in fact represent analytic continuations of $\epsilon_r(s)$ and $\sigma(s)$ out of the RHP for complex s. Consequently, the AC conductivity of a hot plasma, for instance, is obtained by evaluating (16.30) at $s = j\omega$, in which case $\theta = \frac{\omega - k\bar{v}_z}{k\sqrt{2}C}$. That procedure is in effect taking the $\gamma \to 0^+$ limit for $\sigma(s)$.

While the above discussion was focused on electron response, the AC conductivity of a plasma is determined by the motions of all charge carriers, and thus should be expressed as a sum

$$\sigma(\omega, \mathbf{k}) = \sum_s \sigma_s(\omega, \mathbf{k}) \equiv \sum_s \frac{\omega \epsilon_o}{k^2 \lambda_s^2} [j + \theta_s J(\theta_s)], \tag{16.33}$$

where

$$\theta_s \equiv \frac{\omega - k\bar{v}_{zs}}{k\sqrt{2}C_s} \tag{16.34}$$

in terms of thermal speed and mean velocity C_s and $\bar{v}_{zs}$ of charged species s, and λ_s is the corresponding Debye length. Also, the expression in (16.33) within square brackets, i.e.,

$$y(\theta_s) \equiv j + \theta_s J(\theta_s), \tag{16.35}$$

is known as normalized *admittance* for species s and is used to express the equilibrium spectrum of electrostatic waves in a hot plasma (see next section).

Finally, the Gordeyev integral satisfies an initial condition $J(0) = \sqrt{\pi}$ and a differential equation

$$\frac{dJ}{d\theta} + 2\theta J(\theta) = -j2, \tag{16.36}$$

as can be easily confirmed using (16.32). Applying an integrating factor e^{θ^2} to (16.36), and integrating the result from 0 to θ it can be shown that

$$J(\theta) = -j2e^{-\theta^2} \int_0^\theta e^{t^2} dt + \sqrt{\pi} e^{-\theta^2}. \tag{16.37}$$

Using this alternate expression[3] for $J(\theta)$, the dispersion relation

$$\epsilon_r(s) = 0 \tag{16.38}$$

for electrostatic waves can be written as

$$1 + (k\lambda_d)^{-2}[1 - 2\theta e^{-\theta^2} \int_0^\theta e^{t^2} dt - j\sqrt{\pi}\theta e^{-\theta^2}] = 0, \tag{16.39}$$

from which it is easy to see that there can be no real valued roots θ. In fact, all the roots of (16.39) are found in the top-half plane for θ, implying *damped* electrostatic modes in a hot plasma. The collisionless damping of electrostatic waves implied here is known as *Landau damping*, a feature of wave physics which was first explained by *Landau* [1946]. The damping applies to current density waves J_z but not to f_1 oscillations that contribute to J_z — thus, Landau damping is in essence a consequence of phase mixing, randomization of the phases of f_1 contributing to J_z, a dispersion-like phenomenon occurring in time. The absence of "growing" roots of (16.39) is of course a consequence of the absence of free energy in a plasma with Maxwellian velocity distributions.

[3]To compute $J(\theta)$ in this form we can make use of Dawson's integral $F(\theta) = e^{-\theta^2} \int_0^\theta e^{t^2} dt$ defined in MATLAB and *Mathematica* — see `http://mathworld.wolfram.com/DawsonsIntegral.html` — in terms of error function with complex argument, Erfi. Finally, $J(\theta)$ is the same as $jZ(\theta)$, where $Z(\theta)$ is the so-called *plasma dispersion function* tabulated by *Fried and Compte* [1962].

17 Appendix X Electrostatic plasma waves and Landau damping

The AC conductivity of a *cold plasma* — a quasi-neutral gas of free charge carriers envisioned to be stationary in the absence of an applied electric field — is known to be

$$\sigma \equiv \frac{J}{E} \approx \frac{Nq^2}{mj\omega}, \tag{17.1}$$

where m and q are electronic mass and charge, ω the oscillation frequency of co-sinusoidal electric field and current density variations $\sim e^{j\omega t}$ with phasors E and J, and N the average number density of free electrons within the plasma. It is known that a cold plasma can support free oscillations at the *plasma frequency*

$$\omega_p \equiv \sqrt{\frac{Ne^2}{m\epsilon_o}} \tag{17.2}$$

satisfying the so-called e*lectrostatic wave* dispersion relation

$$\sigma + j\omega\epsilon_o = 0 \tag{17.3}$$

with $\omega = \pm\omega_p$ roots. The conductivity model (17.1) facilities an adequate description of transverse *electromagnetic* (TEM) wave propagation in collisionless and non-magnetized plasmas, but it fails short of providing an accurate description of electrostatic waves which turn out to be sensitive to the velocity distribution of plasma particles (free electrons and neutralizing ions).

In a realistic *hot* plasma where the electrons exhibit random thermal motions described by some velocity distribution function $f(\mathbf{r}, \mathbf{v}, t)$, the conductivity σ depends not only on *zeroth moment*

$$N(\mathbf{r}, t) \equiv \int f(\mathbf{r}, \mathbf{v}, t) d\mathbf{v} \tag{17.4}$$

of $f(\mathbf{r}, \mathbf{v}, t)$, but also on high order moments. While (17.4) describes the variation of electron density in a hot plasma across the 3-D volume that it occupies, the distribution $f(\mathbf{r}, \mathbf{v}, t)$ can be viewed as an electron density function in a 6-D volume consisting of space ($\mathbf{r}$) and velocity ($\mathbf{v}$) coordinates, referred to as *phase space*. In general,

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial \mathbf{r}} \cdot \frac{\partial \mathbf{r}}{\partial t} + \frac{\partial f}{\partial \mathbf{v}} \cdot \frac{\partial \mathbf{v}}{\partial t} = \frac{\partial f}{\partial t} + \mathbf{v} \cdot \frac{\partial f}{\partial \mathbf{r}} + \frac{\mathbf{F}}{m} \cdot \frac{\partial f}{\partial \mathbf{v}} \tag{17.5}$$

is the *total* time derivative of $f(\mathbf{r}, \mathbf{v}, t)$ where $\frac{\mathbf{F}}{m} \equiv \frac{\partial \mathbf{v}}{\partial t}$ denotes particle accelerations expressed in terms of an applied force $\mathbf{F}$. A plasma in which the variations of $f(\mathbf{r}, \mathbf{v}, t)$

is constrained by partial differential equation

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \frac{\partial f}{\partial \mathbf{r}} + \frac{\mathbf{F}}{m} \cdot \frac{\partial f}{\partial \mathbf{v}} = 0 \tag{17.6}$$

is said to be *collisionless*[1], as opposed to a plasma where

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \frac{\partial f}{\partial \mathbf{r}} + \frac{\mathbf{F}}{m} \cdot \frac{\partial f}{\partial \mathbf{v}} = (\frac{\delta f}{\delta t})_{coll}, \tag{17.7}$$

with the right-hand-side indicating the rate of change of $f(\mathbf{r}, \mathbf{v}, t)$ due to particle collisions. Above, (17.7) is known as *Boltzmann equation* whereas (17.6) is referred to as *Vlasov equation*.

We will next calculate the AC conductivity of a hot plasma using the Vlasov equation (17.6). What motivates us for this calculation is the question of what happens to electrostatic waves in hot but collisionless plasmas. The extension of the results to include collision effects will be addressed later on. Now, for (i) particles of charge q, (ii) an applied force $q\mathbf{E}$, where

$$\mathbf{E} = \mathbf{E}(\mathbf{r}, t) = \hat{z}\mathrm{Re}\{E_z(t)e^{ikz}\} \tag{17.8}$$

is an electrostatic field that points and varies in direction $\hat{z}$ with some (arbitrary but real valued) wavenumber $k \geq 0$, and (iii) a distribution function that varies as

$$f(\mathbf{r}, \mathbf{v}, t) = Ng_o(\mathbf{v}) + \mathrm{Re}\{f_1(\mathbf{v}, t)e^{ikz}\}, \tag{17.9}$$

we obtain from the Vlasov equation (17.6)

$$\frac{\partial f_1}{\partial t} + ikv_z f_1 + \frac{Nq}{m} E_z \frac{\partial g_o}{\partial v_z} = 0 \tag{17.10}$$

after neglecting a small nonlinear term $\propto E_z f_1$. This *linearized* Vlasov equation, namely (17.10), describes a linear time-invariant system that can sustain finite amplitude $f_1(\mathbf{v}, t)$ oscillations at a resonance frequency of $\omega_o \equiv kv_z$ even in the absence of system input $E_z(t)$; thus the system is non-dissipative and we choose to analyze it using a Laplace transform technique.

To obtain the *zero-state response* of f_1 to forcing E_z we assume that $f_1 = 0$ for $t < 0$. In that case the Laplace transform of (16.9) is

$$-i\omega \hat{f}_1 + ikv_z \hat{f}_1 + \frac{Nq}{m} \hat{E}_z \frac{\partial g_o}{\partial v_z} = 0, \tag{17.11}$$

[1] We can think of $f d\mathbf{r} d\mathbf{v}$ as the number of plasma particles in an infinitesimal 6-D volume element $d\mathbf{r} d\mathbf{v}$ centered about $(\mathbf{r}, \mathbf{v})$ at time t. An instant dt later the volume has translated itself to a new location $(\mathbf{r} + \mathbf{v}dt, \mathbf{v} + \frac{\mathbf{F}}{m}dt)$ in the 6-D phase space, but the number of particles in it has not changed unless binary collisions knock particles in/or/out of the volume. Thus, in a collisionless plasma $\frac{df}{dt} = 0$, if the applied forces $\mathbf{F}$ preserve the size of $d\mathbf{r} d\mathbf{v}$ as it travels through phase space (like the Lorentz force) — see, e.g., *Gurnett and Bhattacharjee* [2004] for more details and insight.

where[2]

$$\hat{f}_1(\mathbf{v},\omega) \equiv \int_{0^-}^{\infty} f_1(\mathbf{v},t)e^{i\omega t}dt, \tag{17.12}$$

and a similar integral defines $\hat{E}_z(\omega)$. The solution

$$\tilde{f}_1 = -\frac{Nq}{m}\frac{\frac{\partial}{\partial v_z}g_o}{-i\omega + ikv_z}\tilde{E}_z. \tag{17.13}$$

of (17.11) is valid only in a portion of the complex ω-plane where the transform integrals converge. This *region of convergence* (ROC) lies above the singularities of $\hat{f}_1$, including the pole at $\omega = kv_z$ on the real axis of ω-plane. Thus the ROC of $\hat{f}_1$ is the upper-half plane (UHP) for complex ω, in the region defined by $\mathrm{Im}\{\omega\} > 0$.

Next, we note that the current density $\tilde{J}_z$ in the Laplace domain driven by $\tilde{E}_z$ is

$$\tilde{J}_z = q\int v_z\tilde{f}_1 d\mathbf{v} = -\frac{Nq^2}{m}\int \frac{v_z\frac{\partial}{\partial v_z}g_o}{-i\omega + ikv_z}d\mathbf{v}\tilde{E}_z \tag{17.14}$$

and thus a transfer function corresponding to the AC conductivity (when evaluated with real valued ω) of the collisionless plasma is obtained as

$$\sigma \equiv \frac{\tilde{J}_z}{\tilde{E}_z} = -\frac{Nq^2}{m}\int \frac{v_z\frac{\partial}{\partial v_z}g_o(v_x,v_y,v_z)}{-i\omega + ikv_z}d\mathbf{v} = \frac{i\epsilon_o\omega_p^2}{k}\int_{-\infty}^{\infty}\frac{v_z G_o'(v_z)}{v_z - \frac{\omega}{k}}dv_z, \tag{17.15}$$

where $\omega_p^2 = Nq^2/m\epsilon_o$ and

$$G_o'(v_z) \equiv \frac{\partial}{\partial v_z}\int dv_x \int dv_y\, g_o(v_x,v_y,v_z). \tag{17.16}$$

Since function $G_o'(v_z)$ is independent of ω, our conductivity (17.15) as well as current density (17.14) formulae contain a single pole at $\omega = kv_z$ on the real axis of the complex ω-plane — this makes the upper half of ω-plane to be the ROC of these results.

Furthermore, because of the very same pole, the conductivity formula (17.15) for real valued ω contains a non-convergent integral — it fails to furnish a σ value when ω is a real number. But when ω is a complex number, residing in upper half (ROC) or lower half (non-ROC) of ω-plane, the integral in (17.15) becomes convergent — since the pole at $v_z = \omega/k$ is then moved away from the integration path along the real axis of the complex v_x-plane — but to conflicting functions of ω of which only the one resulting from the ROC is physically valid[3].

Therefore the correct conductivity formula applicable with real valued ω can be obtained from (17.15) by using it within its ROC, i.e., by performing the integral assuming

[2] We are using Physics notation here, $e^{i\omega t}$ is used instead of e^{-st}, with a complex frequency $\omega = \omega_r + i\omega_i$ in place of complex frequency $s = \sigma + j\omega$ used in ECE literature. RHP of complex s corresponds to the upper half plane (UHP) of ω both of which need to be pole-free in BIBO stable system transfer functions.

[3] Such as *not leading to a conductivity σ with a negative real part* within a Maxwellian plasma, so that there is no wave growth in thermal equilibrium, which would be at odds with being in equilibrium

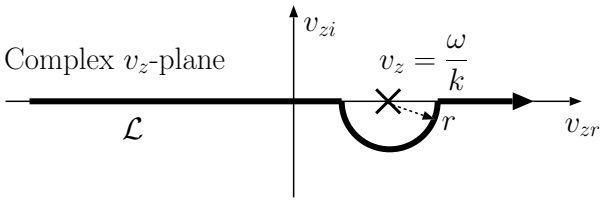


Figure 17.1: Landau contour — $\lim_{r\to 0}$ is taken. Straight path results in a *principal value integral* while the semi-circle contributes half a *residue*. This contour provides analytic continuation from the UHP of complex ω to the rest of the ω-plane. To obtain analytic continuation from LHP the contour should be indented upwards.

that ω is within the upper half plane and then using the resulting function of ω with real valued ω as an "analytic continuation". Alternatively, the same conductivity formula can also be obtained from (17.15) evaluated with real valued ω, provided that the straight line integration contour along v_z is indented as shown in Figure 17.1 so that it goes underneath the real valued pole at $v_z = \omega/k$ — for ease of evaluation of such a contour integral (see below) a semi-circle of a vanishing radius is employed as shown, even though the result of integration is independent of the actual shape and size of the indentation. The reason for this and the freedom to alter a path of integration on the complex plane in general comes from the path independence of complex plane integrals of analytic functions so long as the altered path does not intersect or go across any of the singularities of the integrand. In our particular case as ω descends from the UHP ROC of (17.14) down to the real ω axis, the descending pole of (17.15) integrand at $v_z = \omega/k$ will avoid intersecting the strategically indented integration path shown in Figure 17.1, resulting in, once more, the analytic continuation of the output of (17.15) from the complex UHP of ω to the domain of real ω. The result so achieved for real valued ω, as first derived and explained by Lev Landau, turns out to be an analytic function of ω across the entire complex ω-plane — the function is called "entire" for that reason — and therefore can be evaluated with all ω, real or complex.

Following Landau's recipe, we obtain the collisionless AC conductivity as

$$\sigma = \frac{i\epsilon_o\omega_p^2}{k}\int_{\mathcal{L}} \frac{v_z G_o'(v_z)}{v_z - \frac{\omega}{k}} dv_z \tag{17.17}$$

such that

$$\int_{\mathcal{L}} \frac{v_z G_o'(v_z)}{v_z - \frac{\omega}{k}} dv_z = \mathcal{P}\int_{-\infty}^{\infty} \frac{v_z G_o'(v_z)}{v_z - \frac{\omega}{k}} dv_z + i\pi\frac{\omega}{k} G_o'(\frac{\omega}{k}), \tag{17.18}$$

where

$$\mathcal{P}\int_{-\infty}^{\infty} \frac{v_z G_o'(v_z)}{v_z - \frac{\omega}{k}} dv_z \equiv \lim_{r\to 0} \int_{-\infty}^{\frac{\omega}{k}-r} \frac{v_z G_o'(v_z)}{v_z - \frac{\omega}{k}} dv_z + \int_{\frac{\omega}{k}+r}^{\infty} \frac{v_z G_o'(v_z)}{v_z - \frac{\omega}{k}} dv_z \tag{17.19}$$

is known as a *Cauchy principal value integral*, and

$$i\pi\frac{\omega}{k}G_o'(\frac{\omega}{k}) = \lim_{r\to 0}\int_{\theta=\pi}^{2\pi}\frac{(\frac{\omega}{k}+re^{i\theta})G_o'(\frac{\omega}{k}+re^{i\theta})}{\frac{\omega}{k}+re^{i\theta}-\frac{\omega}{k}}ire^{i\theta}d\theta \tag{17.20}$$

is the integral taken along the semi-circle with vanishing radius going underneath the singularity — the result[4] is *one-half* the *residue* of the integrand $\frac{v_z G_o'(v_z)}{v_z-\frac{\omega}{k}}$ at $v_z = \omega/k$.

Accordingly, the AC conductivity of the hot plasma becomes

$$\sigma(\omega,k) = \frac{\epsilon_o\omega_p^2}{k}[i\mathcal{P}\int_{-\infty}^{\infty}\frac{v_z G_o'(v_z)}{v_z-\frac{\omega}{k}}dv_z - \pi\frac{\omega}{k}G_o'(\frac{\omega}{k})] \tag{17.21}$$

and contains real and imaginary parts in contrast to the purely imaginary conductivity of collisionless cold plasmas (e.g., in ECE 350).

Historical note: Hot plasma conductivity and the dispersion relation of electrostatic waves for such plasmas were first derived by Anatoly Vlasov. To avoid the singularity at $v_z = \omega/k$ Vlasov integrated v_z from $-\infty$ to ω/k and from ω/k to ∞ and thereby obtaining (17.21) but without the last real-valued term coming from the residue of the singularity. Lev Landau's later derivation (Landau, 1946) using the analytic continuation approach led to (17.21) as shown. The concept of Landau contour arose from this work. The surprising result was the "collisionless damping" implied by the real part in (17.21) — this was controversial at the beginning but was confirmed experimentally in the early 1960's in lab plasmas. It turns out that collisionless (non-Ohmic) damping is a possibility because *wave-particle interactions* facilitate a transfer of energy between waves and particles in plasma-like systems — from waves to particles (wave damping) as well as from particles to waves (wave growth), depending on the velocity distribution of particles having comparable velocities with the waves propagating in the system. Collisionless "Landau damping" applies only to electrostatic waves $J_z \propto E_z$ (and charge density waves because of the continuity equation relation between ρ and J_z), but not to f_1 oscillations that contribute to J_z — thus, Landau damping is in essence a consequence of phase mixing, randomization of the phases of f_1 contributing to J_z, a dispersion-like phenomenon occurring in time. In hot plasmas damped electrostatic waves (electric field, current, density) will also be re-generated as a consequence of a very general *fluctuation-dissipation law* in action (a generalized Nyquist noise theorem). Therefore there remains a "thermal level" of random $\rho \propto J_z \propto E_z$ waves in plasmas in thermal equilibrium whose variance is modeled by the *incoherent scatter radar* (ISR) theory. Another confirmation of Landau damping is therefore the consistency of ISR theory (that implicitly includes Landau damping) and ISR measurements of space plasmas described later on.

The dispersion relation of electrostatic waves[5] propagating in z direction follows from the z component of Ampere's law

$$\hat{z}\cdot\nabla\times\tilde{\mathbf{H}} = \sigma\tilde{E}_z - \epsilon_o i\omega\tilde{E}_z = -i\omega\epsilon_o(1-\frac{\sigma}{i\omega\epsilon_o})\tilde{E}_z = 0 \tag{17.22}$$

[4]Notice that the result is remarkably simplified because of the vanishing size of radius r.
[5]Electrostatic waves have $\mathbf{H}$=0.

where we used $\tilde{J}_z = \sigma \tilde{E}_z$. This implies a dispersion relation[6]

$$\epsilon_r \equiv 1 - \frac{\sigma}{i\omega\epsilon_o} = 0, \tag{17.23}$$

which is explicitly — using (17.21) —

$$1 - \frac{\omega_p^2}{\omega k}[\mathcal{P}\int_{-\infty}^{\infty} \frac{v_z G_o'(v_z)}{v_z - \frac{\omega}{k}} dv_z + i\pi \frac{\omega}{k} G_o'(\frac{\omega}{k})] = 0 \tag{17.24}$$

In summary, in hot plasmas (17.21) replaces (17.1) as the relevant conductivity and (17.24) replaces $\omega = \pm\omega_p$ as the relevant electrostatic wave dispersion relation.

All dispersion relations, including (17.24) for electrostatic waves in hot plasmas, are used to find out which ω the waves described by the dispersion relation oscillate with given a wavenumber k. In a cold plasma the answer is $\omega = \pm\omega_p$, independent of value of k, implying undamped electrostatic waves of arbitrary phase velocities $\pm\omega_p/k$. In case of electrostatic waves in hot plasmas, the roots of (17.24) will determine the permissible ω and phase velocities ω/k of what will turn out to be *damped* electrostatic waves (in Maxwellian plasmas) with complex valued $\omega = \omega_r + i\omega_i$ with $\omega_i < 0$. It is essential that the left hand side of (17.24) is an entire function of ω (i.e., not valid only with real ω) for us to be able to reach that conclusion (to reach a complex vaued root) and we owe our understanding of this subtlety to the pioneering work of Lev Landau.

Solutions for Maxwellian plasmas

Let

$$G_o(v_z) = \frac{e^{-\frac{v_z^2}{2C^2}}}{\sqrt{2\pi}C}$$

where $C = \sqrt{KT/m}$ is electron thermal speed — this specifies a Maxwellian plasma at temperature T with a velocity variance $\langle v_z^2 \rangle = C^2$ and an average kinetic energy of $\frac{1}{2}m\langle v_z^2 \rangle = \frac{1}{2}KT$ per degree of freedom of plasma particles. It follows that

$$G_o'(v_z) = -\frac{v_z}{C^2}\frac{e^{-\frac{v_z^2}{2C^2}}}{\sqrt{2\pi}C} = -\frac{v_z}{C^2}G_o(v_z).$$

Now, switching to ECE notation ($i \to -j$), the conductivity becomes

$$\sigma(\omega, k) = \frac{\epsilon_o \omega_p^2}{kC^2}[jP\int_{-\infty}^{\infty} \frac{v_z^2 G_o(v_z)}{v_z - \frac{\omega}{k}} dv_z + \pi \frac{\omega^2}{k^2} G_o(\frac{\omega}{k})] \tag{17.25}$$

and dispersion relation

$$1 + \frac{\omega_p^2}{\omega k C^2}[\mathcal{P}\int_{-\infty}^{\infty} \frac{v_z^2 G_o(v_z)}{v_z - \frac{\omega}{k}} dv_z - j\pi \frac{\omega^2}{k^2} G_o(\frac{\omega}{k})] = 0. \tag{17.26}$$

[6] This is the "high frequency" dispersion relation depending only on electron dynamics. A more complete dispersion relation will have $\sigma = \sigma_e + \sigma_i$, with $\sigma_{e,i}$ denoting electron and ion conductivities modeled as (17.21) but with electron and ion parameters $m_{e,i}$, $G_{e,i}(v_z)$, etc. At high frequencies $\sigma \approx \sigma_e$ since $|\sigma_i| \ll |\sigma_e|$ due to the fact that $m_e \ll m_i$.

Long wavelength approximation and Landau damping

In this section the roots of (17.26) will be derived using the long wavelength approximation to simplify the evaluation of the principal value integral. Initially ignore the Landau term in (17.26) so that

$$1 + \frac{\omega_p^2}{\omega k C^2} \mathcal{P} \int_{-\infty}^{\infty} \frac{v_z^2 G_o(v_z)}{v_z - \frac{\omega}{k}} dv_z = 0. \tag{17.27}$$

Assume $v_z \ll \omega/k$ across the support of $G_o(v_z)$, in which case

$$\frac{1}{1 - \frac{v_z k}{\omega}} \approx 1 + \frac{v_z k}{\omega} + (\frac{v_z k}{\omega})^2 + (\frac{v_z k}{\omega})^3 + \cdots \tag{17.28}$$

within (17.27), leading to

$$1 - \frac{\omega_p^2}{k^2 C^2} \mathcal{P} \int_{-\infty}^{\infty} [1 + \frac{v_z k}{\omega} + (\frac{v_z k}{\omega})^2 + (\frac{v_z k}{\omega})^3 + \cdots] \frac{v_z^2 k^2}{\omega^2} G_o(v_z) dv_z \approx 0. \tag{17.29}$$

The second and fourth terms integrate out to zero because of odd integrands and we are left with

$$1 - \frac{\omega_p^2}{k^2 C^2} \mathcal{P} \int_{-\infty}^{\infty} [(\frac{v_z k}{\omega})^2 + (\frac{v_z k}{\omega})^4] G_o(v_z) dv_z \approx 0 \tag{17.30}$$

leading to (using the central moments of a Gaussian)

$$1 - \frac{\omega_p^2}{k^2 C^2} [(\frac{kC}{\omega})^2 + 3(\frac{kC}{\omega})^4] \approx 0 \;\Rightarrow\; 1 - \frac{\omega_p^2}{\omega^2} [1 + 3(\frac{kC}{\omega})^2] \approx 0. \tag{17.31}$$

For a cold plasma $C = 0$ and therefore $\omega^2 = \omega_p^2$ for electrostatic waves in cold plasmas. We use this within (17.31) to find the corrected (iterated) frequency of electrostatic waves or *Langmuir waves* within hot plasmas — the result is

$$\omega^2 \approx \omega_p^2 + 3k^2 C^2. \tag{17.32}$$

This result is known as *Bohm and Gross* dispersion relation[7].

The condition $\omega/k \approx \omega_p/k \gg v_z$ required for this calculation will hold for most particles in the plasma when v_z is capped by C, in which case we require $\omega_p/k \gg C$ or $1 \gg kh$ where

$$h \equiv \frac{C}{\omega_p} \tag{17.33}$$

is electron Debye length (shielding scale in a plasma). That means that Bohm and Gross formula is valid for perturbations with wavelengths $\lambda = 2\pi/k$ exceeding the Debye length h.

[7] Which permits electrostatic waves when $\omega > \omega_p$ with $v_p v_g = 3C^2$. Can also be derived using fluid theory appropriate for warm plasmas — see *Bittencourt*, Chapter 17. So for warm/hot plasmas ω_p is a cutoff frequency for electrostatic waves just like for electromagnetic waves. Lower frequency electrostatic waves that can exist in a plasma are described by more complete dispersion relations including the effect of ion dynamics — like in ISR theory.

When the Landau term of (17.24) is retained, we obtain, in place of (17.31),

$$1 - \frac{\omega_p^2}{\omega^2}[1 + 3(\frac{Ck}{\omega})^2] + j\pi\frac{\omega_p^2}{k^2}G_o'(\frac{\omega_p}{k}) \approx 0. \tag{17.34}$$

Consequently

$$\omega^2 \approx \omega_p^2 + 3k^2C^2 - j\pi\frac{\omega^2\omega_p^2}{k^2}G_o'(\frac{\omega_p}{k})$$

leading to

$$\omega^2[1 + j\pi\frac{\omega_p^2}{k^2}G_o'(\frac{\omega_p}{k})] \approx \omega_p^2 + 3k^2C^2 \;\Rightarrow\; \omega \approx \sqrt{\omega_p^2 + 3k^2C^2}[1 - j\pi\frac{\omega_p^2}{2k^2}G_o'(\frac{\omega_p}{k})].$$

So, finally, we have

$$\omega \approx \sqrt{\omega_p^2 + 3k^2C^2} - j\pi\frac{\omega_p^3}{2k^2}G_o'(\frac{\omega_p}{k}).$$

Hence electrostatic waves

$$\tilde{E}_z \propto e^{j\omega t} = e^{j\sqrt{\omega_p^2+3k^2C^2}t}e^{\pi\frac{\omega_p^3}{2k^2}G_o'(\frac{\omega_p}{k})t}$$

decay or *grow* in time exponentially depending on $G_o'(\frac{\omega_p}{k}) \lessgtr 0$. In a Maxwellian plasma $G_o'(\frac{\omega_p}{k}) < 0$ and electrostatic waves are damped. In a non-Maxwellian plasma where $G_o(v_z)$ has a high velocity "bump", it is possible to have $G_o'(\frac{\omega_p}{k}) > 0$ — in that case plasma waves grow and we talk about a "bump-on-tail" or "two-stream" type kinetic plasma instabilities.

Physics of Landau damping and bump-on-tail instability

Dawson [Phys. Rev., **118**, 381; 1960] shows that electrostatic waves with phase velocities ω/k interact with particles having velocities $v_z \sim \omega/k$, such that the wave deposits energy into particles with $v_z < \omega/k$ (like a slightly slower surfer[8] extracting energy from a slightly faster ocean wave) while extracting energy from particles with $v_z > \omega/k$. Since in an equilibrium plasma there will be a larger number of particles with slower velocities than faster velocities, the wave will lose a net energy to the particles and damp out. With a "bump-on-tail" the number of particles with higher v_z than ω/k will outnumber those with lower v_z and thus the wave will extract energy from the particles and grow as a function of time — that is the physics of bump-on-tail instability in a nutshell.

Read *Brodin* [AJP, 65, 66; 1997/PRL, **78**, 1263; 1997] and *Vekstein* [AJP, 88, 886; 1998][9] for more insight on these matters that have proven to be quite taxing on generations of plasma physics students and teachers — Landau damping is a non-trivial phenomenon with many subtleties[10].

[8] Note that a stationary surfer will just bob up and down but will not extract forward momentum from a passing wave.

[9] This is a beautiful paper which shows that Landau damping and wind generation of water waves share some common mechanisms. Section IV presents a very straightforward and readable analysis of wave-particle interactions that take place in Landau damping. Understanding how winds generate waves on ocean surfaces via resonant interactions reminiscent of bump-on-tail instability is a bonus!

[10] For instance, waves damp out but particles have a net energy gain and so there is no overall energy

Relaxing the long wavelength approximation - exact calculation!

Let us evaluate

$$\sigma = \frac{i\epsilon_o\omega_p^2}{k}\int_{-\infty}^{\infty}\frac{v_z G_o^{'}(v_z)}{v_z-\frac{\omega}{k}}dv_z \tag{17.35}$$

directly for a Maxwellian plasma without making use of the long wavelength approximation. We will also avoid using the Landau contour and therefore will end up confirming its validity independently using pure calculus (no complex plane magic!).

Here we take

$$G_o(v_z) = \frac{e^{-v_z^2/2C^2}}{\sqrt{2\pi}C} \tag{17.36}$$

and we have reverted back to physics notation using i in place of $-j$.

Given the even symmetry of $G_o(v_z)$ we note that

$$\int_{-\infty}^{\infty}\frac{v_z G_o^{'}(v_z)}{v_z-\frac{\omega}{k}}dv_z = \frac{\omega}{k}\int_{-\infty}^{\infty}\frac{G_o^{'}(v_z)}{v_z-\frac{\omega}{k}}dv_z = -\frac{\omega}{kC^2}\int_{-\infty}^{\infty}\frac{v_z G_o(v_z)}{v_z-\frac{\omega}{k}}dv_z \tag{17.37}$$

$$= -\frac{\omega}{kC^2}(1+\frac{\omega}{k}\int_{-\infty}^{\infty}\frac{G_o(v_z)}{v_z-\frac{\omega}{k}}dv_z) \tag{17.38}$$

so that

$$\sigma = -\frac{i\epsilon_o\omega_p^2\omega}{k^2C^2}(1+\frac{\omega}{k}\int_{-\infty}^{\infty}\frac{G_o(v_z)}{v_z-\frac{\omega}{k}}dv_z). \tag{17.39}$$

Let us next assume ω to be a complex frequency in the upper-half-plane — the ROC of the Laplace transform — so that

$$\int_0^{\infty}e^{-ik(v_z-\frac{\omega}{k})\tau}d\tau = \frac{1}{ik(v_z-\frac{\omega}{k})} \;\Rightarrow\; \frac{1}{v_z-\frac{\omega}{k}} = ik\int_0^{\infty}e^{-ik(v_z-\frac{\omega}{k})\tau}d\tau. \tag{17.40}$$

We can now use this result as follows:

$$\int_{-\infty}^{\infty}\frac{G_o(v_z)}{v_z-\frac{\omega}{k}}dv_z = \int_{-\infty}^{\infty}G_o(v_z)(ik\int_0^{\infty}e^{-ik(v_z-\frac{\omega}{k})\tau}d\tau)dv_z \tag{17.41}$$

$$= ik\int_0^{\infty}d\tau\, e^{i\omega\tau}\int_{-\infty}^{\infty}dv_z\,\frac{e^{-v_z^2/2C^2}}{\sqrt{2\pi}C}e^{-ikv_z\tau} \tag{17.42}$$

$$= ik\int_0^{\infty}d\tau\, e^{i\omega\tau}e^{-\frac{1}{2}k^2C^2\tau^2}, \tag{17.43}$$

loss or *entropy increase* in the system. *Landau growth* can ultimately follow Landau damping as the slope of $f(v_z)$ changes from negative to positive — but non-linear effects come into picture as well as some level of collisions so that waves never grow back, in practice, to their original amplitudes.

where we also made use of the known FT of a Gaussian function[11]. Substituting this result in (17.39) we obtain

$$\sigma = -\frac{i\epsilon_o\omega_p^2\omega}{k^2C^2}(1+\frac{\omega}{k}ik\int_0^\infty d\tau\, e^{i\omega\tau}e^{-\frac{1}{2}k^2C^2\tau^2}) \tag{17.44}$$

$$= \frac{\omega\epsilon_o\omega_p^2}{k^2C^2}(-i+\omega\int_0^\infty d\tau\, e^{i\omega\tau}e^{-\frac{1}{2}k^2C^2\tau^2}) \tag{17.45}$$

$$= \frac{\omega\epsilon_o\omega_p^2}{k^2C^2}(j+\omega\int_0^\infty d\tau\, e^{-j\omega\tau}e^{-\frac{1}{2}k^2C^2\tau^2}) \tag{17.46}$$

where in the last line we have reverted back to ECE notation.

Now we define a *Gordeyev integral*, a one sided Fourier transform (or simply a Laplace transform, if you are a physicist and prefer $-i\omega$ to s ;-),

$$J(\omega) \equiv \int_0^\infty d\tau\, e^{-j\omega\tau}e^{-\frac{1}{2}k^2C^2\tau^2} \tag{17.47}$$

and write down the *normalized* electron conductivity as

$$\frac{\sigma}{j\omega\epsilon_o} = \frac{1-j\omega J(\omega)}{k^2h^2} \tag{17.48}$$

where $h = C/\omega_p$ is the electron Debye length. This is a standard form used in many contexts in kinetic instability and ISR theories when a generalization of Gordeyev integrals is learned and employed.

[11] **Your HW#1 is re-deriving (17.41) under the assumption of a *lower-half-plane* ω and show the result to be less than (17.41) by $2\pi iG_o(\omega/k)$.** This shows "integrating first (within the ROC, upper-half-plane) and evaluating second" does matter by as much as $2\pi iG_o(\omega/k)$**,** and gets you to the causal solution of Landau's, as opposed to an *a*causal solution in conflict with experiments.

More generally, the causal and *analytic* solution for σ, which is "known" within the ROC of $f_1(\omega)$, is *continued* to the rest of the ω-plane by this "integrate first, evaluate later" strategy described here.

This type of "analytic continuation" of an analytic function — a complex-valued $f(z)$ of a complex variable z that is *complex-differentiable* within some domain of complex plane — to the entire complex plane (except the points where the function is singular) is *always* possible as a consequence of the very rigid inner structure of analytic functions imposed by Cauchy-Reimann conditions [see *Byron and Fuller*, "Mathematics of Classical and Quantum Physics", 1970; *Kudeki and Munson*, "Analog Signals and Systems", 2011]. Knowing an analytic function anywhere on the complex plane is sufficient to know it over the entire plane! Landau effectively taught us where to start on the complex ω-plane to reach the "right" σ function needed over the entire ω plane.

As for which σ is the right one, and which one the wrong one, there is an easy test (just in case it is getting confusing): If we picked the wrong σ (by mistake) it would lead to a negative real part of σ for a Maxwellian plasma, implying wave growth in thermal equilibrium, which is at odds with being in equilibrium — you would drop that σ and use the other one!

A useful identity for the computation of (17.47) **to be proven as HW#2** is[12]

$$\int_0^\infty dt\, e^{-j\theta t} e^{-\frac{1}{4}t^2} = \sqrt{\pi} e^{-\theta^2} - j2e^{-\theta^2} \int_0^\theta e^{x^2} dx. \tag{17.49}$$

Since

$$\omega \int_0^\infty d\tau\, e^{-j\omega\tau} e^{-\frac{1}{2}k^2C^2\tau^2} = \theta \int_0^\infty dt\, e^{-j\theta t} e^{-\frac{1}{4}t^2} \tag{17.50}$$

with $\theta = \omega/\sqrt{2}kC$, it follows that

$$\sigma = \frac{\omega\epsilon_o}{k^2h^2}[\sqrt{\pi}\theta e^{-\theta^2} + j(1 - 2\theta e^{-\theta^2} \int_0^\theta e^{x^2} dx)]. \tag{17.51}$$

The first (real) term in (17.51) is the Landau term in

$$\sigma = \frac{\epsilon_o\omega_p^2}{kC^2}[\pi\frac{\omega^2}{k^2}G_o(\frac{\omega}{k}) + j\mathcal{P}\int_{-\infty}^\infty \frac{v_z^2 G_o(v_z)}{v_z - \frac{\omega}{k}} dv_z] \tag{17.52}$$

that was obtained using the Landau contour earlier on (**compare and verify as HW**#3). The imaginary term in (17.51) is evidently an exact evaluation of the Cauchy principal value integral in (17.52) (which was only approximated before for $\lambda \gg h$) — use "Dawson integral" $F(\theta) = e^{-\theta^2}\int_0^\theta e^{x^2}dx$ defined in MATLAB and *Mathematica* in terms of error function with complex argument, Erfi (see `http://mathworld.wolfram.com/DawsonsIntegral.html`) to compute this term with no approximation (**compare and verify** (17.52) **as HW**#4).

In contrasting the hot plasma conductivity expression (17.52) with the purely imaginary collisionless and cold plasma conductivity expression (17.1), you should realize that the real component of (17.52) is vanishingly small at the wave speeds ω/k of TEM waves on the order of the speed of light $c \gg C = \sqrt{KT/m}$, the electron thermal speed (**show that second term in** (17.52) **reduces to** (17.1) **for** $\omega/k \gg C$ **as HW**#5) — this is why studies of TEM wave propagation in plasmas could be carried out with the simple cold plasma model for the plasma conductivity.

Generalization of Gordeyev integral

In

$$J(\omega) \equiv \int_0^\infty d\tau\, e^{-j\omega\tau} e^{-\frac{1}{2}k^2C^2\tau^2} \tag{17.53}$$

the kernel

$$e^{-\frac{1}{2}k^2C^2\tau^2} \text{ happens to match } \langle e^{-jk\Delta r}\rangle = e^{-\frac{1}{2}k^2\langle\Delta r^2\rangle} \tag{17.54}$$

[12]The proof of this identity consists of showing that $P(\theta) \equiv \int_0^\infty dt\, e^{-j\theta t} e^{-\frac{1}{4}t^2}$ satisfies a differential equation $\frac{dP}{d\theta} + 2\theta P(\theta) = -j2$ subject to initial condition $P(0) = \sqrt{\pi}$. Applying an integrating factor e^{θ^2} to the differential equation, and integrating the resulting expression between 0 and θ, RHS of (17.49) is obtained.

where Δr is an electron displacement along a wavevector $\mathbf{k}$ with magnitude k over a time interval τ within a non-collisional and non-magnetized Maxwellian plasma and$\langle\cdot\rangle$ indicates an expected value operation.

To check this claim, we recognize that in such a plasma electrons (and ions) move along straight line trajectories with Maxwellian distributed velocities v. Therefore

$$\Delta r = v\tau \;\text{ and }\; \langle \Delta r^2 \rangle = \langle v^2\tau^2 \rangle = \langle v^2 \rangle \tau^2 = C^2\tau^2. \tag{17.55}$$

The use of this displacement variance in (17.54) — a kernel known as *single particle ACF* — yields the integrand of transformation (17.53). In collisional and magnetized plasmas the displacement variance will be different and consequently the Gordeyev integral as well as conductivity (17.48) are also different. The knowledge of particle variances in such plasmas are sufficient to compute the conductivities needed in such plasmas using the generalized Gordeyev formula written as

$$J(\omega) \equiv \int_0^\infty d\tau\, e^{-j\omega\tau} \langle e^{-jk\Delta r} \rangle \tag{17.56}$$

so long as Δr is a Gaussian random variable and (17.54) holds. If Δr is not a Gaussian random variable, (17.56) still holds but the single particle ACF $\langle e^{-jk\Delta r} \rangle$ needs to be evaluated using the pdf of Δr directly.

Read *Kudeki and Milla* [2011] and *Milla and Kudeki* [2011] to understand the justification and details of this new approach to ISR and plasma instability models.

Kinetic theory of Farley-Buneman instability

Within the partially ionized and magnetized plasma of the E-region ionosphere at equatorial and auroral latitudes there exists so-called "electrojets" wherein the electrons stream through the background ions at speeds exceeding the sound speed. The result is a kinetic *modified two-stream instability* first described, independently, by Don Farley (Cornell) and Oscar Buneman (Stanford) in the early 1960's. Given our intro to plasma kinetic theory above, we will be able to provide a brief description of the Farley-Buneman instability in this section. The phenomenon is similar to the bump-on-tail instability we mentioned earlier, but still different because there are no bumps in this case, just displaced Maxwellian's for the electrons and ions [see, e.g., *Farley*, JATP, 47, 729; 1985].

The general electrostatic wave dispersion relation is

$$\epsilon_r = 1 + \frac{\sigma_e + \sigma_i}{j\omega\epsilon_o} = 0 \tag{17.57}$$

We will obtain $\sigma_{e,i}$ for this equation by using the appropriate Gordeyev integrals for electron and ion populations within their own reference frames.

In a magnetized plasma

$$\langle e^{j\mathbf{k}\cdot\Delta\mathbf{r}} \rangle = e^{-\frac{1}{2}k_\parallel^2 \langle \Delta r^2 \rangle} \times e^{-\frac{1}{2}k_\perp^2 \langle \Delta p^2 \rangle} \tag{17.58}$$

where $k_\parallel$ and $k_\perp$ are parallel and perpendicular components of wavevector $\mathbf{k}$ with respect to the geomagnetic field $\mathbf{B}_o$ in the background. Also, the parallel and perpendicular variances are

$$\begin{aligned}\langle \Delta r^2 \rangle =& C^2\tau^2 \\ \langle \Delta p^2 \rangle =& \frac{4C^2}{\Omega^2}\sin^2(\Omega\tau/2)\end{aligned} \tag{17.59}$$

where $\Omega = eB_o/m$ is gyro-frequency of the particles [see *Kudeki and Milla*, 2011 — **HW#6: derive (17.59) with minimal glances to Appendix IV in KM11**]. Thus,

$$\langle e^{j\mathbf{k}\cdot\Delta\mathbf{r}} \rangle = e^{-\frac{1}{2}k_\parallel^2 C^2\tau^2}\; e^{-\frac{2k_\perp^2 C^2}{\Omega^2}\sin^2(\Omega\tau/2)} \equiv \langle e^{j\mathbf{k}\cdot\Delta\mathbf{r}} \rangle_m. \tag{17.60}$$

When charged particles in a magnetized plasma go through binary collisions with neutrals at a rate ν, it is found that

$$\langle e^{j\mathbf{k}\cdot\Delta\mathbf{r}} \rangle \;=\; \sum_{n=0}^{\infty} \nu^n \underbrace{g(\tau) * \cdots * g(\tau)}_{\substack{n+1 \text{ terms, } n \text{ convolutions}}} \tag{17.61}$$

with

$$g(\tau) = u(\tau)e^{-\nu\tau}\,\langle e^{j\mathbf{k}\cdot\Delta\mathbf{r}} \rangle_m \;\;\leftrightarrow\;\; G(\omega) = \int e^{-j\omega\tau} g(\tau) d\tau = \int_0^\infty e^{-j\omega\tau} e^{-\nu\tau}\,\langle e^{j\mathbf{k}\cdot\Delta\mathbf{r}} \rangle_m d\tau. \tag{17.62}$$

Consequently, the Gordeyev integral, one-sided Fourier transform of (17.61), can be computed as

$$J(\omega) = \frac{G(\omega)}{1 - \nu G(\omega)} \tag{17.63}$$

as shown in *Milla and Kudeki* [2009].

For the dispersion relation (17.57) we need *normalized admittance* $y \equiv \sigma/j\omega\epsilon_o$ given by (17.48) computed for electron and ions. If the calculation is to be carried out in the ion frame (the usual practice since ion frame is typically the Earth frame where measurements are conducted using an Earth fixed radar), then the electron Gordeyev integral needs to be Doppler shifted from the electron to the ion frame via the usage of

$$\omega' \equiv \omega - \mathbf{k}\cdot\mathbf{V} = \omega - k_\perp V \tag{17.64}$$

in the expression for $\sigma_e/j\omega\epsilon_o$ given by (17.48) — here $\mathbf{V}$ is the mean velocity vector of the electrons in the ion frame with a magnitude $V = |\mathbf{V}|$ and a direction orthogonal to the ambient magnetic field $\mathbf{B}_o$, while $\mathbf{k}$ is assumed to lie on the $\mathbf{VB_o}$-plane making a small *magnetic aspect angle* $k_\parallel/k \ll 1$. The dispersion relation in that case becomes

$$\epsilon_r(k, k_\perp, \omega, \omega') = 1 + \frac{1 - j\omega' J_e(\omega')}{k^2 h_e^2} + \frac{1 - j\omega J_i(\omega)}{k^2 h_i^2} = 0 \tag{17.65}$$

with

$$J_{e,i}(\omega) = \frac{G_{e,i}(\omega)}{1-\nu_{e,i}G_{e,i}(\omega)} \text{ and } G_{e,i}(\omega) = \int_0^\infty e^{-j\omega\tau} e^{-\nu_{e,i}\tau}\ e^{-\frac{1}{2}k_\parallel^2 C_{e,i}^2\tau^2}\ e^{-\frac{2k_\perp^2 C_{e,i}^2}{\Omega_{e,i}^2}\sin^2(\Omega_{e,i}\tau/2)} d\tau. \tag{17.66}$$

Instability analysis based on dispersion relation (17.65) can be conducted as follows: Make 3D surface plots of $|\epsilon_r(k,k_\perp,\omega,\omega')|^{-2}$ as a function of real ω and ω', with real ω and ω' implying *marginal stability* at the solution point[13] corresponding to the peak of $|\epsilon_r(k,k_\perp,\omega,\omega')|^{-2}$ surface — equivalent to a zero of $\epsilon_r(k,k_\perp,\omega,\omega')$, a root of the dispersion relation. The threshold velocity V_{th} for instability is then obtained from $k_\perp V_{th} = \omega - \omega'$ computed with ω and ω' corresponding to the peak of $|\epsilon_r(k,k_\perp,\omega,\omega')|^{-2}$. For $V > V_{th}$ the plasma is unstable. Also the phase velocity of marginally stable waves in the ion frame is simply ω/k.

Analysis shows that roughly $V_{th} \sim \omega/k \sim C_s \equiv \sqrt{K(T_e+T_i)/m_i}$, which is the so-called *ion-acoustic speed*. In summary, Farley-Buneman instability produces *magnetic field aligned* ($k_\parallel/k \ll 1$) ion-acoustic waves when the electrons stream through the background ions with supersonic speeds — $V > C_s$ — across the magnetic field lines (driven by $\mathbf{E}_o \times \mathbf{B}_o$ motion at E-region altitudes). Exact numbers depend on the magnetic aspect angle $k_\parallel/k$ as well as kh_e. If $kh_e \ll 1$, this kinetic description can be replaced by a simpler fluid description, but at short wavelengths $kh_e \gtrsim 1$ the fluid description fails. For a first detailed kinetic description see *Farley* [JGR, **68**, 6083; 1963].

Here is a ***python script for two-stream instability analysis*** including Gordeyev integral code that can also be used for plotting theoretical ISR spectra discussed in the next section. **Your HW#7 is to download the script and "experiment" with it. Use the code to find the threshold velocity for some aspect angle. Also for larger than threshold velocity show that the roots of the dispersion relation are located in the LHP.**

More details + ISR theory - highlights from KM & MK papers

While the previous sections emphasized a kinetic theory based approach for obtaining $\sigma_{e,i}$, we will outline here an alternate approach based on the *fluctuation-dissipation theorem*. The theorem is a generalization of the *Nyquist noise theorem* that stipulates that the open-circuit voltage $v(t)$ and short-circuit current $i(t)$ of an electrical circuit element with an input impedance $Z(\omega) = Y^{-1}(\omega)$ will exhibit thermally driven fluctuations having the two-sided power spectra $\langle|V(\omega)|^2\rangle$ and $\langle|I(\omega)|^2\rangle$, respectively, with associated

[13] We know that for $\mathbf{V} = 0$ the plasma is stable and thus $\epsilon_r(k,k_\perp,\omega,\omega')$ has zeroes — dispersion relation solutions — only in the upper-half $\omega = \omega_r + j\omega_i$ plane because $e^{j\omega t} = e^{j\omega_r t}e^{-\omega_i t}$ is damped (stable) *iff* $\omega_i > 0$. As $V \equiv |\mathbf{V}|$ is increased these zeroes will migrate out of upper-half-plane into the unstable lower-half-plane by crossing the real ω axis. Value of V at which $\epsilon_r(k,k_\perp,\omega,\omega')$ has a real valued zero in ω-plane is therefore the threshold velocity V_{th} of the instability for the assumed aspect angle $k_\parallel/k$.

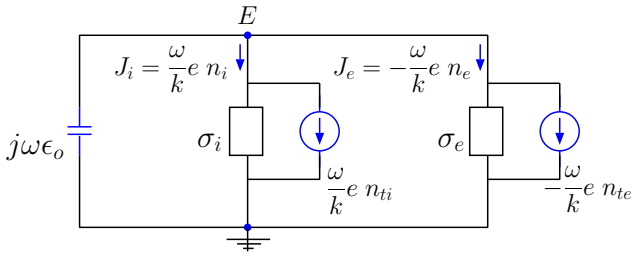


Figure 17.2: Equivalent circuit of electrostatic plasma wave system.

variances

$$\langle |v(t)|^2 \rangle = \int_{-\infty}^{\infty} \langle |V(\omega)|^2 \rangle \frac{d\omega}{2\pi} \quad \text{and} \quad \langle |i(t)|^2 \rangle = \int_{-\infty}^{\infty} \langle |I(\omega)|^2 \rangle \frac{d\omega}{2\pi}, \tag{17.67}$$

such that

$$\langle |V(\omega)|^2 \rangle = 2KT\mathrm{Re}\{Z(\omega)\} \quad \text{and} \quad \langle |I(\omega)|^2 \rangle = 2KT\mathrm{Re}\{Y(\omega)\} \tag{17.68}$$

for frequencies $f = \frac{\omega}{2\pi}$ satisfying $h|f| \ll KT$, where T is the equilibrium temperature of the element while h and K are Planck and Boltzmann constants, respectively. Fluctuation-dissipation theorem *in turn* stipulates that these results can be applied to the equivalent circuit of any linear dissipative system in thermal equilibrium in order model the pertinent fluctuation spectra in analogy with $\langle |V(\omega)|^2 \rangle$ and $\langle |I(\omega)|^2 \rangle$.

Now, the *equivalent circuit model* of a two component plasma sustaining electrostatic electric field fluctuations $\mathbf{E}$ can be obtained from Ampere's law (in plane wave form and using ECE notation)

$$-j\mathbf{k} \times \mathbf{H} = \mathbf{J} + j\omega\epsilon_o \mathbf{E}. \tag{17.69}$$

The component of this vector equation in the direction of a wavevector $\mathbf{k}$, can be written as

$$0 = (\sigma_i + \sigma_e)E + \frac{\omega}{k} e(n_{ti} - n_{te}) + j\omega\epsilon_o E \tag{17.70}$$

in terms of conductivities $\sigma_{e,i}$ as well thermally *impressed* currents[14] $\frac{\omega}{k} e n_{ti}$ and $-\frac{\omega}{k} e n_{te}$ obtained from the continuity equation $j\omega\rho_t - jkJ_t = 0$ applied for each species with $\rho_{te,i} = \mp e n_{te,i}$.

Note that (17.70) has the form of Kirchhoff's current law (KCL) applied at the top node of a lumped-element equivalent circuit shown in Figure 17.2 having the branch "admittances" $\sigma_{e,i}$ and a branch "capacitance" ϵ_o, as well as impressed current sources $\propto n_{te,ti}$ representing independent thermal drivers within the plasma. The branch currents $\frac{\omega}{k} e n_i$ and $-\frac{\omega}{k} e n_e$ indicated on the circuit diagram refer to the total ion and electron current density components J_e and J_i, including the macroscopic currents driven by E.

Let's apply the fluctuation-dissipation theorem to the branches of the equivalent circuit in Figure 17.2. Specifically, in terms of short circuit current spectra $\frac{\omega^2}{k^2} e^2 \langle |n_{te,i}(\mathbf{k}, \omega)|^2 \rangle$

[14] In analogy with noise generators inserted in parallel with hot resistors in circuit noise calculations.

of branches having admittances $\sigma_{e,i}$, the Nyquist formula

$$\langle |I(\omega)|^2 \rangle = 2KT\mathrm{Re}\{Y(\omega)\} \tag{17.71}$$

becomes

$$\frac{\omega^2}{k^2} e^2 \langle |n_{te,i}(\mathbf{k},\omega)|^2 \rangle = 2KT_{e,i}\mathrm{Re}\{\sigma_{e,i}(\mathbf{k},\omega)\}, \tag{17.72}$$

from which

$$\mathrm{Re}\{\sigma_{e,i}(\mathbf{k},\omega)\} = \frac{1}{2KT_{e,i}} \frac{\omega^2}{k^2} e^2 \langle |n_{te,i}(\mathbf{k},\omega)|^2 \rangle. \tag{17.73}$$

The imaginary part of $\sigma_{e,i}$ can be deduced from the real part (17.73) by virtue of the causality of the system (see discussion below about Hilbert transform). Therefore (17.73) will provide us with complex $\sigma_{e,i}$ once thermally driven spectra $\langle |n_{te,i}(\mathbf{k},\omega)|^2 \rangle$ are properly modeled.

To model $\langle |n_{te}(\mathbf{k},\omega)|^2 \rangle$ we first note that the *microscopic* density function of the electrons and its spatial Fourier transform can be expressed as

$$n_e(\mathbf{r},t) = \sum_{p=1}^{N_o \Delta V} \delta(\mathbf{r} - \mathbf{r}_p(t)) \quad \leftrightarrow \quad n_e(\mathbf{k},t) = \int d\mathbf{r}\, n_e(\mathbf{r},t)\, e^{j\mathbf{k}\cdot\mathbf{r}} = \sum_{p=1}^{N_o \Delta V} e^{j\mathbf{k}\cdot\mathbf{r}_p(t)} \tag{17.74}$$

across a volume of size ΔV having an average density of N_o electrons per meter cube. The space-time spectrum of $n_e(\mathbf{k},t)$ is defined to be time Fourier transform of its ACF as specified in

$$\langle |n_e(\mathbf{k},\omega)|^2 \rangle \equiv \int d\tau e^{-j\omega\tau} \frac{1}{\Delta V} \langle n_e^*(\mathbf{k},t) n_e(\mathbf{k},t+\tau) \rangle. \tag{17.75}$$

Using (17.74), the spectrum (17.75) can be expressed as

$$\begin{aligned} \langle |n_e(\mathbf{k},\omega)|^2 \rangle &= \int d\tau e^{-j\omega\tau} \frac{1}{\Delta V} \langle \sum_{p=1}^{N_o \Delta V} e^{-j\mathbf{k}\cdot\mathbf{r}_p(t)} \sum_{p=1}^{N_o \Delta V} e^{j\mathbf{k}\cdot\mathbf{r}_p(t+\tau)} \rangle \\ &= \int d\tau e^{-j\omega\tau} \frac{1}{\Delta V} \sum_{p=1}^{N_o \Delta V} \sum_{q=1}^{N_o \Delta V} \langle e^{-j\mathbf{k}\cdot\mathbf{r}_p(t)} e^{j\mathbf{k}\cdot\mathbf{r}_q(t+\tau)} \rangle. \end{aligned} \tag{17.76}$$

Individual electrons in ΔV follow independent random trajectories due to random thermal motions so that the expected value within (17.76) is zero except for when $q = p$. With that stipulation the thermal density spectrum of the electrons is

$$\langle |n_{te}(\mathbf{k},\omega)|^2 \rangle = N_o \int d\tau e^{-j\omega\tau} \langle e^{j\mathbf{k}\cdot\Delta\mathbf{r}} \rangle, \tag{17.77}$$

where

$$\Delta\mathbf{r} \equiv \mathbf{r}_p(t+\tau) - \mathbf{r}_p(t) \tag{17.78}$$

denotes the random displacement vector of the electrons over time lags τ along their independent trajectories $\mathbf{r}_p(t)$.

Now, the use of (17.77) in (17.73) for electrons yields

$$\mathrm{Re}\{\sigma_e(\mathbf{k},\omega)\} = \frac{N_o e^2}{2KT_e}\frac{\omega^2}{k^2}\int_{-\infty}^{\infty} d\tau e^{-j\omega\tau}\langle e^{j\mathbf{k}\cdot\Delta\mathbf{r}}\rangle. \tag{17.79}$$

Consequently,

$$\frac{\mathrm{Re}\{\sigma_e(\mathbf{k},\omega)\}}{j\omega\epsilon_o} = -\frac{j\omega}{2k^2h_e^2}\int_{-\infty}^{\infty} d\tau e^{-j\omega\tau}\langle e^{j\mathbf{k}\cdot\Delta\mathbf{r}}\rangle. \tag{17.80}$$

where $h_e^2 = C_e^2/\omega_e^2$, $C_e = \sqrt{KT_e/m_e}$, and $\omega_e = \sqrt{N_o e^2/m_e\epsilon_o}$. With *Gordeyev integral* defined as

$$J_e(\omega) = \int_0^{\infty} d\tau e^{-j\omega\tau}\langle e^{j\mathbf{k}\cdot\Delta\mathbf{r}}\rangle \tag{17.81}$$

we see that

$$\frac{\sigma_R}{j\omega\epsilon_o} = -\frac{j\omega}{k^2h_e^2}J_R, \tag{17.82}$$

with σ_R and J_R referring to real parts of conductivity $\sigma(\mathbf{k},\omega)$ and the Gordeyev integral $J_e(\omega)$, respectively. Using a Hilbert transform procedure outlined in Appendix I and II of *Kudeki and Milla* [2011], the complex conductivities $\sigma_{e,i}$ can be expressed as (and we already know this from our earlier kinetic theory work)

$$\frac{\sigma_e}{j\omega\epsilon_o} = \frac{1-j\omega J_e(\omega)}{k^2h_e^2} \text{ and } \frac{\sigma_i}{j\omega\epsilon_o} = \frac{1-j\omega J_i(\omega)}{k^2h_i^2}. \tag{17.83}$$

Consequently

$$1+\frac{\sigma_e+\sigma_i}{j\omega\epsilon_o} = 0 \;\Rightarrow\; 1+\frac{1-j\omega J_e(\omega)}{k^2h_e^2}+\frac{1-j\omega J_i(\omega)}{k^2h_i^2} = 0. \tag{17.84}$$

As a final note, let us return to the equivalent circuit of electrostatic waves shown in Figure 17.2. Solving the circuit problem for "voltage" E, and then using E in the "electron current" expression $-\frac{\omega}{k}en_e = E\sigma_e - \frac{\omega}{k}en_{te}$, we find that the electron density fluctuation amplitude is

$$n_e(\mathbf{k},\omega) = \frac{(j\omega\epsilon_o+\sigma_i)n_{te}(\mathbf{k},\omega)+\sigma_e n_{ti}(\mathbf{k},\omega)}{j\omega\epsilon_o+\sigma_e+\sigma_i}. \tag{17.85}$$

The roots ω of the denominator in (17.85) and solutions of the dispersion relation (17.84) are one and the same — that means that the normal mode frequencies of electrostatic electron density waves in the two component plasma coincide with the poles of the Laplace transform $n_e(\mathbf{k},\omega)$. Also (17.85) implies that the electron density spectrum $\langle|n_e(\mathbf{k},\omega)|^2\rangle$ needed in ISR spectral models must be weighted superpositions of $\langle|n_{te}(\mathbf{k},\omega)|^2\rangle$ *and* $\langle|n_{ti}(\mathbf{k},\omega)|^2\rangle$ — not simply $\langle|n_e(\mathbf{k},\omega)|^2\rangle = \langle|n_{te}(\mathbf{k},\omega)|^2\rangle$ because of *collective effects* caused by E — which we already know how to compute via (17.77) and single particle ACF's $\langle e^{j\mathbf{k}\cdot\Delta\mathbf{r}}\rangle$, but

$$\langle|n_e(\mathbf{k},\omega)|^2\rangle = \frac{|j\omega\epsilon_o+\sigma_i|^2\langle|n_{te}(\mathbf{k},\omega)|^2\rangle+|\sigma_e|^2\langle|n_{ti}(\mathbf{k},\omega)|^2\rangle}{|j\omega\epsilon_o+\sigma_e+\sigma_i|^2}. \tag{17.86}$$

18 Appendix XI: Incoherent scatter theory — some aspects outdated, see KM and MK papers

Electrostatic wave spectrum

Let $n_{te}(\omega, \mathbf{k})$ and $n_{ti}(\omega, \mathbf{k})$ denote the complex amplitudes of electron and ion number density waves $\sim e^{j(\omega t - \mathbf{k}\cdot\mathbf{r})}$ generated in a hypothetical plasma of non-interacting particles (meaning that electrostatic interactions are assumed to be absent) with random thermal motions. Density waves $n_{e,i}(\omega, \mathbf{k})$ in a real plasma where electrostatic interactions do take place can then be expressed in terms of $n_{te,i}(\omega, \mathbf{k})$ as follows[1]:

An imbalance between independent random variables $n_{te}(\omega, \mathbf{k})$ and $n_{ti}(\omega, \mathbf{k})$ produces a space charge density $e(n_{ti}(\omega, \mathbf{k}) - n_{te}(\omega, \mathbf{k}))$ which in turn produces a fluctuating electric field $E(\omega, \mathbf{k})$ (in the direction of $\mathbf{k}$) satisfying the Poisson's equation

$$-jk\epsilon(\omega, \mathbf{k})E(\omega, \mathbf{k}) = -jk\epsilon_o(1 + \frac{\sigma(\omega, \mathbf{k})}{j\omega\epsilon_o})E(\omega, \mathbf{k}) = e(n_{ti}(\omega, \mathbf{k}) - n_{te}(\omega, \mathbf{k})) \tag{18.1}$$

written in plane wave form. Thus

$$E(\omega, \mathbf{k}) = \frac{e(n_{ti}(\omega, \mathbf{k}) - n_{te}(\omega, \mathbf{k}))}{-jk\epsilon_o - \frac{k}{\omega}\sigma(\omega, \mathbf{k})}, \tag{18.2}$$

which in turn drives divergent electron and ion currents $\sigma_{e,i}(\omega, \mathbf{k})E(\omega, \mathbf{k})$ giving rise to polarization charge densities $\delta\rho_{e,i}(\omega, \mathbf{k})$ related by the continuity equation

$$j\omega\delta\rho_{e,i}(\omega, \mathbf{k}) - jk\sigma_{e,i}(\omega, \mathbf{k})E(\omega, \mathbf{k}) = 0 \tag{18.3}$$

also written in plane wave from. Hence, for example,

$$\delta\rho_e(\omega, \mathbf{k}) = \frac{k}{\omega}\sigma_e(\omega, \mathbf{k})E(\omega, \mathbf{k}) = -\sigma_e(\omega, \mathbf{k})\frac{e(n_{ti}(\omega, \mathbf{k}) - n_{te}(\omega, \mathbf{k}))}{j\omega\epsilon_o + \sigma(\omega, \mathbf{k})} \equiv -e\delta n_e(\omega, \mathbf{k}), \tag{18.4}$$

and, therefore, the total electron number density wave (i.e., correcting the unrealistic scenario of non-interacting particles)

$$\begin{aligned} n_e(\omega, \mathbf{k}) &= n_{te}(\omega, \mathbf{k}) + \delta n_e(\omega, \mathbf{k}) = n_{te}(\omega, \mathbf{k}) + \sigma_e(\omega, \mathbf{k})\frac{(n_{ti}(\omega, \mathbf{k}) - n_{te}(\omega, \mathbf{k}))}{j\omega\epsilon_o + \sigma(\omega, \mathbf{k})} \\ &= \frac{(j\omega\epsilon_o + \sigma_i(\omega, \mathbf{k}))n_{te}(\omega, \mathbf{k}) + \sigma_e(\omega, \mathbf{k})n_{ti}(\omega, \mathbf{k})}{j\omega\epsilon_o + \sigma_e(\omega, \mathbf{k}) + \sigma_i(\omega, \mathbf{k})}. \end{aligned} \tag{18.5}$$

[1] We follow here a derivation given by *Fejer* [1985].

The total ion density wave $n_i(\omega, \mathbf{k})$ can also be expressed in a similar way.

Finally, we notice that in view of statistical independence of random variables $n_{te}(\omega, \mathbf{k})$ and $n_{ti}(\omega, \mathbf{k})$, linear relation (18.5) between $n_e(\omega, \mathbf{k})$ and $n_{te,i}(\omega, \mathbf{k})$ implies a total electron density spectrum

$$\langle |n_e(\omega, \mathbf{k})|^2 \rangle = \frac{|j\omega\epsilon_o + \sigma_i(\omega, \mathbf{k})|^2 \langle |n_{te}(\omega, \mathbf{k})|^2 \rangle + |\sigma_e(\omega, \mathbf{k})|^2 \langle |n_{ti}(\omega, \mathbf{k})|^2 \rangle}{|j\omega\epsilon_o + \sigma_e(\omega, \mathbf{k}) + \sigma_i(\omega, \mathbf{k})|^2} \quad (18.6)$$

$$= \frac{|jk^2\lambda_e^2 + \mu y(\theta_i)|^2 \langle |n_{te}(\omega, \mathbf{k})|^2 \rangle + |y(\theta_e)|^2 \langle |n_{ti}(\omega, \mathbf{k})|^2 \rangle}{|jk^2\lambda_e^2 + y(\theta_e) + \mu y(\theta_i)|^2}, \quad (18.7)$$

where $\mu \equiv \frac{T_e}{T_i}$ and $y(\theta_s) \equiv j + \theta_s J(\theta_s)$. We next have the task of determining the density spectra $\langle |n_{te,i}(\omega, \mathbf{k})|^2 \rangle$ expected for non-interacting particles to use in (18.7).

The number density function of a collection of non-interacting free electrons with individual trajectories $\mathbf{r}_p(t)$ can be expressed as

$$n_{te}(t, \mathbf{r}) = \sum_{p=1}^{P} \delta(\mathbf{r} - \mathbf{r}_p(t)), \quad (18.8)$$

where P is the number electrons in the collection and the individual electrons have been idealized as point particles, an acceptable notion when de Broglie wavelength of an average electron is small compared to the mean interparticle distance $\sim (V/P)^{1/3}$ of P particles spread out over a volume V. Now, the spatial Fourier transform of (18.8) over volume V gives

$$n_{te}(t, \mathbf{k}) = \int_V d\mathbf{r} n_{te}(t, \mathbf{r}) e^{j\mathbf{k}\cdot\mathbf{r}} = \sum_{p=1}^{P} e^{j\mathbf{k}\cdot\mathbf{r}_p(t)}, \quad (18.9)$$

and

$$\frac{\langle n_{te}^*(t, \mathbf{k}) n_{te}(t+\tau, \mathbf{k}) \rangle}{V} = \frac{1}{V} \sum_{p=1}^{P} \langle e^{j\mathbf{k}\cdot(\mathbf{r}_p(t+\tau) - \mathbf{r}_p(t))} \rangle = \frac{P}{V} \langle e^{j\mathbf{k}\cdot(\mathbf{r}_p(t+\tau) - \mathbf{r}_p(t))} \rangle \quad (18.10)$$

specifies — in the limit of sufficiently large V — the wavenumber spectrum *and* temporal auto-correlation function (ACF) of density fluctuations (18.8). Note that in the right-hand side of (18.10) we expressed the product of two sums as a single sum, because the premise of this calculation is that individual particles have independent dynamics and the cross-terms average out to zero.

Since

$$\langle e^{j\mathbf{k}\cdot(\mathbf{r}_p(t+\tau) - \mathbf{r}_p(t))} \rangle = \langle e^{j\mathbf{k}\cdot\mathbf{v}_p\tau} \rangle \equiv \langle e^{jkv\tau} \rangle = \int dv G_o(v) e^{jkv\tau} = e^{jk\bar{v}_{ze}\tau} e^{-\frac{C_e^2 k^2 \tau^2}{2}} \quad (18.11)$$

for a collection of electrons with straight line trajectories and a Maxwellian velocity distribution, we re-write (18.10) as

$$\frac{\langle n_{te}^*(t, \mathbf{k}) n_{te}(t+\tau, \mathbf{k}) \rangle}{V} = \frac{P}{V} e^{jk\bar{v}_{ze}\tau} e^{-\frac{C_e^2 k^2 \tau^2}{2}} = N e^{jk\bar{v}_{ze}\tau} e^{-\frac{C_e^2 k^2 \tau^2}{2}}, \quad (18.12)$$

where N is the *mean density* of the particles in the collection over the volume V. Finally, the *frequency spectrum*[2] $\langle|n_{te}(\omega,\mathbf{k})|^2\rangle$ of thermal fluctuations $n_{te}(t,\mathbf{r})$ is simply the Fourier transform of (18.12) over lag τ, that is

$$\begin{aligned}\langle|n_{te}(\omega,\mathbf{k})|^2\rangle &= N\int_{-\infty}^{\infty} d\tau e^{-j\omega\tau}e^{jk\bar{v}_{ze}\tau}e^{-\frac{C_e^2k^2\tau^2}{2}} = \frac{N}{k\sqrt{2}C_e}\int_{-\infty}^{\infty} dt e^{-j\frac{\omega-k\bar{v}_{ze}}{k\sqrt{2}C_e}t}e^{-t^2/4}\\ &= \frac{N}{k\sqrt{2}C_e}2\mathrm{Re}\{J(\theta_e)\} = \frac{2N\theta_e\mathrm{Re}\{J(\theta_e)\}}{\omega-k\bar{v}_{ze}} = \frac{2N\mathrm{Re}\{y(\theta_e)\}}{\omega-k\bar{v}_{ze}},\end{aligned} \tag{18.13}$$

where $\theta_e = \frac{\omega-k\bar{v}_{ze}}{k\sqrt{2}C_e}$. Likewise, we also have

$$\langle|n_{ti}(\omega,\mathbf{k})|^2\rangle = \frac{N}{k\sqrt{2}C_i}2\mathrm{Re}\{J(\theta_i)\} = \frac{2N\theta_i\mathrm{Re}\{J(\theta_i)\}}{\omega-k\bar{v}_{zi}} = \frac{2N\mathrm{Re}\{y(\theta_i)\}}{\omega-k\bar{v}_{zi}}. \tag{18.14}$$

Finally[3], combining (18.7) with (18.13) and (18.14), we find that

$$\begin{aligned}\langle|n_e(\omega,\mathbf{k})|^2\rangle &= 2N\frac{|jk^2\lambda_e^2+\mu y(\theta_i)|^2\frac{\mathrm{Re}\{y(\theta_e)\}}{\omega-k\bar{v}_{ze}}+|y(\theta_e)|^2\frac{\mathrm{Re}\{y(\theta_i)\}}{\omega-k\bar{v}_{zi}}}{|jk^2\lambda_e^2+y(\theta_e)+\mu y(\theta_i)|^2}\\ &= 2N\frac{|j(k^2\lambda_e^2+\mu)+\mu\theta_iJ(\theta_i)|^2\frac{\mathrm{Re}\{J(\theta_e)\}}{k\sqrt{2}C_e}+|j+\theta_eJ(\theta_e)|^2\frac{\mathrm{Re}\{J(\theta_i)\}}{k\sqrt{2}C_i}}{|j(k^2\lambda_e^2+1+\mu)+\theta_eJ(\theta_e)+\mu\theta_iJ(\theta_i)|^2}\end{aligned} \tag{18.15}$$

Note that (18.15) can be evaluated exclusively in terms of electron and ion Gordeyev integrals $J(\theta_{e,i})$.

Radar backscatter from ionospheric density waves

Let

$$E_s = \sum_{p=1}^{P}|E_i|a_pe^{-j2k_or_p} \tag{18.16}$$

denote the back*s*cattered field from a collection of free electrons at radar ranges r_p far away from a radar antenna, where $|E_i|$ is the magnitude of the transmitted radar field incident on the electrons, k_o the propagation constant of radar transmission, $|a_p|^2 \equiv \frac{\sigma_e}{4\pi r_p^2}$,

[2]Note that $\langle|n_{te}(\omega,\mathbf{k})|^2\rangle$ is *not* the expected value of magnitude-square of some Fourier transform $n_{te}(\omega,\mathbf{k})$ — as our notation may misleadingly suggest — given that an ordinary Fourier transform $n_{te}(\omega,\mathbf{k})$ *does not exist* for our random signal $n_{te}(t,\mathbf{r})$ of an infinite duration. Rather, $\langle|n_{te}(\omega,\mathbf{k})|^2\rangle$ — and the same comment also applies to $\langle|n_e(\omega,\mathbf{k})|^2\rangle$ — *denotes* the Fourier transform over lag τ of an ACF function related to $n_{te}(t,\mathbf{r})$ (in this case its *spatial* Fourier transform which does exist because of finite V). The linear relation between $n_e(\omega,\mathbf{k})$ and $n_{te,i}(\omega,\mathbf{k})$, on the other hand, which we used to relate $\langle|n_e(\omega,\mathbf{k})|^2\rangle$ to $\langle|n_{te,i}(\omega,\mathbf{k})|^2\rangle$ is valid and justified in the generalized Fourier-Stieltjes sense — see *Lumley and Panofsky* [1964].

[3]these results given for a plasma with a single ion species can be readily generalized for a multi ion plasma by replacing the $\mu y(\theta_i)$ and $\frac{\mathrm{Re}\{y(\theta_i)\}}{\omega-k\bar{v}_{zi}}$ terms by their weighted sums over the existing ion species. For example, in a two-species plasma with species densities N_1 and N_2 use $\mu y(\theta_i) \to \frac{N_1}{N}\mu_1y(\theta_1)+\frac{N_2}{N}\mu_2y(\theta_2)$ and $\frac{N\mathrm{Re}\{y(\theta_i)\}}{\omega-k\bar{v}_{zi}} \to \frac{N_1\mathrm{Re}\{y(\theta_1)\}}{\omega-k\bar{v}_{z1}}+\frac{N_2\mathrm{Re}\{y(\theta_2)\}}{\omega-k\bar{v}_{z2}}$.

and σ_e the backscatter radar cross section (RCS) of a single electron (not to be confused with conductivity, also denoted using the Greek sigma). *Assuming* non-interacting electrons with independent trajectories $\mathbf{r}_p(t)$ and using (18.10) for $\langle e^{j\mathbf{k}\cdot(\mathbf{r}_p(t+\tau)-\mathbf{r}_p(t))}\rangle$, the ACF of backscattered field (18.16) over a time lag τ can be denoted and modeled as

$$\rho_E(\tau) = \sum_{p=1}^{P} |E_i|^2 |a_p|^2 \langle e^{j\mathbf{k}\cdot(\mathbf{r}_p(t+\tau)-\mathbf{r}_p(t))}\rangle \approx \frac{|E_i|^2}{4\pi r^2} V\sigma_e \frac{\langle n_{te}^*(t,\mathbf{k}) n_{te}(t+\tau,\mathbf{k})\rangle}{V}, \quad (18.17)$$

where $\mathbf{k} \equiv -2k_o\hat{r}$, $\hat{r}$ is a unit vector in radar viewing direction, $k \equiv 2k_o$, and r the average value of all r_p, i.e., the so-called radar range $r \gg V^{-1/3}$. The corresponding frequency spectrum is the Fourier transform of (18.17), denoted as

$$\langle |E_s(\omega)|^2\rangle = \frac{|E_i|^2}{4\pi r^2} V\sigma_e \langle |n_{te}(\omega, -2k_o\hat{r})|^2\rangle. \quad (18.18)$$

The basic conclusion here is that the backscattered-field frequency spectrum $\langle |E_s(\omega)|^2\rangle$ is linearly proportional to the spectrum of electron density fluctuations in the *scattering volume*[4] V evaluated at a wavenumber $2k_o$, known as *Bragg* wavenumber. Clearly, for scattering from a real plasma — where charged particles do interact via electrostatic fields — the *Bragg scatter* result above should be written as

$$\langle |E_s(\omega)|^2\rangle = \frac{|E_i|^2}{4\pi r^2} V\sigma_e \langle |n_e(\omega, -2k_o\hat{r})|^2\rangle, \quad (18.19)$$

where $\langle |n_e(\omega,\mathbf{k})|^2\rangle$ is now the actual density spectrum taking into account the interactions[5].

It is in general convenient to re-write (18.19) as

$$\langle |E_s(\omega)|^2\rangle = \frac{|E_i|^2}{4\pi r^2} V\sigma(\omega) \quad (18.20)$$

and define

$$\langle |E_s|^2\rangle \equiv \int \frac{d\omega}{2\pi} \langle |E_s(\omega)|^2\rangle = \frac{|E_i|^2}{4\pi r^2} V \int \frac{d\omega}{2\pi} \sigma(\omega), \quad (18.21)$$

where

$$\sigma(\omega) \equiv \sigma_e \langle |n_e(\omega, -2k_o\hat{r})|^2\rangle \quad (18.22)$$

is, by definition, the *volumetric* RCS spectrum of density irregularities $n_e(\omega,\mathbf{k})$ within the radar field of view. In backscatter from hot plasmas in thermal equilibrium (Maxwellian

[4] In general, the scattering volume $V \sim r^2\delta\Omega\delta r$ for a pulsed backscatter radar, where $\delta\Omega$ is the solid angle of "two-way beam" of radar antenna and $\delta r = \frac{c\delta t}{2}$ is *range resolution* related to the transmitted pulse duration δt.

[5] This new expression can be verified independently starting from a continuum model for the propagation of a transmitted radar pulse E_i and then using the *first-Born approximation* to calculate the scattered field E_s — see, e.g., *Woodman* [1991].

distributions, unequal temperatures allowed) — known as *incoherent scatter* — the RCS spectrum equals

$$\sigma(\omega) = \sigma_e 2N \frac{|j(k^2\lambda_e^2+\mu)+\mu\theta_i J(\theta_i)|^2 \frac{\mathrm{Re}\{J(\theta_e)\}}{k\sqrt{2}C_e} + |j+\theta_e J(\theta_e)|^2 \frac{\mathrm{Re}\{J(\theta_i)\}}{k\sqrt{2}C_i}}{|j(k^2\lambda_e^2+1+\mu)+\theta_e J(\theta_e)+\mu\theta_i J(\theta_i)|^2}, \tag{18.23}$$

while the total volumetric RCS, defined as

$$\sigma_T \equiv \int \frac{d\omega}{2\pi}\sigma(\omega), \tag{18.24}$$

is found to be $\sim \frac{\sigma_e N}{1+\mu}$ for $k\lambda_e \ll 1$ and μ is the range of 1 to 5 (applies to non-magnetized plasmas only — see, e.g., *Farley* [1971]). In backscatter from an ionosphere *not* in thermal equilibrium, σ_T could be orders of magnitude larger — that case is usually referred to as *coherent scatter*.

Fluctuation-dissipation theorem and Kramers-Kronig relations

We have established that that in collisionless and non-magnetized plasmas, the AC conductivities of charged species are

$$\sigma_{e,i}(\omega) = \frac{\omega\epsilon_o}{k^2\lambda_{e,i}^2}[j+\theta_{e,i}J(\theta_{e,i})], \tag{18.25}$$

while thermally driven density spectra of the same set of carriers in the plasma (ignoring collective interactions via electrostatic fields) are

$$\langle|n_{te,i}(\omega,\mathbf{k})|^2\rangle = \frac{2N_{e,i}\mathrm{Re}\{J(\theta_{e,i})\}}{k\sqrt{2}C_{e,i}}. \tag{18.26}$$

Combining (18.25) and (18.26) for $\bar{v}_{ze,i}=0$ we note that

$$q^2\frac{\omega^2}{k^2}\langle|n_{te,i}(\omega,\mathbf{k})|^2\rangle = 2KT_{e,i}\mathrm{Re}\{\sigma_{e,i}(\omega)\}. \tag{18.27}$$

It turns out that (18.27) is not specific to collisionless and non-magnetized plasmas; instead, it is a very general relation which is valid for each species of any plasma (i.e., including collisional and/or magnetized cases) in its own reference frame so long as the species are in thermal equilibrium. In fact, (18.27) is an *instance* of the *fluctuation-dissipation* theorem[6] of equilibrium statistical mechanics (see *Farley* [1966] and references

[6] Circuit formulae $\langle|I(\omega)|^2\rangle = 2KT\mathrm{Re}\{Y(\omega)\}$ and $\langle|V(\omega)|^2\rangle = 2KT\mathrm{Re}\{Z(\omega)\}$ describing the "thermal noise" spectra of *short-circuit current* I of an admittance Y and *open-circuit voltage* V of an impedance Z at a temperature T are well known examples of the fluctuation-dissipation theorem — note that the continuity equation implies $q^2\frac{\omega^2}{k^2}\langle|n_{te,i}(\omega,\mathbf{k})|^2\rangle = \langle|J_{e,i}(\omega,\mathbf{k})|^2\rangle$, and hence $\langle|J_{e,i}(\omega,\mathbf{k})|^2\rangle = 2KT_{e,i}\mathrm{Re}\{\sigma_{e,i}(\omega)\}$, leading to $\langle|I(\omega)|^2\rangle = 2KT\mathrm{Re}\{Y(\omega)\}$ in lumped circuits.

therein). The generality of (18.27) can be utilized to obtain $\langle|n_{te,i}(\omega,\mathbf{k})|^2\rangle$ after $\sigma_{e,i}(\omega)$ has been modified to include, for example, collision or DC magnetic field effects in the plasma. Conversely, we can also use (18.27) to determine $\mathrm{Re}\{\sigma_{e,i}(\omega)\}$ after modifying $\langle|n_{te,i}(\omega,\mathbf{k})|^2\rangle$ to include collision and/or magnetic field effects. Given $\mathrm{Re}\{\sigma_{e,i}(\omega)\}$, the complex $\sigma_{e,i}(\omega)$ needed in the spectrum formula (18.6) can be obtained using *Kramers-Kronig relations*

$$(\mathrm{Re,\ Im})\{\sigma_{e,i}(\omega)\} = \frac{1}{\pi}\lim_{\epsilon\to 0^+}\{\int_{-\infty}^{\omega-\epsilon} + \int_{\omega+\epsilon}^{\infty}\}\frac{(\text{-Im, Re})\{\sigma_{e,i}(x)\}}{x-\omega}dx, \tag{18.28}$$

another general principle known to be a consequence of causality (see, e.g., *Yeh and Liu*, [1972]) .

Clearly, the fluctuation-dissipation theorem (18.27) and Kramers-Kronig relations (18.28) provide a framework where $\langle|n_{te,i}(\omega,\mathbf{k})|^2\rangle$ and $\sigma_{e,i}(\omega)$ can be uniquely inferred from one another. In the following sections we will choose to derive first the fluctuation spectra $\langle|n_{te,i}(\omega,\mathbf{k})|^2\rangle$ by considering the motions of individual plasma particles in a given background (in the reference frame of the particle species and not including the fluctuating electric field response of the plasma). That, as we have already seen in (18.13), amounts to multiplying the temporal Fourier transform of a single particle correlation function $\langle e^{-j\mathbf{k}\cdot(\mathbf{r}_p(t+\tau)-\mathbf{r}_p(t))}\rangle$ with the mean particle density $N_{e,i}$. Once $\langle|n_{te,i}(\omega,\mathbf{k})|^2\rangle$ has been established, $\sigma_{e,i}(\omega)$ can be deduced and electron density spectrum constructed using the general framework outlined above. In practice, however, once $\langle e^{-j\mathbf{k}\cdot(\mathbf{r}_p(t+\tau)-\mathbf{r}_p(t))}\rangle$ is determined, the exercise shifts to identifying *modified* Gordeyev integrals $J(\theta_{e,i})$ which can be directly used in (18.15) and (18.23) for density and RCS spectra — an explicit use of Kramers-Kronig relations will not even be needed.

Modified Gordeyev integral and single particle ACF

Note that from (18.13) and (18.26) we have, effectively,

$$\langle|n_{ts}(\omega,\mathbf{k})|^2\rangle = N_s\int_{-\infty}^{\infty}d\tau e^{-j\omega\tau}\langle e^{j\mathbf{k}\cdot(\mathbf{r}_p(t+\tau)-\mathbf{r}_p(t))}\rangle = \frac{N_s}{\omega}\theta_s 2\mathrm{Re}\{J(\theta_s)\} \tag{18.29}$$

in the reference frame of any species s with a thermal speed C_s and individual particle trajectories $\mathbf{r}_p(t)$. With a change of variables $t' \equiv k\sqrt{2}C_s\tau$ the rightmost equality in (18.29) implies

$$\frac{N_s}{\omega}\theta_s\int_{-\infty}^{\infty}dt' e^{-j\theta_s t'}\langle e^{j\mathbf{k}\cdot(\mathbf{r}_p(t+\frac{t'}{k\sqrt{2}C_s})-\mathbf{r}_p(t))}\rangle = \frac{N_s}{\omega}\theta_s 2\mathrm{Re}\{J(\theta_s)\}, \tag{18.30}$$

and, consequently,

$$2\mathrm{Re}\{J(\theta)\} = \int_{-\infty}^{\infty}dt' e^{-j\theta t'}\langle e^{j\mathbf{k}\cdot(\mathbf{r}_p(t+\frac{t'}{k\sqrt{2}C})-\mathbf{r}_p(t))}\rangle. \tag{18.31}$$

Now, replacing the Fourier transform in (18.31) with a Laplace transform (justified because of Kramers-Kronig relations) and also swapping t and t', we identify the required generalized form of Gordeyev integral as

$$J(\theta) = \int_0^\infty dt e^{-j\theta t} \langle e^{j\mathbf{k}\cdot(\mathbf{r}_p(t'+\frac{t}{k\sqrt{2}C})-\mathbf{r}_p(t'))} \rangle. \tag{18.32}$$

Note that the Gordeyev integral (18.32) is the transform of the *single particle ACF*

$$\rho_p(\frac{t}{k\sqrt{2}C}) \equiv \langle e^{j\mathbf{k}\cdot(\mathbf{r}_p(t'+\frac{t}{k\sqrt{2}C})-\mathbf{r}_p(t'))} \rangle \tag{18.33}$$

of $e^{j\mathbf{k}\cdot\mathbf{r}_p(t)}$ — i.e., the Bragg scattered signal from a single particle — where the expected value is taken with respect to a 3-D Maxwellian velocity distribution

$$g_o(\mathbf{v}) = \frac{e^{-\frac{\mathbf{v}\cdot\mathbf{v}}{2C^2}}}{(2\pi)^{3/2}C^3} \tag{18.34}$$

(in the reference frame of species s). We emphasize here that the single particle ACF is none other than the normalized ACF (to unity peak) a backscatter radar would measure if a probed plasma (e.g., the ionosphere) were to behave as a collection of non-interacting particles (as originally envisioned by *Gordon* [1958]).

$J(\theta)$ for a magnetized but collisionless plasma

In the presence of a DC magnetic field $\mathbf{B} \equiv B\hat{z}$, free electrons will follow helical trajectories (right handed wrt $\mathbf{B}$) of the form

$$\mathbf{r}_p(t) = \mathrm{Re}\{(\hat{x} - j\hat{y})\frac{v_\perp}{\Omega}e^{j\Omega t}\} + \hat{z}(v_z t + z_o), \tag{18.35}$$

where $\Omega = \frac{|e|B}{m}$ is the electron gyrofrequency and the first term describes a circular orbit with a radius $\frac{|v_\perp|}{\Omega}$ and speed $|v_\perp|$ in the plane perpendicular to $\mathbf{B}$. Clearly then,

$$\begin{aligned} \mathbf{r}_p(t+\tau) - \mathbf{r}_p(t) &= \mathrm{Re}\{v_\perp e^{j\Omega t}(\hat{x} - j\hat{y})\frac{(e^{j\Omega\tau}-1)}{\Omega}\} + \hat{z}v_z\tau \\ &= \frac{v_\perp e^{j\Omega t}}{\sqrt{2}}\frac{\hat{x}-j\hat{y}}{\sqrt{2}}\frac{(e^{j\Omega\tau}-1)}{\Omega} + \frac{v_\perp^* e^{-j\Omega t}}{\sqrt{2}}\frac{\hat{x}+j\hat{y}}{\sqrt{2}}\frac{(e^{-j\Omega\tau}-1)}{\Omega} + \hat{z}v_z\tau \end{aligned} \tag{18.36}$$

Defining

$$\mathbf{M} \equiv \begin{bmatrix} \frac{(e^{j\Omega\tau}-1)}{\Omega} & 0 & 0 \\ 0 & \frac{(e^{-j\Omega\tau}-1)}{\Omega} & 0 \\ 0 & 0 & \tau \end{bmatrix} \text{ and } \mathbf{v} \equiv \begin{bmatrix} v_\perp e^{j\Omega t}/\sqrt{2} \\ v_\perp^* e^{-j\Omega t}/\sqrt{2} \\ v_z \end{bmatrix}, \tag{18.37}$$

where particle velocity $\mathbf{v}$ is now expressed in terms of its components along orthonormal basis vectors $\frac{\hat{x}-j\hat{y}}{\sqrt{2}}$, $\frac{\hat{x}+j\hat{y}}{\sqrt{2}}$, and $\hat{z}$, we can write

$$\mathbf{r}_p(t+\tau) - \mathbf{r}_p(t) = \mathbf{M}\mathbf{v} \tag{18.38}$$

to represent vector $\mathbf{r}_p(t+\tau) - \mathbf{r}_p(t)$ in the same basis as a vector matrix product. Also, a propagation vector at a *magnetic aspect angle* of α can be expressed as

$$\mathbf{k} \equiv k\cos\alpha\hat{x} + k\sin\alpha\hat{z} = \begin{bmatrix} k\cos\alpha/\sqrt{2} \\ k\cos\alpha/\sqrt{2} \\ k\sin\alpha \end{bmatrix} \tag{18.39}$$

in the same basis, and thus

$$\mathbf{k}\cdot(\mathbf{r}_p(t+\tau) - \mathbf{r}_p(t)) = \mathbf{k}^T\mathbf{M}\mathbf{v} = \mathbf{a}^T\mathbf{v} = \mathbf{a}\cdot\mathbf{v} \tag{18.40}$$

with

$$\mathbf{a} \equiv \begin{bmatrix} k\frac{(e^{j\Omega\tau}-1)}{\sqrt{2}\Omega}\cos\alpha \\ k\frac{(e^{-j\Omega\tau}-1)}{\sqrt{2}\Omega}\cos\alpha \\ k\tau\sin\alpha \end{bmatrix} = \mathrm{Re}\{(\hat{x}-j\hat{y})k\tau\frac{(e^{j\Omega\tau}-1)}{\Omega\tau}\cos\alpha\} + \hat{z}k\tau\sin\alpha \tag{18.41}$$

and

$$\mathbf{a}\cdot\mathbf{a} = k^2\tau^2[\sin^2\alpha + \mathrm{sinc}^2(\frac{\Omega\tau}{2})\cos^2\alpha]. \tag{18.42}$$

The single particle ACF in the presence of a DC magnetic field can therefore be obtained as

$$\begin{aligned} \rho_p(\tau) &\equiv \langle e^{j\mathbf{k}\cdot(\mathbf{r}_p(t+\tau)-\mathbf{r}_p(t))}\rangle = \langle e^{j\mathbf{a}.\mathbf{v}}\rangle = \int d\mathbf{v}\frac{e^{-\frac{\mathbf{v}\cdot\mathbf{v}}{2C^2}}}{(2\pi)^{3/2}C^3}e^{j\mathbf{a}.\mathbf{v}} \\ &= e^{-\frac{C^2\mathbf{a}.\mathbf{a}}{2}} = e^{-\frac{k^2C^2\tau^2}{2}[\sin^2\alpha+\mathrm{sinc}^2(\frac{\Omega\tau}{2})\cos^2\alpha]}, \end{aligned} \tag{18.43}$$

which yields

$$\rho_p(\frac{t}{k\sqrt{2}C}) = e^{-\frac{t^2}{4}[\sin^2\alpha+\mathrm{sinc}^2(\frac{\phi t}{2})\cos^2\alpha]} = e^{-[\frac{t^2}{4}\sin^2\alpha+\frac{1}{\phi^2}\sin^2(\frac{\phi t}{2})\cos^2\alpha]} \tag{18.44}$$

for the modified Gordeyev integral appropriate for a magnetized plasma, with

$$\phi \equiv \frac{\Omega}{k\sqrt{2}C} \tag{18.45}$$

defining a normalized electron gyrofrequency. Thus, the modified integral for any species s is of the form

$$J(\theta) = \int_0^\infty dt e^{-j\theta t}e^{-[\frac{t^2}{4}\sin^2\alpha+\frac{1}{\phi^2}\sin^2(\frac{\phi t}{2})\cos^2\alpha]}, \tag{18.46}$$

where, as already noted, α is the magnetic aspect angle, the complement of the angle between propagation vector $\mathbf{k}$ and magnetic field $\mathbf{B}$. This result has been obtained using a variety of approaches (see *Farley* [1971] and references therein) even though the derivation here is closest to one given by *Hagfors* [2003].

Notice that the transformed function in Gordeyev integral (18.46) is not absolute integrable in $\alpha \to 0$ limit (and thus the integral does not converge for real θ) as opposed to the

case for a finite α or non-magnetized plasma. As a consequence spectrum models based on above $J(\theta)$ fail to account for experimental data (e.g., ionospheric radar backscatter observations) taken in $\alpha \to 0$ limit (although otherwise perform remarkably well). In physical terms the failure has to do with undamped nature of electrostatic modes propagating exactly perpendicular to **B** in a collisionless plasma. Collisionless damping does not work when $\alpha = 0$ and we need to include the very weak (and ordinarily negligible) collision effects (see next section) to ensure absolute integrability of (18.46) integrand for all α in the presence of a DC magnetic field.

$J(\theta)$ with Coulomb collisions and open questions

Woodman [1967] examined the effect of collisions between charged particles — so called *Coulomb collisions* — on density fluctuations in a magnetized plasma using a Fokker-Planck type collision operator (added to the right-hand-side of Vlasov equation). Woodman's spectral formula (derived using a special method developed only for $T_e = T_i$ case) is consistent with a modified Gordeyev integral that can be expressed as [*Kudeki et al.*, 1999]

$$J(\theta) = \int_0^\infty dt e^{-j\theta t} e^{-[\frac{\psi t - 1 + e^{-\psi t}}{2\psi^2} \sin^2\alpha + \frac{\cos(2\gamma) + \psi t - e^{-\psi t}\cos(\phi t - 2\gamma)}{2(\psi^2+\phi^2)} \cos^2\alpha]}, \tag{18.47}$$

where

$$\psi \equiv \frac{\nu}{k\sqrt{2C}} \tag{18.48}$$

denotes the normalized collision frequency of charged species and

$$\gamma \equiv \tan^{-1}\frac{\nu}{\Omega} = \tan^{-1}\frac{\psi}{\phi}. \tag{18.49}$$

Notice that integrand in (18.47) is absolute integrable except in $\psi \to 0$ limit when it relaxes to the collisionless result of the previous section (as it should). Also, the non-magnetized limit ($\phi \to 0$) of the integrand agrees with the collisional but non-magnetized result of *Hagfors and Brockelman* [1971] (and others) derived using a Brownian motion model (for collisions with neutral particles in mind).

We conjecture here that the collisional *and* magnetized Gordeyev integral of Woodman derived for the case of $\mu = \frac{T_e}{T_i} = 1$ should be valid also for $\mu > 1$ — after all single particle ACF's used to obtain $J(\theta)$'s are immune to variations of μ. However, the definition of Coulomb collision frequency ν remains far from settled. *Woodman* [2004] proposed an aspect angle dependent electron collision frequency $\nu_e = \nu_e(\alpha)$ to model electron collisions (with all charged particles), which is in fact just a "best fit" of Woodman's Gordeyev integral to a library of collisional electron Gordeyev integrals obtained by *Sulzer and Gonzales* [1999] based on Monte-Carlo estimates of single particle ACF's. Monte Carlo calculations do not extend to very small aspect angles α, which is a regime that needs further attention, in particular in view of recent incoherent scatter radar data from ALTAIR [e.g., *Kudeki and Milla*, 2005] and Jicamarca examining the variation of radar backscatter from the daytime F-region ionosphere at small aspect angles and $\mu > 1$.

Another unsolved problem, enhanced backscatter from the 150-km altitude region, also concerns $\alpha \to 0$ limit in a region of strong μ gradients.

Note added in Jan 2010: The above problem has been addressed in the PhD thesis work of UIUC graduate student Marco Milla [see Kudeki and Milla (2010) and Milla and Kudeki (2010) — IEEE TGRS]. Updated ECE 458 Class Notes (Sp10 version) will tell the revised story.

19 Appendix XI|: Ionogram inversions

20 Appendix XII|: Partial reflection radars

The result in Chapter 3 Section 3.6.4 implies that the steady-state response of the ionosphere to a transmitted (and up-going) field $\sim e^{j\omega t}$ is a reflected (down-coming) field $\sim R_p e^{j\omega t}$. Thus, a transmitted narrowband field

$$f(t)\cos(\omega_o t) = \frac{1}{2}\int \frac{d\omega}{2\pi} F(\omega-\omega_o)e^{j\omega t} + c.c. \;\; \leftrightarrow \;\; \frac{1}{2}F(\omega-\omega_o) + \frac{1}{2}F(\omega+\omega_o)$$

generates a reflected field that varies as

$$\frac{1}{2}\int \frac{d\omega}{2\pi} F(\omega-\omega_o)R_p e^{j\omega t} + c.c. = \frac{1}{2}\int_0^\infty dz\delta N(z)\int \frac{d\omega}{2\pi} F(\omega-\omega_o) j\lambda_o r_e e^{j(\omega t - 2k_o z)} + c.c.,$$

which is the real part of

$$\int_0^\infty dz\delta N(z)\int \frac{d\omega}{2\pi} F(\omega-\omega_o) j\lambda_o r_e e^{j(\omega t - 2k_o z)},$$

where $k_o = \frac{\omega}{c}$ and $\lambda_o = \frac{2\pi}{k_o}$ both depend on ω. However, if the bandwidth of $f(t) \leftrightarrow F(\omega)$ is small compared to ω_o (which is what we will assume) then λ_o above can be calculated for $\omega = \omega_o$ and the partial reflected field can be approximated as

$$-\lambda_o r_e \int_0^\infty dz\delta N(z) f(t-\frac{2z}{c})\sin(\omega_o(t-\frac{2z}{c})).$$

In a practical setting (e.g., at a radar station) the down-coming field will be *mixed* with a signal like $-2je^{-j\omega_o t}$ (digitally, nowadays, in a DSP card or computer supporting complex arithmetic) and *low-pass filtered* to obtain a complex baseband signal, say,

$$\begin{aligned} V(t) &= \lambda_o r_e \int_0^\infty dz\delta N(z) \underbrace{\int \frac{d\omega}{2\pi} F(\omega-\omega_o) e^{j(\omega-\omega_o)(t-\frac{2z}{c})}}_{f(t-\frac{2z}{c})} e^{-j\omega_o\frac{2z}{c}} \\ &= \lambda_o r_e \int_0^\infty dz\delta N(z) f(t-\frac{2z}{c}) e^{-j2k_o z}, \end{aligned}$$

where in the last line $k_o = \frac{\omega}{c}$ is for $\omega = \omega_o$. If the low-pass filter *impulse response* $h(t)$ is taken into account properly, this result should be re-written as

$$V(t) = \lambda_o r_e \int_0^\infty dz\delta N(z) W(t-\frac{2z}{c}) e^{-j2k_o z},$$

where

$$W(t) \equiv h(t) * f(t)$$

is a windowing function that depends on both the transmitted pulse shape $f(t)$ and filter shape $h(t)$.

In any case, assuming that $W(t)$ peaks around $t = 0$, the above result indicates that complex response $V(t)$ measured at the filter output at time t (and stored in a computer file) is representative of Bragg scale density waves in the vicinity of ionospheric height $z = \frac{ct}{2}$. Furthermore, if the width of $W(t)$ is some Δt, then the width of the pertinent altitude region about $z = \frac{ct}{2}$ is some $\Delta z = \frac{c\Delta t}{2}$. Finally, it should be mentioned that filter $h(t)$ is employed (among other reasons) to limit the contributions of additive background noise to the baseband signal $V(t)$.