Full Waveform Inversion — Personal Notes

Jayin Khanna

The Fréchet Derivative – A pre-req

Let $g : M \to D$ be a **nonlinear operator** between two normed spaces.

Let $m \in M$. Then g is **Fréchet differentiable** at m if there exists a *bounded linear operator* $J : M \to D$ such that

$$\lim_{\|h\|_M \to 0} \frac{\|g(m+h) - g(m) - Jh\|_D}{\|h\|_M} = 0.$$

Intuition: Fréchet Derivative

J is the *best linear approximation* to the change in g near m. Exactly like the total derivative in finite dimensions, but now the "perturbation" h and the "output" $g(\cdot)$ both live in potentially infinite-dimensional spaces. The condition says the error of the linear approximation is *sub-linear* in $\|h\|$.

FWI Case

Subsurface Domain. $\Omega \subset \mathbb{R}^2$ is the subsurface domain:

$$\Omega = \big\{(x, z) : 0 \le x \le X,\ 0 \le z \le Z\big\},$$

where x is the horizontal coordinate and z is depth.

Example

OpenFWI uses a 70×70 grid.

Function space for velocity. $L^2(\Omega)$ is the set of all functions $f : \Omega \to \mathbb{R}$ with

$$\|f\|^2_{L^2(\Omega)} = \int_\Omega |f(x, z)|^2 \, dx \, dz < \infty \qquad \text{(finite energy)},$$

equipped with the inner product

$$\langle f,\, g\rangle_{L^2(\Omega)} = \int_\Omega f(x, z)\, g(x, z) \, dx \, dz.$$

$L^2(\Omega)$ is a **Hilbert space** (complete inner product space). The completeness (every Cauchy sequence converges in L^2) follows from the Riesz–Fischer theorem. *Open question from notes:* Can the structure of this inner product — in particular Riesz representation — restrict or constrain the gradient descent in FWI? (Hint: yes — the choice of inner product determines what "gradient" means via the Riesz representation theorem, which is exactly why the adjoint-state method works.)

Velocity model. The velocity model $v(x, z)$ lives in $L^2(\Omega)$ — it is a function that assigns a wave speed to every point of the subsurface.

Since v is bounded and Ω is bounded,

$$v \in L^\infty(\Omega) \subset L^2(\Omega).$$

Data Space

$L^2([0, T] \times \Gamma)$ is the space of all seismograms $d : [0, T] \times \Gamma \to \mathbb{R}$, where Γ is the set of receiver locations, with

$$\|d\|^2 = \int_0^T \sum_{r \in \Gamma} |d(t, r)|^2 \, dt < \infty,$$

and inner product

$$\langle d_1, \, d_2 \rangle = \int_0^T \sum_{r \in \Gamma} d_1(t, r) \, d_2(t, r) \, dt.$$

What is a seismogram?

At each receiver location $r \in \Gamma$, the sensor records the pressure $d(t, r)$ as a function of time $t \in [0, T]$. The full collection of recordings across all receivers and all times is a single element of $L^2([0, T] \times \Gamma)$.

Terms in FWI

1. **Velocity model:** $v \in L^2(\Omega)$, with $v(x, z) > 0$.
2. **Pressure wavefield:** $p : \Omega \times [0, T] \to \mathbb{R}$; the solution to the forward (wave) problem.
3. **Source term:** $s : \Omega \times [0, T] \to \mathbb{R}$, with $s \in L^2(\Omega \times [0, T])$, given by

$$s(x, z, t) = w(t) \cdot \delta_\varepsilon(x - x_s, \, z - z_s),$$

where (x_s, z_s) is the **source location** and $w(t)$ is the **Ricker wavelet** (a band-limited pulse used in seismic acquisition).

The Ricker (Mexican hat) wavelet $w(t) = A\big(1 - 2\pi^2 f_0^2 t^2\big)e^{-\pi^2 f_0^2 t^2}$ is the second derivative of a Gaussian. It is the standard source signature in synthetic seismic experiments. Key question: *What is the bandwidth / spectral content?* How does the dominant frequency f_0 affect the resolution of the reconstructed velocity model?

4. **Observed seismogram:** $d^{\text{obs}} \in L^2([0,T] \times \Gamma)$, the data recorded in the field.

The Forward Problem

Given v, **find** p. The initial-boundary value problem (IBVP) is

$$\boxed{\nabla^2 p - \frac{1}{v(x,z)^2}\,\partial_{tt} p = s(x,z,t) \qquad \text{on } \Omega \times [0,T],}$$

with initial conditions (simplest form):

$$p(x,z,0) = 0, \qquad \partial_t p(x,z,0) = 0 \quad \text{on } \Omega,$$

and **absorbing (PML) boundary condition**:

$$\mathcal{B}[p] = 0 \quad \text{on } \partial\Omega \times [0,T].$$

Apparent contradiction: the PDE is linear (in p), yet the forward operator g is nonlinear (in v). Resolution:
For *fixed* v, the coefficient $\frac{1}{v(x,z)^2}$ is a known function, so the PDE is linear in the single unknown p.
However, $g(v) = d^{\text{obs}}$ depends on v *nonlinearly*: if you double v, the solution p does not double. Concretely, $g(v) = R \circ S(v)$ where $S(v) = p$ solves the PDE with coefficient v^{-2}; the map $v \mapsto v^{-2}$ is already nonlinear before we even solve.
Well-posedness (to look up): For fixed $v > 0$ and $s \in L^2(\Omega \times [0,T])$, the IBVP has a unique solution p depending continuously on v and s.

The Operators

Three operators organise the FWI pipeline:

$$S : L^2(\Omega) \longrightarrow L^2(\Omega \times [0,T]), \quad v \longmapsto p \qquad \text{(solution operator)}$$

$$R : L^2(\Omega \times [0,T]) \longrightarrow L^2([0,T] \times \Gamma), \quad p \longmapsto p\big|_{[0,T]\times\Gamma} \qquad \text{(restriction operator)}$$

$$g = R \circ S \qquad \text{(forward operator)}$$

Operator pipeline

S solves the wave PDE to get the full pressure field everywhere in the domain. R then reads off only what the receivers actually see — the restriction to $[0, T] \times \Gamma$. Their composition g maps a velocity model directly to a predicted seismogram.

The Optimization Problem

Setup. The forward map g is known but the true velocity v is not. What we have:

1. d^{obs} — the seismogram recorded by the receivers.
2. g — the forward operator.

FWI asks: *which v would have produced d^{obs}?* That is, find v such that $g(v) = d^{\mathrm{obs}}$.

Misfit functional. Since we cannot invert g directly, we instead minimise a data-fit criterion. Closeness is measured in the L^2 norm on data space:

$$\mathcal{X}(v) = \frac{1}{2} \left\| g(v) - d^{\mathrm{obs}} \right\|^2_{L^2([0,T]\times\Gamma)}$$

This is the **misfit functional**: it measures how far the predicted seismogram $g(v)$ is from the observed seismogram d^{obs} for any candidate model v. The FWI optimisation problem is then

$$\min_{v \in L^2(\Omega)} \mathcal{X}(v), \qquad \mathcal{X} : L^2(\Omega) \to \mathbb{R}.$$

The L^2 norm is standard but has known weaknesses — in particular it is sensitive to cycle-skipping (phase shifts between predicted and observed waveforms can create spurious local minima). Alternatives actively studied:

- **Optimal transport / Wasserstein distance**: treats seismograms as measures and compares their mass distributions geometrically. Potentially convex in regimes where L^2 is not.
- **Envelope / instantaneous-amplitude misfit**: less sensitive to phase.
- **WGANs**: use a learned critic as a surrogate loss.

Key research question: *Which difference metric on data space is most suited to FWI?* Loss function analysis (comparing landscapes, convexity, gradient behaviour) is an active direction. The phrase "set matching" in the notes also points toward optimal transport as a framing.

Gradient descent update. To minimise $\mathcal{X}$ we would like to run

$$v_{k+1} = v_k - \alpha_k \nabla_v \mathcal{X},$$

but v is a *function*, not a vector, so $\nabla_v \mathcal{X}$ cannot be written as a finite list of partial derivatives. This forces us to work with the Fréchet derivative. Three things are needed:

- **(i)** A notion of **differentiating $\mathcal{X}$ w.r.t.** v, where v is a function not a vector.
- **(ii)** A **formula for the gradient** $\nabla_v \mathcal{X}$ in terms of g and S.
- **(iii)** A **computationally feasible way** to evaluate this gradient without explicitly forming the Jacobian J.

Why We Cannot Simply Compute $g^{-1}(d^{\text{obs}})$

Three fundamental obstacles:

1. **g is nonlinear.**

   $g = R \circ S$ involves solving a PDE with v as a coefficient. Nonlinear operators generally do not admit clean closed-form inverses.

2. **g is not injective.**

   Multiple distinct velocity models $v_1 \neq v_2$ can produce the same surface seismogram:

   $$g(v_1) = g(v_2).$$

   Physically: receivers only see the wavefield at the surface; many different subsurface structures are consistent with the same surface data. Hence g^{-1} does not exist as a function.

3. **d^{obs} is noisy.**

   Even if a true model v^{true} exists with $g(v^{\text{true}}) = d^{\text{obs}}$, the measured data contain irreducible noise:

   $$d^{\text{obs}} = g(v^{\text{true}}) + \varepsilon,$$

   so exact inversion would fit the noise, not the signal.

Why optimisation instead of inversion

The three obstacles above — nonlinearity, non-injectivity, noise — are the standard ill-posed inverse problem triad (Hadamard's conditions fail). Reformulating as an optimisation over the misfit $\mathcal{X}(v)$ side-steps the need for a closed-form inverse: we search for the "best" v in a regularised sense.

Fréchet Differentiability of $\mathcal{X}$

Finite-dimensional warm-up. For $f : \mathbb{R}^n \to \mathbb{R}$ at a point $v \in \mathbb{R}^n$, differentiability means finding the linear map that best approximates the change in f near v:

$$f(v+h) = f(v) + \nabla f(v)^T h + O(\|h\|).$$

Generalising to $f : \mathbb{R}^n \to \mathbb{R}$ at $v \in \mathbb{R}^n$:

$$f(v+h) = f(v) + \nabla f(v)^T h + O(\|h\|),$$

where $\nabla f(v)^T h$ is the hyperplane describing the first-order change — it is a bounded linear functional on $\mathbb{R}^n$.

Moving to FWI: $\mathcal{X} : L^2(\Omega) \to \mathbb{R}$. In FWI, $v \in L^2(\Omega)$ is a *function*, not a vector. There are uncountably many directions $\delta v \in L^2(\Omega)$ one can perturb v in, so $\nabla_v \mathcal{X}$ cannot be written as a vector of partial derivatives.

Definition. $\mathcal{X}$ is *Fréchet differentiable* at v if there exists a bounded linear functional $D_{\mathcal{X}}[v] : L^2(\Omega) \to \mathbb{R}$ such that

$$\boxed{\mathcal{X}(v+\delta v) = \mathcal{X}(v) + D_{\mathcal{X}}[v](\delta v) + O(\|\delta v\|).}$$

Geometric picture

Just as in finite dimensions the gradient $\nabla f(v)^T h$ defines a tangent hyperplane at v, here $D_{\mathcal{X}}[v](\delta v)$ defines a tangent "hyperplane" in function space: the unique bounded linear functional that best approximates how $\mathcal{X}$ changes as you move away from v. By the Riesz representation theorem for Hilbert spaces, every such bounded linear functional on $L^2(\Omega)$ can be written as an inner product

$$D_{\mathcal{X}}[v](\delta v) = \langle \nabla_v \mathcal{X},\, \delta v \rangle_{L^2(\Omega)},$$

and $\nabla_v \mathcal{X} \in L^2(\Omega)$ is what we call the *gradient* of $\mathcal{X}$ at v. This is precisely the object the adjoint-state method computes.

Fréchet Differentiability of g

Since $\mathcal{X}(v)$ depends on v only through $g(v)$, computing $D_{\mathcal{X}}[v]$ requires first knowing how $g(v)$ changes when $v \mapsto v + \delta v$.

g is Fréchet differentiable at v if there exists a bounded linear operator

$$J : L^2(\Omega) \longrightarrow L^2([0,T] \times \Gamma)$$

such that

$$g(v+\delta v) = g(v) + J\,\delta v + O(\|\delta v\|).$$

$J\,\delta v$ is the **first-order change in the predicted seismogram** caused by the velocity perturbation δv.

Deriving J: the Linearised Wave Equation

Perturbing the PDE coefficient. Perturbing $v \to v + \delta v$ changes the PDE coefficient:

$$\frac{1}{v^2} \longrightarrow \frac{1}{(v+\delta v)^2} \approx \frac{1}{v^2} - \frac{2\,\delta v}{v^3} + O(\delta v^2).$$

This changes the wavefield solution $p \to p + \delta p$. To first order in δv, the perturbation δp satisfies the **linearised wave equation**:

$$\boxed{\nabla^2(\delta p) - \frac{1}{v^2}\,\partial_{tt}(\delta p) = -\frac{2\,\delta v}{v^3}\,\partial_{tt} p \qquad \text{on } \Omega \times [0,T],}$$

with zero initial conditions (since the unperturbed p already satisfies the ICs), and the same absorbing boundary conditions as before.

What the linearised equation says

The right-hand side $-\frac{2\,\delta v}{v^3}\,\partial_{tt} p$ acts as a *secondary source*: the perturbation in velocity scatters the existing wavefield p, generating a new wavefield δp that propagates through the same medium. This is the Born approximation at the PDE level.

The Jacobian action. Once δp is solved from the linearised equation above, the action of J on δv is simply the restriction of δp to the receivers:

$$J\,\delta v = \delta p\big|_{[0,T]\times\Gamma}.$$

Why J Cannot Be Stored or Applied Directly

J maps $L^2(\Omega) \to L^2([0,T]\times\Gamma)$. In the discrete OpenFWI setting:

- Velocity grid: $70 \times 70 = 4{,}900$ unknowns (input dimension).
- Seismogram space: $5 \times 1000 \times 70 = 350{,}000$ entries (5 sources, 1000 time steps, 70 receivers).
- The matrix J therefore has $350{,}000 \times 4{,}900 \approx 1.7 \times 10^9$ entries.

This matrix **cannot be stored**, let alone applied to every possible δv. A fundamentally different approach is needed to compute $\nabla_v \mathcal{X}$ — this is precisely what the **adjoint-state method** provides.

The key insight is that $\nabla_v \mathcal{X}$ is a single element of $L^2(\Omega)$ — we never need all columns of J separately. By the chain rule,

$$D_{\mathcal{X}}[v](\delta v) = \langle g(v) - d^{\mathrm{obs}},\, J\,\delta v\rangle_{L^2([0,T]\times\Gamma)}.$$

Writing J^* for the L^2-adjoint of J, this equals $\langle J^*(g(v)-d^{\mathrm{obs}}),\, \delta v\rangle_{L^2(\Omega)}$, so $\nabla_v \mathcal{X} = J^*(g(v)-d^{\mathrm{obs}})$. Computing J^* applied to the data residual requires solving one *backward-in-time* (adjoint) wave equation — the same cost as one forward solve, and no matrix storage at all.