Chapter 3

Scattering series

In this chapter we describe the nonlinearity of the map $c \mapsto u$ in terms of a perturbation (Taylor) series. To first order, the linearization of this map is called the Born approximation. Linearization and scattering series are the basis of most inversion methods, both direct and iterative.

The idea of perturbation permeates imaging for physical reasons as well. In radar imaging for instance, the background velocity is $c_0 = 1$ (speed of light), and the *reflectivity* of scatterers is viewed as a deviation in $c(x)$. The assumption that $c(x)$ does not depend on t is a strong one in radar: it means that the scatterers do not move. In seismology, it is common to consider a smooth background velocity $c_0(x)$ (rarely well known), and explain the scattered waves as reflections due to a "rough" (singular/oscillatory) perturbations to this background. In both cases, we will write

$$\frac{1}{c^2(x)} = m(x), \qquad \frac{1}{c_0^2(x)} = m_0(x), \qquad m \text{ for "model"},$$

and, for some small number ε,

$$m(x) = m_0(x) + \varepsilon m_1(x). \tag{3.1}$$

Note that, when perturbing $c(x)$ instead of $m(x)$, an additional Taylor approximation is necessary:

$$c(x) = c_0(x) + \varepsilon c_1(x) \qquad \Rightarrow \qquad \frac{1}{c^2(x)} \simeq \frac{1}{c_0^2(x)} - 2\varepsilon \frac{c_1(x)}{c_0^3(x)}.$$

While the above is common in seismology, we avoid making unnecessary assumptions by choosing to perturb $m(x) = 1/c^2(x)$ instead.

Perturbations are of course not limited to the wave equation with a single parameter c. The developments in this chapter clearly extend to more general wave equations.

3.1 Perturbations and Born series

Let

$$m(x)\frac{\partial^2 u}{\partial t^2} - \Delta u = f(x,t), \tag{3.2}$$

with zero initial conditions and $x \in \mathbb{R}^n$. Perturb $m(x)$ as in (3.1). The wavefield u correspondingly splits into

$$u(x) = u_0(x) + u_{sc}(x),$$

where u_0 solves the wave equation in the undisturbed medium m_0,

$$m_0(x)\frac{\partial^2 u_0}{\partial t^2} - \Delta u_0 = f(x,t). \tag{3.3}$$

We say u is the total field, u_0 is the incident field[1], and u_{sc} is the scattered field, i.e., anything but the incident field.

We get the equation for u_{sc} by subtracting (3.3) from (3.2), and using (3.1):

$$m_0(x)\frac{\partial^2 u_{sc}}{\partial t^2} - \Delta u_{sc} = -\varepsilon\, m_1(x)\frac{\partial^2 u}{\partial t^2}. \tag{3.4}$$

This equation is implicit in the sense that the right-hand side still depends on u_{sc} through u. We can nevertheless reformulate it as an implicit integral relation by means of the Green's function:

$$u_{sc}(x,t) = -\varepsilon \int_0^t\!\!\int_{\mathbb{R}^n} G(x,y;t-s) m_1(y) \frac{\partial^2 u}{\partial t^2}(y,s)\, dyds.$$

Abuse notations slightly, but improve conciseness greatly, by letting

- G for the operator of space-time integration against the Green's function, and

[1]Here and in the sequel, u_0 is not the initial condition. It is so prevalent to introduce the source as a right-hand side f in imaging that it is advantageous to free the notation u_0 and reserve it for the incident wave.

- m_1 for the operator of multiplication by m_1.

Then $u_{sc} = -\varepsilon \, G \, m_1 \frac{\partial^2 u}{\partial t^2}$. In terms of u, we have the implicit relation

$$u = u_0 - \varepsilon \, G \, m_1 \frac{\partial^2 u}{\partial t^2},$$

called a *Lippmann-Schwinger* equation. The field u can be formally[2] expressed in terms of u_0 by writing

$$u = \left[I + \varepsilon \, G \, m_1 \frac{\partial^2}{\partial t^2} \right]^{-1} u_0. \tag{3.5}$$

While this equation is equivalent to the original PDE, it shines a different light on the underlying physics. It makes explicit the link between u_0 and u, as if u_0 "generated" u via scattering through the medium perturbation m_1.

Writing $[I + A]^{-1}$ for some operator A invites a solution in the form of a Neumann series $I - A + A^2 - A^3 + \ldots$, provided $\|A\| < 1$ in some norm. In our case, we write

$$u = u_0 - \varepsilon \left(G \, m_1 \frac{\partial^2}{\partial t^2} \right) u_0 + \varepsilon^2 \left(G \, m_1 \frac{\partial^2}{\partial t^2} \right) \left(G \, m_1 \frac{\partial^2}{\partial t^2} \right) u_0 + \ldots$$

This is called a *Born series*. The proof of convergence, based on the "weak scattering" condition $\varepsilon \|G \, m_1 \frac{\partial^2}{\partial t^2}\|_* < 1$, in some norm to be determined, will be covered in the next section. It retroactively justifies why one can write (3.5) in the first place.

The Born series carries the physics of *multiple scattering*. Explicitly,

$$
\begin{aligned}
u = {} & u_0 \qquad\qquad \text{(incident wave)} \\
& - \varepsilon \int_0^t \int_{\mathbb{R}^n} G(x, y; t - s) m_1(y) \frac{\partial^2 u_0}{\partial t^2}(y, s) \, dy ds \\
& \qquad\qquad \text{(single scattering)} \\
& + \varepsilon^2 \int_0^t \int_{\mathbb{R}^n} G(x, y_2; t - s_2) m_1(y_2) \frac{\partial^2}{\partial s_2^2} \left[\int_0^{s_2} \int_{\mathbb{R}^n} G(y_2, y_1; s_2 - s_1) m_1(y_1) \frac{\partial^2 u_0}{\partial t^2}(y_1, s_1) \, dy_1 ds_1 \right] dy_2 ds_2 \\
& \qquad\qquad \text{(double scattering)} \\
& + \ldots
\end{aligned}
$$

[2]For mathematicians, "formally" means that we are a step ahead of the rigorous exposition: we are only interested in inspecting the *form* of the result before we go about proving it. That's the intended meaning here. For non-mathematicians, "formally" often means rigorous, i.e., the opposite of "informally"!

We will naturally summarize this expansion as

$$u = u_0 + \varepsilon u_1 + \varepsilon^2 u_2 + \ldots \tag{3.6}$$

where εu_1 represent single scattering, $\varepsilon^2 u_2$ double scattering, etc. For instance, the expression of u_1 can be physically read as "the incident wave initiates from the source at time $t = 0$, propagates to y where it scatters due to $m(y)$ at time $t = s$, then further propagates to reach x at time t." The expression of u_2 can be read as "the incident wave initiates from the source at $t = 0$, propagates to y_1 where it first scatters at time $t = s_1$, them propagates to y_2 where it scatters a second time at time $t = s_2$, then propagates to x at time t, where it is observed." Since scatterings are not a priori prescribed to occur at fixed points in space and time, integrals must be taken to account for all physically acceptable scattering scenarios.

The approximation

$$u_{sc}(x) \simeq \varepsilon u_1(x)$$

is called the *Born approximation*. From $u_1 = -Gm_1 \frac{\partial^2 u_0}{\partial t^2}$, we can return to the PDE and obtain the equation for the primary reflections:

$$m_0(x) \frac{\partial^2 u_1}{\partial t^2} - \Delta u_1 = -m_1(x) \frac{\partial^2 u_0}{\partial t^2}. \tag{3.7}$$

The only difference with (3.4) is the presence of u_0 in place of u in the right-hand side (and ε is gone, by choice of normalization of u_1). Unlike (3.4), equation (3.7) is explicit: it maps m_1 to u_1 in a linear way. The incident field u_0 is determined from m_0 alone, hence "fixed" for the purpose of determining the scattered fields.

It is informative to make explicit the dependence of $u_1, u_2, \ldots$ on m_1. To that end, the Born series can be seen as a Taylor series of the *forward map*

$$u = \mathcal{F}[m],$$

in the sense of the calculus of variations. Denote by $\frac{\delta \mathcal{F}}{\delta m}[m_0]$ the "functional gradient" of $\mathcal{F}$ with respect to m, evaluated at m_0. It is an operator acting from model space (m) to data space (u). Denote by $\frac{\delta^2 \mathcal{F}}{\delta m^2}[m_0]$ the "functional Hessian" of $\mathcal{F}$ with respect to m, evaluated at m_0. It is a bilinear form from model space to data space. See the appendix for background on functional derivatives. Then the functional version of the Taylor expansion enables to express (3.6) in terms of the various derivatives of $\mathcal{F}$ as

$$u = u_0 + \varepsilon \frac{\delta \mathcal{F}}{\delta m}[m_0]\, m_1 + \frac{\varepsilon^2}{2} \langle \frac{\delta^2 \mathcal{F}}{\delta m^2}[m_0]\, m_1, m_1 \rangle + \ldots$$

It is convenient to denote the *linearized forward map* by (print) F:

$$F = \frac{\delta \mathcal{F}}{\delta m}[m_0],$$

or, for short, $F = \frac{\partial u}{\partial m}$. It is a linear operator. The point of F is that it makes explicit the linear link between m_1 and u_1:

$$u_1 = F m_1.$$

While $\mathcal{F}$ is supposed to completely model data (up to measurement errors), F would properly explain data only in the regime of the Born approximation.

Let us show that the two concepts of linearized scattered field coincide, namely

$$u_1 = \frac{\delta \mathcal{F}}{\delta m}[m_0]\, m_1 = -G m_1 \frac{\partial^2 u_0}{\partial t^2}.$$

This will justify the first term in the Taylor expansion above. For this purpose, let us take the $\frac{\delta}{\delta m}$ derivative of (3.2). As previously, write $u = \mathcal{F}(m)$ and $F = \frac{\delta \mathcal{F}}{\delta m}[m]$. We get the operator-valued equation

$$\frac{\partial^2 u}{\partial t^2} I + m \frac{\partial^2}{\partial t^2} F - \Delta F = 0.$$

Evaluate the functional derivatives at the base point m_0, so that $u = u_0$. Applying each term as an operator to the function m_1, and defining $u_1 = F m_1$, we obtain

$$m_1 \frac{\partial^2 u_0}{\partial t^2} + m_0 \frac{\partial^2 u_1}{\partial t^2} - \Delta u_1 = 0,$$

which is exactly (3.7). Applying G on both sides, we obtain the desired conclusion that $u_1 = -G m_1 \frac{\partial^2 u_0}{\partial t^2}$.

3.2 Convergence of the Born series (math)

We are faced with two very interrelated questions: justifying convergence of the Born series, and showing that the Born approximation is accurate when the Born series converges. The answers can either take the form of mathematical theorems (this section), or physical explanations (next section). As of 2013, the community's mathematical understanding is not yet up to par with the physical intuition!

Let us describe what is known mathematically about convergence of Born series in a simple setting. To keep the notations concise, it is more convenient to treat the wave equation in first-order hyperbolic form

$$M\frac{\partial w}{\partial t} - Lw = f, \qquad L^* = -L, \tag{3.8}$$

for some inner product $\langle w, w' \rangle$. The conserved energy is then $E = \langle w, Mw \rangle$. See one of the exercises at the end of chapter 1 to illustrate how the wave equation can be put in precisely this form, with $\langle w, w' \rangle$ the usual L^2 inner product and M a positive diagonal matrix.

Consider a background medium M_0, so that $M = M_0 + \varepsilon M_1$. Let $w = w_0 + \varepsilon w_1 + \ldots$ Calculations very similar to those of the previous section (a good exercise) show that

- The Lippmann-Schwinger equation is

  $$w = w_0 - \varepsilon G M_1 \frac{\partial w}{\partial t},$$

  with the Green's function $G = (M_0 \frac{\partial}{\partial t} - L)^{-1}$.

- The Neumann series of interest is

  $$w = w_0 - \varepsilon G M_1 \frac{\partial w_0}{\partial t} + \varepsilon^2 G M_1 \frac{\partial}{\partial t} G M_1 \frac{\partial w_0}{\partial t} + \ldots$$

  We identify $w_1 = -G M_1 \frac{\partial w_0}{\partial t}$.

- In differential form, the equations for the incident field w_0 and the primary scattered field w_1 are

  $$M_0 \frac{\partial w_0}{\partial t} - L w_0 = f, \qquad M_0 \frac{\partial w_1}{\partial t} - L w_1 = -M_1 \frac{\partial w_0}{\partial t}, \tag{3.9}$$

- Convergence of the Born series occurs when

  $$\varepsilon \| G M_1 \frac{\partial}{\partial t} \|_* < 1,$$

  in some induced operator norm, i.e., when $\varepsilon \|w_1\|_* < \|w_0\|_*$ for arbitrary w_0, and $w_1 = -G M_1 \frac{\partial w_0}{\partial t}$, for some norm $\| \cdot \|_*$.

Notice that the condition $\varepsilon \|w_1\|_* < \|w_0\|_*$ is precisely one of weak scattering, i.e., that the primary reflected wave εw_1 is weaker than the incident wave w_0.

While any induced norm over space and time in principle works for the proof of convergence of the Neumann series, it is convenient to use

$$\|w\|_* = \max_{0 \le t \le T} \sqrt{\langle w, M_0 w \rangle} = \max_{0 \le t \le T} \|\sqrt{M_0} w\|.$$

Note that it is a norm in space *and* time, unlike $\|w\| = \sqrt{\langle w, w \rangle}$, which is only a norm in space.

Theorem 3. *(Convergence of the Born series) Assume that the fields w, w_0, w_1 are bandlimited with bandlimit*[3] *Ω. Consider these fields for $t \in [0, T]$. Then the weak scattering condition $\varepsilon \|w_1\|_* < \|w_0\|_*$ is satisfied, hence the Born series converges, as soon as*

$$\varepsilon \, \Omega T \, \| \frac{M_1}{M_0} \|_\infty < 1.$$

Proof. We compute

$$\begin{aligned}
\frac{d}{dt} \langle w_1, M_0 w_1 \rangle &= 2 \langle w_1, M_0 \frac{\partial w_1}{\partial t} \rangle \\
&= 2 \langle w_1, L w_1 - M_1 \frac{\partial w_0}{\partial t} \rangle \\
&= -2 \langle w_1, M_1 \frac{\partial w_0}{\partial t} \rangle \qquad \text{because } L^* = -L \\
&= -2 \langle \sqrt{M_0} w_1, \frac{M_1}{\sqrt{M_0}} \frac{\partial w_0}{\partial t} \rangle.
\end{aligned}$$

Square roots and fractions of positive diagonal matrices are legitimate operations. The left-hand-side is also $\frac{d}{dt} \langle w_1, M_0 w_1 \rangle = 2 \|\sqrt{M_0} w_1\|_2 \frac{d}{dt} \|\sqrt{M_0} w_1\|_2$. By Cauchy-Schwarz, the right-hand-side is majorized by

$$2 \|\sqrt{M_0} w_1\|_2 \, \| \frac{M_1}{\sqrt{M_0}} \frac{\partial w_0}{\partial t} \|_2.$$

Hence

$$\frac{d}{dt} \|\sqrt{M_0} w_1\|_2 \le \| \frac{M_1}{\sqrt{M_0}} \frac{\partial w_0}{\partial t} \|_2.$$

[3] A function of time has bandlimit Ω when its Fourier transform, as a function of ω, is supported in $[-\Omega, \Omega]$.

$$\|\sqrt{M_0}w_1\|_2 \leq \int_0^t \|\frac{M_1}{\sqrt{M_0}}\frac{\partial w_0}{\partial t}\|_2(s)\,ds.$$

$$\|w_1\|_* = \max_{0\leq t\leq T}\|\sqrt{M_0}w_1\|_2 \leq T\max_{0\leq t\leq T}\|\frac{M_1}{\sqrt{M_0}}\frac{\partial w_0}{\partial t}\|_2$$
$$\leq T\|\frac{M_1}{M_0}\|_\infty \max_{0\leq t\leq T}\|\sqrt{M_0}\frac{\partial w_0}{\partial t}\|_2.$$

This last inequality is almost, but not quite, what we need. The right-hand side involves $\frac{\partial w_0}{\partial t}$ instead of w_0. Because time derivatives can grow arbitrarily large in the high-frequency regime, this is where the bandlimited assumption needs to be used. We can invoke a classical result known as Bernstein's inequality[4], which says that $\|f'\|_\infty \leq \Omega\|f\|_\infty$ for all Ω-bandlimited f. Then

$$\|w_1\|_* \leq \Omega T\|\frac{M_1}{M_0}\|_\infty\|w_0\|_*.$$

In view of our request that $\varepsilon\|w_1\|_* < \|w_0\|_*$, it suffices to require

$$\varepsilon\,\Omega T\,\|\frac{M_1}{M_0}\|_\infty < 1.$$

$\square$

See the book Inverse Acoustic and Electromagnetic Scattering Theory by Colton and Kress for a different analysis that takes into account the size of the support of M_1.

Note that the beginning of the argument, up to the Cauchy-Scwharz inequality, is called an *energy estimate* in math. See an exercise at the end of this chapter. It is a prevalent method to control the size of the solution of many initial-value PDE, including nonlinear ones.

The weak scattering condition $\varepsilon\|w_1\|_* < \|w_0\|_*$ encodes the idea that the primary reflected field εw_1 is small compared to the incident field w_0. It is satisfied when ε is small, and when w_1 is not so large that it would undo the smallness of ε (via the factors ΩT, for instance). It turns out that

- the full scattered field $w_{sc} = w - w_0$ is also on the order of $\varepsilon\Omega T\|M_1\|_\infty$ — namely the high-order terms don't compromise the weak scattering situation; and

[4]The same inequality holds with the L^p norm for all $1 \leq p \leq \infty$.

- the remainder $w_{sc} - \varepsilon w_1 = w - w_0 - \varepsilon w_1$ is on the order of $\varepsilon^2 (\Omega T \|M_1\|_\infty)^2$.

Both claims are the subject of an exercise at the end of the chapter. The second claim is the mathematical expression that the Born approximation is accurate (small $w_{sc} - \varepsilon w_1$ on the order of ε^2) precisely when scattering is weak (εw_1 and w_{sc} on the order of ε.)

3.3 Convergence of the Born series (physics)

Let us explain why the criterion $\varepsilon \Omega T < 1$ (assuming the normalization $\|M_1/M_0\|_\infty = 1$) is adequate in some cases, and why it is grossly pessimistic in others.

- Instead of m or M, consider the wave speed $c_0 = 1$. Consider a constant perturbation $c_1 = 1$, so that $c = c_0 + \varepsilon c_1 = 1 + \varepsilon$. In one spatial dimension, $u(x, T) = f(x - cT)$. As a Taylor series in ε, this is

  $$u(x,T) = f(x-(1+\varepsilon)T) = f(x-T) - \varepsilon T f'(x-T) + \frac{\varepsilon^2}{2} T^2 f''(x-T) + \ldots$$

  We identify $u_0(x,T) = f(x-T)$ and $u_1(x,T) = -T f'(x-T)$. Assume now that f is a waveform with bandlimit Ω, i.e., wavelength $2\pi/\Omega$. The Born approximation

  $$f(x-(1+\varepsilon)T) - f(x-T) \simeq -\varepsilon T f'(x-T)$$

  is only good when the translation step εT between the two waveforms on the left is a small fraction of a wavelength $2\pi/\Omega$, otherwise the subtraction $f(x-(1+\varepsilon)T) - f(x-T)$ will be out of phase and will not give rise to values on the order of ε. The requirement is $\varepsilon T \ll 2\pi/\Omega$, i.e.,

  $$\varepsilon \Omega T \ll 2\pi,$$

  which is exactly what theorem 3 is requiring. We could have reached the same conclusion by requiring either the first or the second term of the Taylor expansion to be $o(1)$, after noticing that $|f'| = O(\Omega)$ or $|f''| = O(\Omega^2)$. In the case of a constant perturbation $c_1 = 1$, the waves undergo a shift which quickly becomes nonlinear in the perturbation. This is the worst case: the requirement $\varepsilon \Omega T < 1$ is sharp.

- As a second example, consider $c_0 = 1$ and $c_1(x) = H(x)$. The profile of reflected and transmitted waves was studied in equations (1.20) and (1.21). The transmitted wave will undergo a shift as in the previous example, so we expect $\varepsilon \Omega T < 1$ to be sharp for it. The full reflected wave, on the other hand, is

$$u_r(x,T) = R_\varepsilon f(-x-T), \qquad R_\varepsilon = \frac{\varepsilon}{2+\varepsilon}.$$

  Notice that ε only appears in the reflection coefficient R_ε, not in the waveform itself. As $\varepsilon \to 0$, u_r expands as

$$u_r(x,T) = \frac{\varepsilon}{2} f(-x-T) - \frac{\varepsilon^2}{4} f(-x-T) + \ldots$$

  We recognize $u_1 = \frac{1}{2} f(-x-T)$. The condition for weak scattering and accuracy of the Born approximation is now simply $\varepsilon < 1$, which is in general much weaker than $\varepsilon \Omega T < 1$.

- In the case when $c_0 = 1$ and c_1 is the indicator function of a thin slab in one dimension, or a few isolated scatterers in several dimensions, the Born approximation is often very good. That's when the interpretation of the Born series in terms of multiple scattering is the most relevant. Such is the case of small isolated objects in synthetic aperture radar: double scattering from one object to another is often negligible.

The Born approximation is often satisfied in the low-frequency regime (small Ω), by virtue of the fact that cycle skipping is not as much of an issue. In the high-frequency regime, the heuristics for validity of the Born approximation are that

1. c_0 or m_0 should be *smooth*.
2. c_1 or m_1 should be *localized*, or better yet, localized and oscillatory (zero mean).

The second requirement is the most important one: it prohibits transmitted waves from propagating in the wrong velocity for too long. We do not yet have a way to turn these empirical criteria and claims into rigorous mathematical results. Seismologists typically try to operate in the regime of this heuristic when performing imaging with migration (see chapter on seismic imaging).

Conversely, there are a few settings in which the Born approximation is clearly violated: (i) in radar, when waves bounce multiple times before being recorded (e.g. on the ground and on the face of a building, or in cavities such as airplane engines), (ii) in seismology, when trying to optimize over the small-wavenumber components of $m(x)$ (model velocity estimation), or when dealing with multiple scattering (internal multiples). However, note that multiple reflections from features already present in the modeling (such as ghosts due to reflections at the ocean-air interface) do not count as non-linearities.

Scattered waves that do not satisfy the Born approximation have long been considered a nuisance in imaging, but have recently become the subject of some research activity.

3.4 A first look at optimization

In the language of the previous sections, the forward map is denoted

$$d = \mathcal{F}[m], \qquad d = \text{data}, \qquad m = \text{model},$$

where $d_{r,s}(t) = u_s(x_r, t)$,

- x_r is the position of receiver r,
- s indexes the source,
- and t is time.

The inverse problem of imaging is that of solving for m in the system of nonlinear equations $d = \mathcal{F}[m]$. No single method will convincingly solve such a system of nonlinear equations efficiently and in all regimes.

The very prevalent *least-squares* framework formulate the inverse problem as finding m as the solution of the minimization problem

$$\min_m J[m], \qquad \text{where} \qquad J[m] = \frac{1}{2}\|d - \mathcal{F}[m]\|_2^2, \tag{3.10}$$

where $\|d\|_2^2 = \sum_{r,s} \int_0^T |d_{r,s}(t)|^2$ is the L^2 norm squared in the space of vectors indexed by r, s (discrete) and t (continuous, say). J is called the output least-squares criterion, or objective, or cost.

In the sequel we consider iterative schemes based on the variations of J at a base point m_0, namely the functional gradient $\frac{\delta J}{\delta m}[m_0]$, a linear functional in m space; and the functional Hessian $\frac{\delta^2 J}{\delta m^2}[m_0]$, also called wave-equation Hessian, an operator (or bilinear form) in m space. The appendix contains a primer on functional calculus.

Two extreme scenarios cause problems when trying to solve for m as the minimizer of a functional J:

- The inverse problem is called *ill-posed* when there exist directions m_1 in which $J(m)$ has a zero curvature, or a very small curvature, in the vicinity of the solution m^*. Examples of such directions are the eigenvectors of the Hessian of J associated to *small* eigenvalues. The curvature is then twice the eigenvalue, i.e., twice the second directional derivative in the eigen-direction. Small perturbations of the data, or of the model $\mathcal{F}$, induce modifications of J that may result in large movements of its global minimum in problematic directions in the "near-nullspace" of the Hessian of J.

- Conversely, the inverse problem may suffer from severe non-convexity when the abundance of local minima, or local "valleys", hinders the search for the global minimum. This happens when the Hessian of J alternates between having *large* positive and negative curvatures in some direction m_1.

Many inversion problems in high-frequency imaging suffer from some (not overwhelming) amount of ill-posedness, and can be quite non-convex. These topics will be further discussed in chapter 9.

The gradient descent method[5] applied to J is simply

$$m^{(k+1)} = m^{(k)} - \alpha \frac{\delta J}{\delta m}[m^{(k)}]. \tag{3.11}$$

The choice of α is a balance between stability and speed of convergence – see two exercises at the end of the chapter. In practice, a line search for α is often a good idea.

The usual rules of functional calculus give the expression of $\frac{\delta J}{\delta m}$, also known as the "sensitivity kernel" of J with respect to m.

[5] Also called Landweber iteration in this nonlinear context.

Proposition 4. *Put* $F = \frac{\delta \mathcal{F}}{\delta m}[m]$. *Then*

$$\frac{\delta J}{\delta m}[m] = F^*(\mathcal{F}[m] - d).$$

Proof. Since $\mathcal{F}[m+h] = \mathcal{F}[m] + Fh + O(\|h\|^2)$, we have

$$\langle \mathcal{F}[m+h]-d, \mathcal{F}[m+h]-d\rangle = \langle \mathcal{F}[m]-d, \mathcal{F}[m]-d\rangle + 2\langle Fh, \mathcal{F}[m]-d\rangle + O(\|h\|^2).$$

Therefore

$$\begin{aligned} J[m+h] - J[m] &= \frac{1}{2} 2\langle Fh, \mathcal{F}[m] - d\rangle + O(\|h\|^2) \\ &= \langle h, F^*(\mathcal{F}[m] - d)\rangle + O(\|h\|^2). \end{aligned}$$

We conclude by invoking (A.1). $\square$

With some care, calculations involving functional derivatives are more efficiently done using the usual rules of calculus in $\mathbb{R}^n$. For instance, the result above is more concisely justified from

$$\begin{aligned} \langle \frac{\delta}{\delta m}\left(\frac{1}{2}\langle \mathcal{F}[m] - d, \mathcal{F}[m] - d\rangle\right), m_1\rangle &= \langle F m_1, \mathcal{F}[m] - d\rangle \\ &= \langle F^*(\mathcal{F}[m] - d), m_1\rangle. \end{aligned}$$

The reader may still wish to use a precise system for bookkeeping the various free and dummy variables for longer calculations – see the appendix for such a system.

The problem of computing F^* will be completely addressed in the next chapter.

The *Gauss-Newton iteration* is Newton's method applied to J:

$$m^{(k+1)} = m^{(k)} - \left(\frac{\delta^2 J}{\delta m^2}[m^{(k)}]\right)^{-1} \frac{\delta J}{\delta m}[m^{(k)}]. \tag{3.12}$$

Here $\left(\frac{\delta^2 J}{\delta m^2}[m^{(k)}]\right)^{-1}$ is an operator: it is the inverse of the functional Hessian of J.

Any iterative scheme based on a local descent direction may converge to a wrong local minimum when J is nonconvex. Gradient descent typically converges slowly – a significant impediment for large-scale problems. The

Gauss-Newton iteration converges faster than gradient descent in the neighborhood of a local minimum, when the Hessian of J is (close to being) positive semi-definite, but may otherwise result in wrong update directions. It is in general much more complicated to set up Gauss-Newton than a gradient descent since the wave-equation Hessian is a large matrix, costly to store and costly to invert. Good practical alternatives include quasi-Newton methods such as LBFGS, which attempt to partially invert the wave-equation Hessian.

3.5 Exercises

1. Repeat the development of section (3.1) in the frequency domain (ω) rather than in time.

2. Derive Born series with a multiscale expansion: write $m = m_0 + \varepsilon m_1$, $u = u_0 + \varepsilon u_1 + \varepsilon^2 u_2 + \ldots$, substitute in the wave equation, and equate like powers of ε. Find the first few equations for u_0, u_1, and u_2.

3. Write the Born series for the acoustic system, i.e., find the linearized equations that the first few terms obey. [Hint: repeat the reasoning of section 3.1 for the acoustic system, or equivalently expand on the first few three bullet points in section 3.2.]

4. At the end of section 3.1 we found the equation that u_1 obeys by differentiating (3.2) with respect to m. Now, differentiate (3.2) twice in two different directions m_1, m_1' to find the equation for the Hessian $\frac{\delta^2 \mathcal{F}}{\delta m_1 \delta m_1'}$, as a bilinear form of two functions m_1 and m_1'. Check that (up to a factor 2) your answer reduces to the equation for u_2 obtained in exercise 2 when $m_1 = m_1'$. The Hessian of $\mathcal{F}$ reappears in the next chapter as we describe accelerated descent methods for the inversion problem.

   Solution. A first derivative with respect to m_1 gives

   $$\frac{\delta m}{\delta m_1} \frac{\partial^2 \mathcal{F}(m)}{\partial t^2} + \left(m \frac{\partial^2}{\partial t^2} - \Delta \right) \frac{\delta \mathcal{F}(m)}{\delta m_1} = 0.$$

   The notation $\frac{\delta m}{\delta m_1}$ means the linear form that takes a function m_1 and returns the operator of multiplication by m_1. We may also write it as the identity I_{m_1} "expecting" a trial function m_1. A second derivative with respect to m_1' gives

$$\frac{\delta m}{\delta m_1}\frac{\partial^2}{\partial t^2}\frac{\delta \mathcal{F}(m)}{\delta m_1'} + \frac{\delta m}{\delta m_1'}\frac{\partial^2}{\partial t^2}\frac{\delta \mathcal{F}(m)}{\delta m_1} + \left(m\frac{\partial^2}{\partial t^2} - \Delta\right)\frac{\delta^2 \mathcal{F}(m)}{\delta m_1 \delta m_1'} = 0.$$

We now evaluate the result at the base point $m = m_0$, and perform the pairing with two trial functions m_1 and m_1'. Denote

$$v = \langle \frac{\delta^2 \mathcal{F}(m_0)}{\delta m_1 \delta m_1'} m_1, m_1' \rangle.$$

Then the equation for v is

$$\left(m_0\frac{\partial^2}{\partial t^2} - \Delta\right) v = -m_1 \frac{\partial^2 u_1'}{\partial t^2} - m_1' \frac{\partial^2 u_1}{\partial t^2},$$

where u_1, u_1' are the respective linearized reflection fields generated by m_1, m_1'. In this formulation, the computation of v requires solving four wave equations, for v, u_1, u_1', and u_0 (which appears in the equations for u_1 and u_1'). Notice that $v = 2u_2$ when $m_1 = m_1'$.

5. Compute $\frac{\delta^2 \mathcal{F}}{\delta m^2}$ in an alternative way by polarization: find the equations for the second-order field u_2 when the respective model perturbations are $m_1 + m_1'$ and $m_1 - m_1'$, and take a combination of those two fields.

6. Consider the setting of section 3.2 in the case $M = I$. No perturbation will be needed for this exercise (no decomposition of M into $M_0 + \varepsilon M_1$). Prove the following energy estimate for the solution of (3.8):

$$E(t) \leq \left(\int_0^t \|f\|(s)\, ds\right)^2, \tag{3.13}$$

where $E(t) = \langle w, Mw \rangle$ and $\|f\|^2 = \langle f, f \rangle$. [Hint: repeat and adapt the beginning of the proof of theorem 3.]

7. Consider (3.8) and (3.9) in the special case when $M_0 = I$. Let $\|w\| = \sqrt{\langle w, w \rangle}$ and $\|w\|_* = \max_{0 \leq t \leq T} \|w\|$. In this exercise we show that $w - w_0 = O(\varepsilon)$, and that $w - w_0 - w_1 = O(\varepsilon^2)$.

(a) Find an equation for $w - w_0$. Prove that

$$\|w - w_0\|_* \leq \varepsilon \|M_1\|_\infty \Omega T \|w\|_*$$

[Hint: repeat and adapt the proof of theorem 3.]

(b) Find a similar inequality to control the time derivative of $w - w_0$.

(c) Find an equation for $w - w_0 - w_1$. Prove that

$$\|w - w_0 - w_1\|_* \leq (\varepsilon \|M_1\|_\infty \Omega T)^2 \|w\|_*$$

8. Consider the gradient descent method applied to the linear least-squares problem $\min_x \|Ax - b\|_2$. Show that

$$\alpha = \frac{1}{\|A^* A\|}$$

is a safe choice in the sense that the resulting gradient step is a contraction, i.e., the distance between successive iterates decreases monotonically.

9. Consider $J(m)$ any smooth, locally convex function of m.

(a) Show that the specific choice

$$\alpha = \frac{\langle \frac{\delta J}{\delta m}[m^{(k)}], \frac{\delta J}{\delta m}[m^{(k)}] \rangle}{\langle \frac{\delta J}{\delta m}[m^{(k)}], \frac{\delta J^2}{\delta m^2}[m^{(k)}] \, \frac{\delta J}{\delta m}[m^{(k)}] \rangle}$$

for the gradient descent method results from approximating J by a quadratic function *in the direction of* $\delta J/\delta m$, near $m^{(k)}$, and finding the minimum of that quadratic function.

(b) Show that the Gauss-Newton iteration (3.12) results from approximating J by a quadratic near $m^{(k)}$, and finding the minimum of that quadratic function.

10. Prove the following formula for the wave-equation Hessian $\frac{\delta^2 J}{\delta m_1 \delta m_1'}$ in terms of F and its functional derivatives:

$$\frac{\delta^2 J}{\delta m_1 \delta m_1'} = F^* F + \langle \frac{\delta^2 \mathcal{F}}{\delta m_1 \delta m_1'}, \mathcal{F}[m] - d \rangle. \tag{3.14}$$

Note: F^*F is called the normal operator.

Solution. To compute Hessians, it is important to expand the notation to keep track of the different variables, i.e., we compute $\frac{\delta^2 J}{\delta m_1 \delta m_1'}$. A first derivative gives

$$\frac{\delta J}{\delta m_1} = \langle \frac{\delta \mathcal{F}(m)}{\delta m_1}, \mathcal{F}(m) - d \rangle,$$

where the inner product bears on $\mathcal{F}$ in each factor. A second derivative gives

$$\frac{\delta^2 J}{\delta m_1 \delta m_1'} = \langle \frac{\delta \mathcal{F}(m)}{\delta m_1}, \frac{\delta \mathcal{F}(m)}{\delta m_1'} \rangle + \langle \frac{\delta^2 \mathcal{F}(m)}{\delta m_1 \delta m_1'}, \mathcal{F}(m) - d \rangle.$$

This result is then evaluated at the base point $m = m_0$, where $\frac{\delta \mathcal{F}(m_0)}{\delta m_1} = F$. The second term in the right-hand side already has the desired form. The first term in the right-hand-side, when paired with m_1 and m_1', gives

$$\langle F m_1, F m_1' \rangle = \langle F^* F m_1, m_1' \rangle,$$

hence it can be seen as F^*F, turned into a bilinear form by application to m_1 and inner product with m_1'. Notice that, if we pair the whole equation with m_1 and m_1', and evaluate at $m = m_0$, we arrive at the elegant expression.

$$\langle \frac{\delta^2 J}{\delta m_1 \delta m_1'} m_1, m_1' \rangle = \langle u_1, u_1' \rangle + \langle v, u_0 - d \rangle, \tag{3.15}$$

where v was defined in the solution of an earlier exercise as

$$v = \langle \frac{\delta^2 \mathcal{F}(m_0)}{\delta m_1 \delta m_1'} m_1, m_1' \rangle.$$

11. Show that the spectral radius of the Hessian operator $\frac{\delta^2 J}{\delta m^2}$, when data are (essentially) limited by $t \leq T$ and $\omega \leq \Omega$, is bounded by a constant times $(\Omega T)^2$.