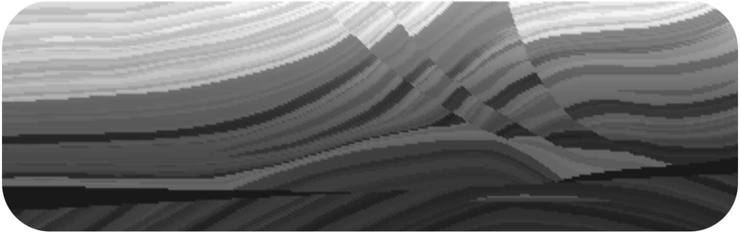
Inner product preconditioned optimization methods for full waveform inversion



X . Adriaens[1], L. Métivier[2] and C. Geuzaine[1]

[1]Université de Liège
[2]Université Grenoble Alpes

March 2022

Consider

- a model parameter m
- a wave propagation operator F
- a wavefield u
- a mesurement operator R
- a dataset d

Full wave inversion consists in finding m^* such that

$$R(u) = d \text{ with } F(u, m^*) = f$$

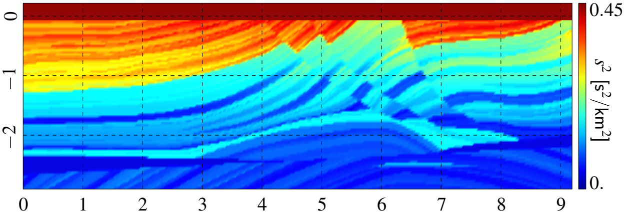
through the **optimization problem**

$$m^* = \arg\min_m J(m) \triangleq \arg\min_m \operatorname{dist}(R(u(m)), d)$$

The distribution of the **slowness squared** is here chosen to be the unknown model, *i.e.*

$$m \triangleq s^2(\boldsymbol{x}) = 1/v^2(\boldsymbol{x}).$$

The **Marmousi model**[1] is a typical example of distributions that are sought in the context of geophysics.



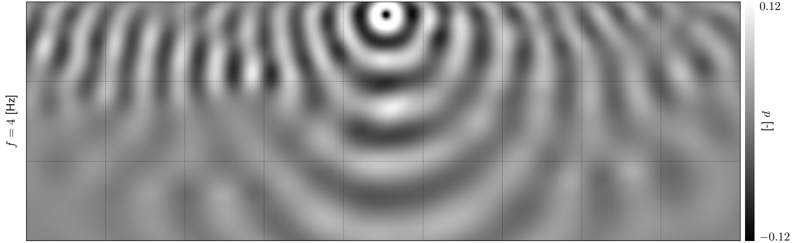
[1]Versteeg, "The Marmousi experience: Velocity model determination on a synthetic complex data set".

In acoustics, the wavefield is a **pressure field**, *i.e.*

$$u \triangleq p(\boldsymbol{x})$$

whose propagation can be modelled by the **Helmholtz equation**, *i.e.*

$$F(u,m) \triangleq \Delta p + \omega^2 s^2 p$$

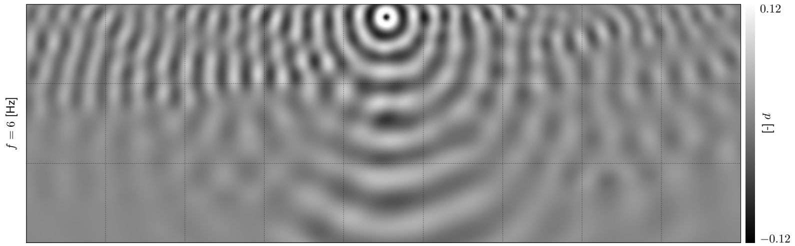


In acoustics, the wavefield is a **pressure field**, *i.e.*

$$u \triangleq p(\boldsymbol{x})$$

whose propagation can be modelled by the **Helmholtz equation**, *i.e.*

$$F(u, m) \triangleq \Delta p + \omega^2 s^2 p$$

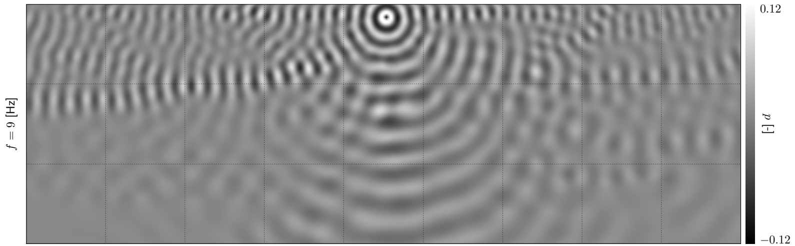


In acoustics, the wavefield is a **pressure field**, *i.e.*

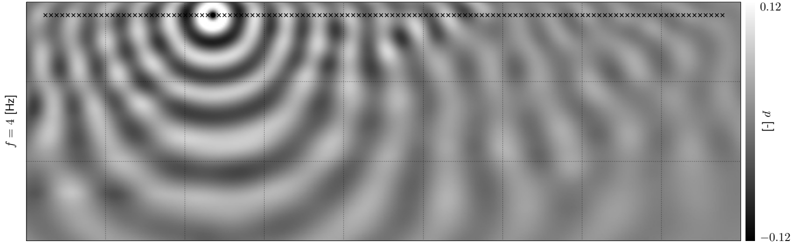
$$u \triangleq p(\boldsymbol{x})$$

whose propagation can be modelled by the **Helmholtz equation**, *i.e.*

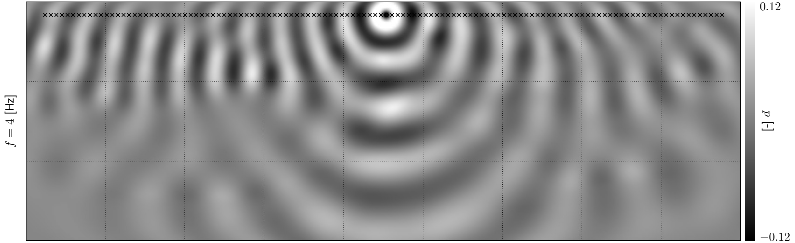
$$F(u, m) \triangleq \Delta p + \omega^2 s^2 p$$



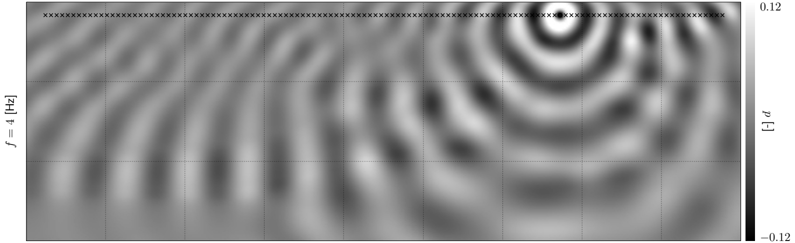
Several (n_{er}) emitters-receivers lie on the surface of the model. Each e-r is successively excited ($\bullet$) and the response ($\times$) is recorded at all the other e-r, for several frequencies (n_ω).



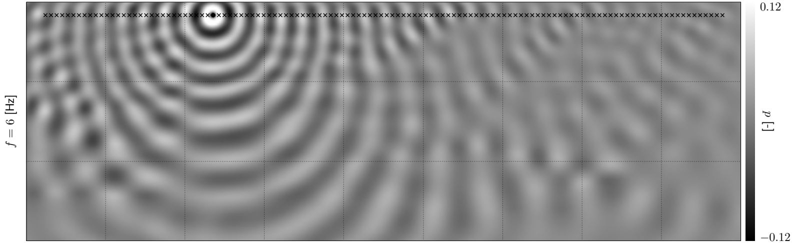
Several (n_{er}) emitters-receivers lie on the surface of the model. Each e-r is successively excited ($\bullet$) and the response ($\times$) is recorded at all the other e-r, for several frequencies (n_ω).



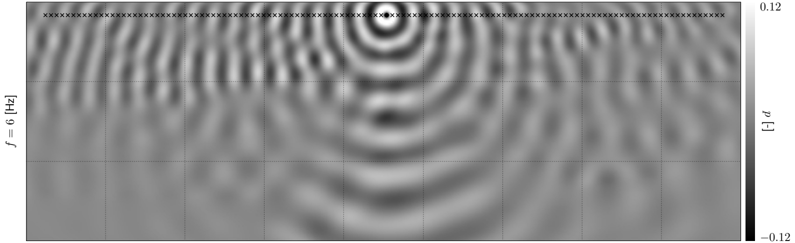
Several (n_{er}) emitters-receivers lie on the surface of the model. Each e-r is successively excited ($\bullet$) and the response ($\times$) is recorded at all the other e-r, for several frequencies (n_ω).



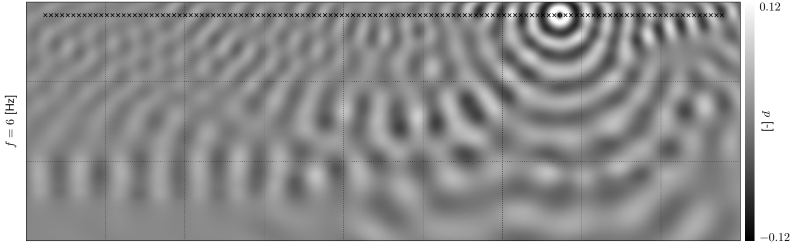
Several (n_{er}) emitters-receivers lie on the surface of the model. Each e-r is successively excited (•) and the response ($\times$) is recorded at all the other e-r, for several frequencies (n_{ω}).



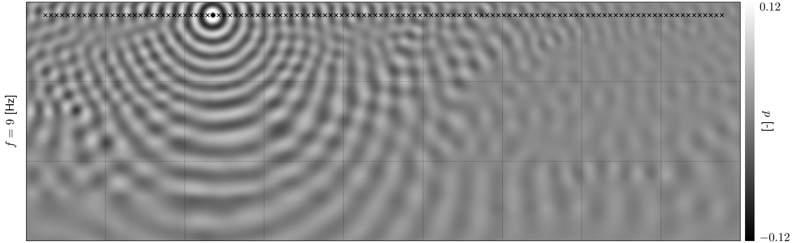
Several (n_{er}) emitters-receivers lie on the surface of the model. Each e-r is successively excited ($\bullet$) and the response ($\times$) is recorded at all the other e-r, for several frequencies (n_ω).



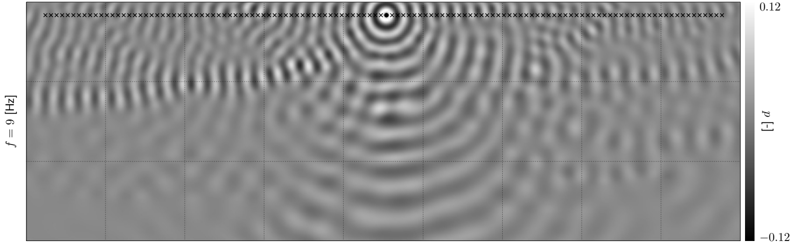
Several (n_{er}) emitters-receivers lie on the surface of the model. Each e-r is successively excited ($\bullet$) and the response ($\times$) is recorded at all the other e-r, for several frequencies (n_ω).



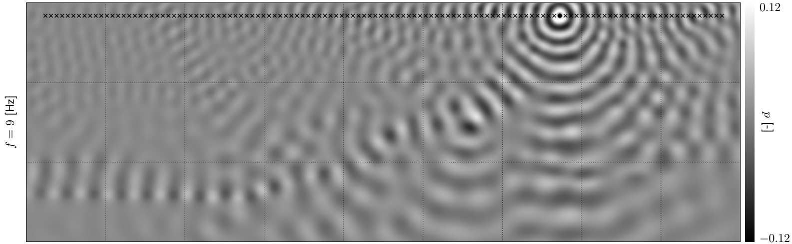
Several (n_{er}) emitters-receivers lie on the surface of the model. Each e-r is successively excited ($\bullet$) and the response ($\times$) is recorded at all the other e-r, for several frequencies (n_ω).



Several (n_{er}) emitters-receivers lie on the surface of the model. Each e-r is successively excited ($\bullet$) and the response ($\times$) is recorded at all the other e-r, for several frequencies (n_ω).



Several (n_{er}) emitters-receivers lie on the surface of the model. Each e-r is successively excited ($\bullet$) and the response ($\times$) is recorded at all the other e-r, for several frequencies (n_ω).

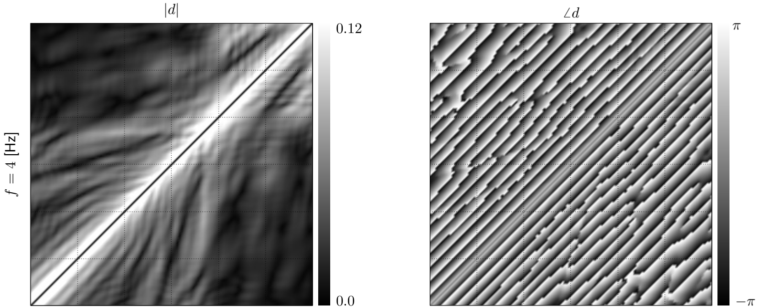


A dataset d is thus a $n_s \times n_r \times n_\omega$ **complex-valued matrix**, *i.e*

$$d \in \mathbb{C}^{n_s \times n_r \times n_\omega}$$

which can be obtained by **point-wise** measurements at the receivers, *i.e.*

$$[R(u)]_{s,r,\omega} \triangleq \int p_{s,\omega}(\boldsymbol{x})\delta(\boldsymbol{x} - \boldsymbol{x}_r)\, d\boldsymbol{x} = p_{s,\omega}(\boldsymbol{x}_r).$$

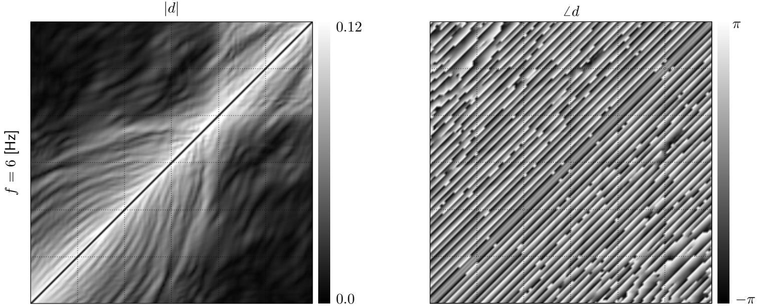


A dataset d is thus a $n_s \times n_r \times n_\omega$ **complex-valued matrix**, *i.e*

$$d \in \mathbb{C}^{n_s \times n_r \times n_\omega}$$

which can be obtained by **point-wise** measurements at the receivers, *i.e.*

$$[R(u)]_{s,r,\omega} \triangleq \int p_{s,\omega}(\boldsymbol{x})\delta(\boldsymbol{x} - \boldsymbol{x}_r)\, d\boldsymbol{x} = p_{s,\omega}(\boldsymbol{x}_r).$$



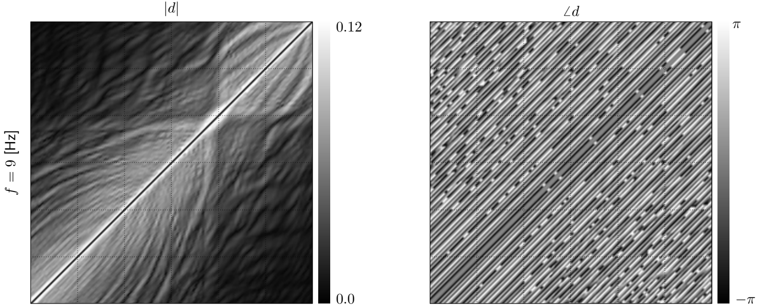
Problem statement: data space II

A dataset d is thus a $n_s \times n_r \times n_\omega$ **complex-valued matrix**, *i.e*

$$d \in \mathbb{C}^{n_s \times n_r \times n_\omega}$$

which can be obtained by **point-wise** measurements at the receivers, *i.e.*

$$[R(u)]_{s,r,\omega} \triangleq \int p_{s,\omega}(\boldsymbol{x})\delta(\boldsymbol{x}-\boldsymbol{x}_r)\,d\boldsymbol{x} = p_{s,\omega}(\boldsymbol{x}_r).$$



Full waveform inversion relies on the solution of an optimization problem

$$m^* = \arg\min_m J(m)$$

Local optimization techniques are used because the search space is typically large

Local optimization techniques originate from a second order expansions of the misfit

$$\begin{aligned} J(m+\delta m) &\approx J(m) + \{D_m J\}(\delta m) + \frac{1}{2}\{D^2_{mm} J\}(\delta m, \delta m) \\ &\approx J(m) + \langle j', \delta m \rangle_M \quad + \frac{1}{2}\langle H\delta m, \delta m \rangle_M \end{aligned}$$

provided some inner product $\langle , \rangle_M$ is chosen for the model space

Full waveform inversion relies on the solution of an optimization problem

$$m^* = \arg\min_m J(m)$$

Local optimization techniques are used because the search space is typically large

Based on this expansion, the descent direction p is then chosen as

$$p_N = \arg\min J(m) + \langle j', p \rangle_M + \frac{1}{2} \left\langle \tilde{H} p, p \right\rangle_M$$

or equivalently

$$\tilde{H} p_N = -j'$$

for some approximate Hessian operator $\tilde{H}$.

- $\tilde{H} \approx I$ (steepest descent)
- $\tilde{H} \approx B$ (Broyden-Fletcher-Goldfarb-Shanno method)
- $\tilde{H} \approx H_{(\mathrm{GN})}$ ((Gauss-)Newton conjugate gradient method)

Local optimization methods

Full waveform inversion relies on the solution of an optimization problem

$$m^* = \arg\min_m J(m)$$

Local optimization techniques are used because the search space is typically large

In addition, a strategy for scaling this descent direction must be chosen.
Such strategies ensure convergence towards the nearest local minimum.

- Line search: $m = m + \gamma p$

  with $p = -\tilde{H}^{-1} j'$ and $\gamma \approx \arg\min J(m + \gamma p)$

- Trust region: $m = m + p$

  with $p = \arg\min_p J(m) + \langle j', p \rangle_M + \frac{1}{2} \left\langle \tilde{H} p, p \right\rangle_M$ and $\|p\|_M \leq \Delta$

Local optimization methods

Full waveform inversion relies on the solution of an optimization problem

$$m^* = \operatorname{arg\,min}_m J(m)$$

Local optimization techniques are used because the search space is typically large

In summary, a local optimization procedure requires three ingredients

1. A globalization strategy to control their lengths
2. A method to compute descent directions
3. An inner product for the model space

Local optimization techniques are based on a local misfit expansion

$$\begin{aligned} J(m+\delta m) &\approx J(m) + \{D_m J\}(\delta m) + \frac{1}{2}\{D^2_{mm} J\}(\delta m, \delta m) \\ &\approx J(m) + \left\langle j', \delta m \right\rangle_M \quad + \frac{1}{2}\left\langle H\delta m, \delta m \right\rangle_M \end{aligned}$$

Equivalence between both expansions is granted by the gradient j' and the Hessian operator H defining property

$$\left\langle j', \delta m \right\rangle_M \triangleq \{D_m J\}(\delta m),\ \forall \delta m$$

and

$$\left\langle H\delta m_1, \delta m_2 \right\rangle_M \triangleq \{D^2_{mm} J\}(\delta m_1, \delta m_2),\ \forall \delta m_1\, \forall \delta m_2$$

that **strongly depend on the chosen inner product** $\langle \cdot, \cdot \rangle_M$. Changing the inner product therefore modify both the gradient and the Hessian operator.

Local optimization techniques are based on a local misfit expansion

$$
\begin{aligned}
J(m+\delta m) &\approx J(m) + \{D_m J\}(\delta m) + \frac{1}{2}\{D^2_{mm} J\}(\delta m, \delta m) \\
&\approx J(m) + \left\langle j', \delta m \right\rangle_M \quad + \frac{1}{2}\left\langle H\delta m, \delta m \right\rangle_M
\end{aligned}
$$

By transitivity, the link with the conventional (L_2) inner product ($\langle \cdot, \cdot \rangle$) is straightforward

$$
\left\langle j', \delta m \right\rangle_M = \left\langle j'_{L_2}, \delta m \right\rangle, \ \forall \delta m
$$

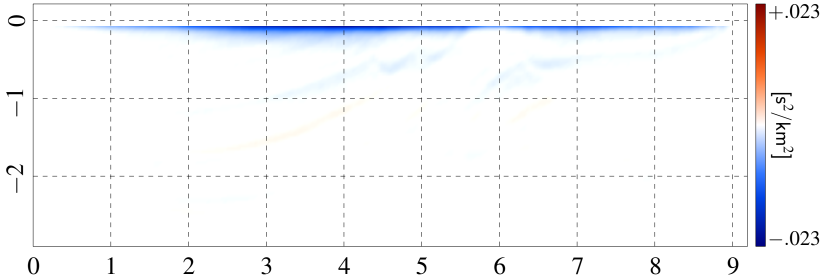
and

$$
\left\langle H\delta m_1, \delta m_2 \right\rangle_M = \left\langle H_{L_2}\delta m_1, \delta m_2 \right\rangle \ \forall \delta m_1 \delta m_2
$$

The conventional choice is a least squares inner product.

$$\langle m_2, m_1 \rangle_M = \langle m_2, m_1 \rangle := \int_\Omega m_1(\boldsymbol{x}) m_2(\boldsymbol{x}) \; d\Omega$$

$$\Rightarrow j' = j'_{L_2}$$

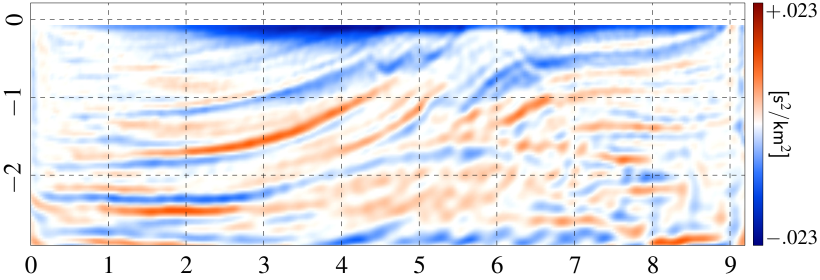


Balance between shallow and deep contributions is broken.

An appropriate **spatial weight** $w(\boldsymbol{x})$ is often applied, *e.g.* the diagonal of the Gauss-Newton Hessian[2].

$$\langle m_1, m_2 \rangle_M = \langle \sqrt{w}\, m_1, \sqrt{w}\, m_2 \rangle$$

$$\Rightarrow j' = w^{-1} j'_{L_2}$$



This inner product choice restores balance between gradient contributions everywhere.

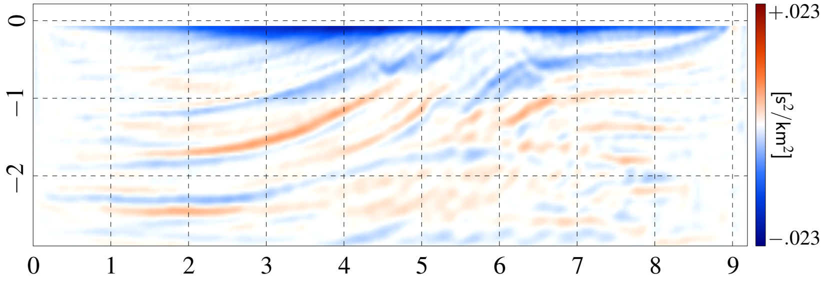
[2] Pan, Innanen, and Liao, "Accelerating Hessian-free Gauss-Newton full-waveform inversion via l-BFGS preconditioned conjugate-gradient algorithm".

Inner product: example III: scaled and thresholded

The diagonal of the Gauss-Newton Hessian can be close to zero.
Therefore a threshold ϵ is added to prevent instabilities.

$$\langle m_1, m_2 \rangle_M = \langle \sqrt{w}\, m_1, \sqrt{w}\, m_2 \rangle + \epsilon \langle m_1, m_2 \rangle$$

$$\Rightarrow j' = (w + \epsilon)^{-1} j'_{L_2}$$

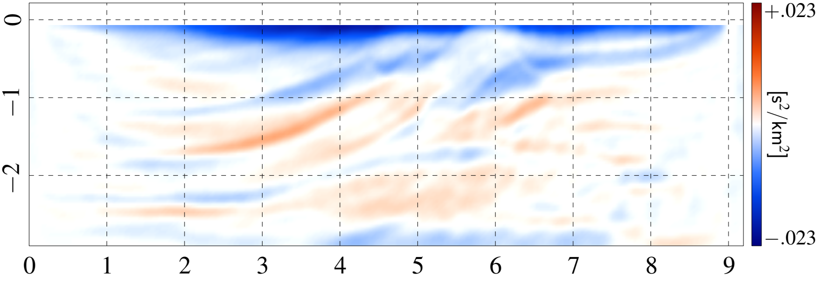


Boundary and corner contributions are silenced.

A stabilization term penalizing rough models can also be added

$$\langle m_1, m_2 \rangle_M = \langle \sqrt{w}\, m_1, \sqrt{w}\, m_2 \rangle + \epsilon l_c^2 \langle \boldsymbol{\nabla} m_1, \boldsymbol{\nabla} m_2 \rangle$$

$$\Rightarrow j' = (w - \epsilon l_c^2 \Delta)^{-1} j'_{L_2}$$



Gradient w.r.t this inner product are therefore smoother.
Encouraging smooth updates early in the inversion process is a strategy to avoid local minima trapping[3].

[3] Zuberi and Pratt, "Mitigating nonlinearity in full waveform inversion using scaled-Sobolev pre-conditioning".

In general, any inner product that can be expressed through some preconditioner P

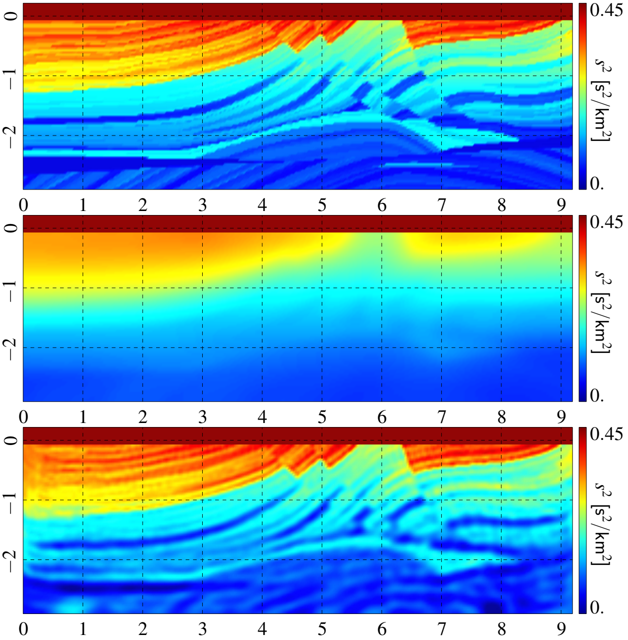
$$\langle m_1, m_2 \rangle_M = \langle P\, m_1, m_2 \rangle$$

yields a preconditioned gradient and a preconditioned Hessian operator

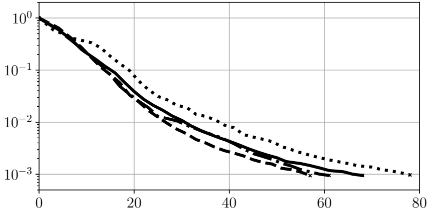
$$j' = P^{-1} j'_{L_2} \quad \text{and} \quad H = P^{-1} H_{L_2}$$

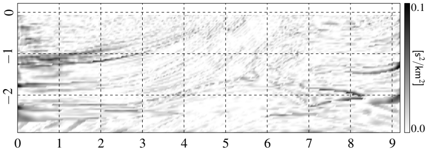
Changing the inner product

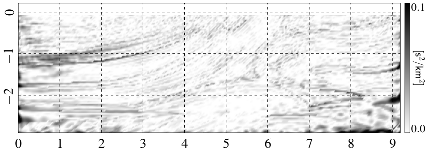
- is formally equivalent to preconditioning
- modifies lengths in the model space
- is mathematically rigorous (and elegant (?))
- makes preconditioning nearly invisible inside the optimization algorithms


0
−1
−2
0 1 2 3 4 5 6 7 8 9
0.45
s^2 [s^2/km^2]
0.
0
−1
−2
0 1 2 3 4 5 6 7 8 9
0.45
s^2 [s^2/km^2]
0.
0
−1
−2
0 1 2 3 4 5 6 7 8 9
0.45
s^2 [s^2/km^2]
0.

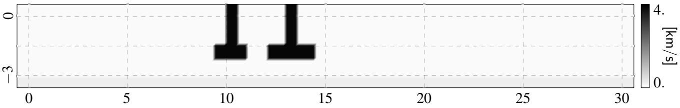
		Wave sol. (tot)	Error rms ([s^2/km^2])
Conventional		78	0.0174
Weighted	only	61	0.0202
	and thresholded	57	0.0174
	and smoothed	68	0.0173



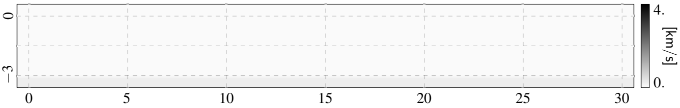




Model is composed of two close T-shaped structures and a bottom reflector.
Non negligible multiple scattering between them.



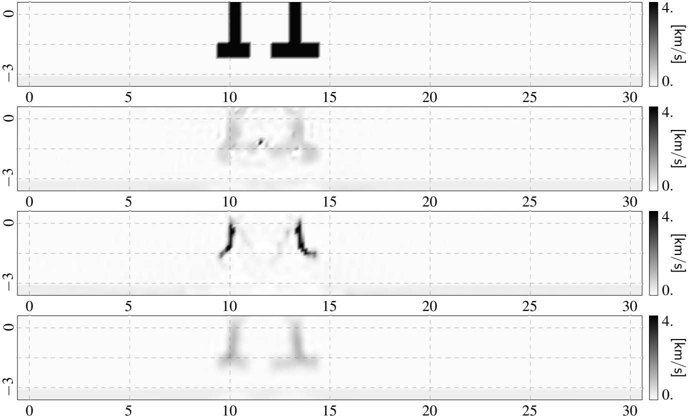
Initial model is an empty background.



unscaled

scaled

scaled & smoothed



Scaled and smoothed inner product only reaches a minimizer close to the true model.

Conclusions

- Selecting the inner product appropriately accelerates the convergence.
- More robust inversion path are obtain with preconditioning.

Perspectives

- Inner product preconditioning is an efficient strategy to reduce the influence of noise in the data.
- More sophisticated inner product (*e.g.* edge preserving adaptive smoothing) yield even better reconstructions.

Conclusions

- Selecting the inner product appropriately accelerates the convergence.
- More robust inversion path are obtain with preconditioning.

Perspectives

- Inner product preconditioning is an efficient strategy to reduce the influence of noise in the data.
- More sophisticated inner product (*e.g.* edge preserving adaptive smoothing) yield even better reconstructions.

Thank you for your attention