Optimal-transport objective functions in FWI

Yunan Yang[1], Björn Engquist[1], Junzhe Sun[2], Brittany Froese[3] and Lingyun Qiu[4]

May 2, 2017

[1]The University of Texas at Austin, USA

[2]Formerly UT-Austin, currently ExxonMobil Upstream Research Company

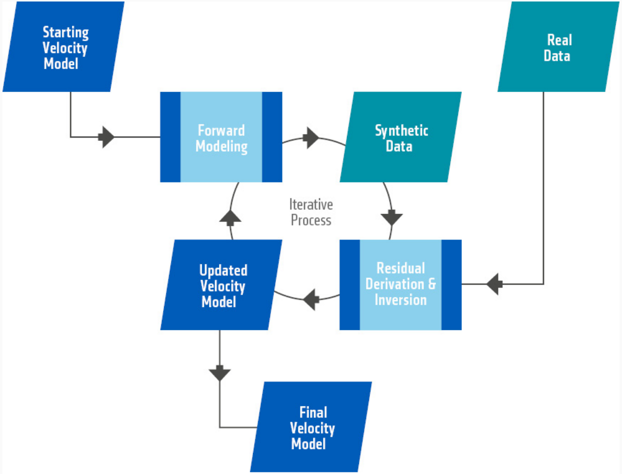
[3]New Jersey Institute of Technology, USA

[4]Petroleum Geo-Services, Inc

Table of contents

Background

Full Waveform Inversion (FWI): a PDE-constrained optimization



$$m^\star = \underset{m}{\operatorname{argmin}} \; \chi(m),$$

$\chi(m)$ is the objective function.

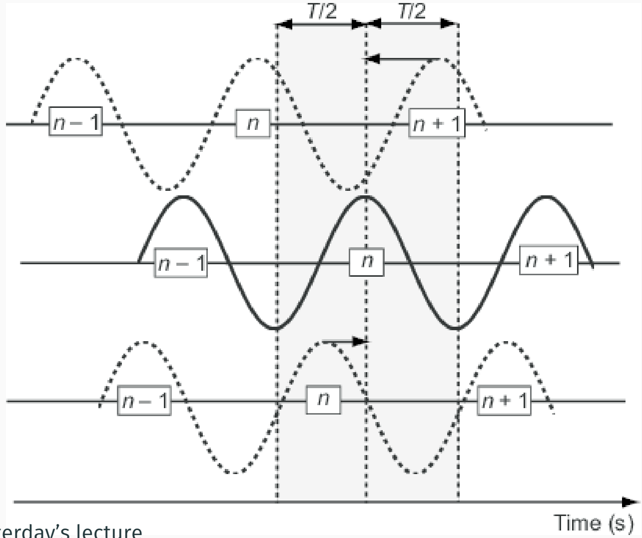
Example: L^2 norm as misfit function

We may define a least squares waveform misfit measure as:

$$\chi(m) = \frac{1}{2} \sum_r \int |s(x_r, t; m) - d(x_r, t)|^2 \, dt,$$

- observed data d,
- simulated data s,
- receiver x_r,
- the model parameter m.

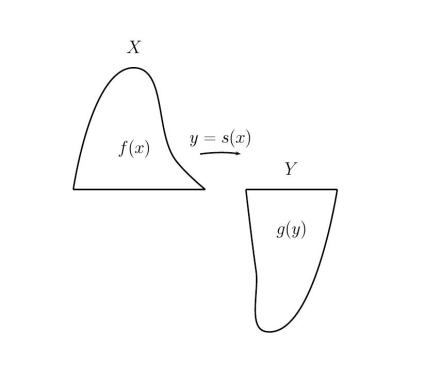
Headache of FWI: Cycle skipping



Virieux *el al*, 2009 and Yesterday's lecture

Optimal Transport

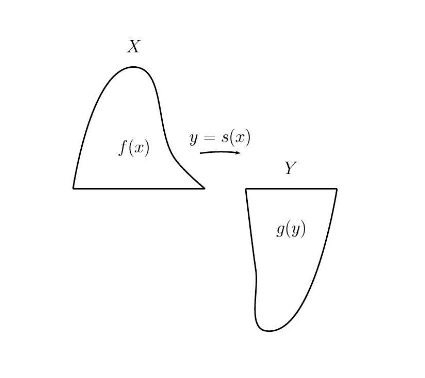
Optimal Transport



Brought up by Monge in 1781

- Monge (1781)
- Kantorovich (1975)
- Brenier, Caffarelli, Gangbo, McCann, Benamou, Otto, Villani, Figalli, etc. (1990s - present)

Optimal Transport

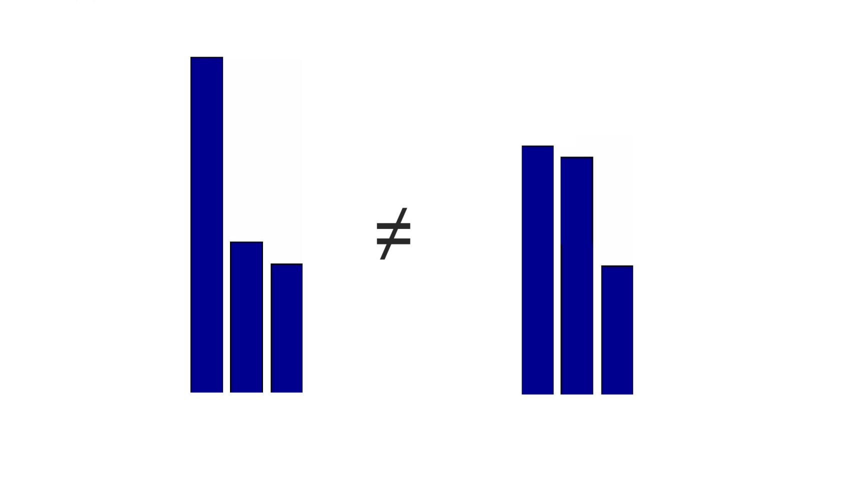


Brought up by Monge in 1781

- Monge (1781)
- Kantorovich (1975)
- Brenier, Caffarelli, Gangbo, McCann, Benamou, Otto, Villani, Figalli, etc. (1990s - present)

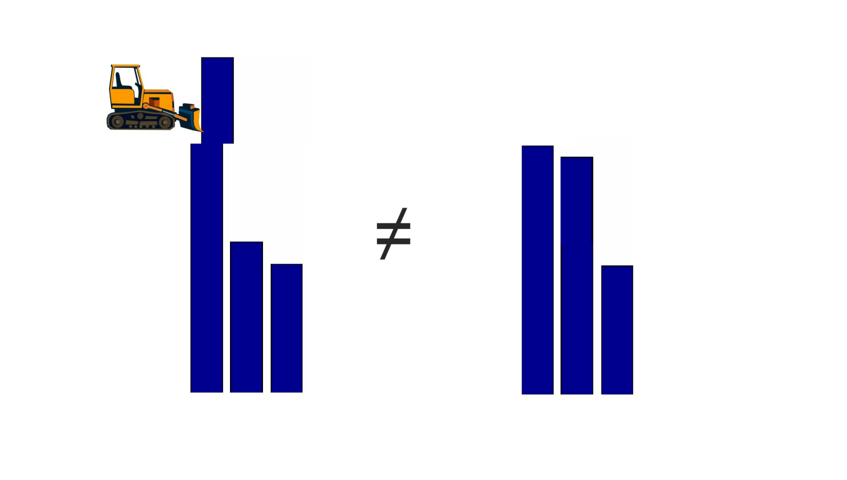
- Computer Science (EMD)
- Imaging processing and registration
- Machine learning (vs. KL-divergence)

Optimal transport



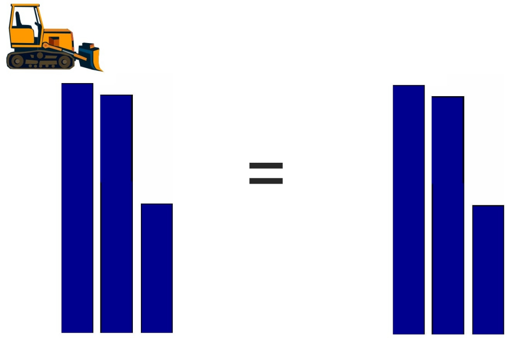
Synthetic data f (left) and observed data g (right)

Optimal transport



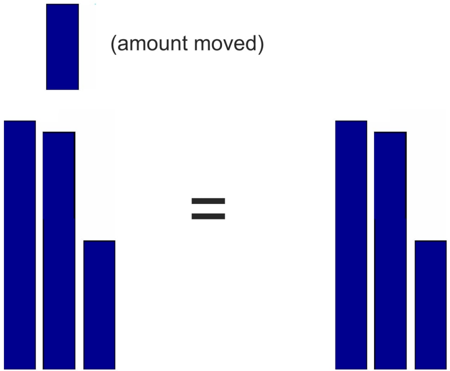
Synthetic data f (left) and observed data g (right)

Optimal transport



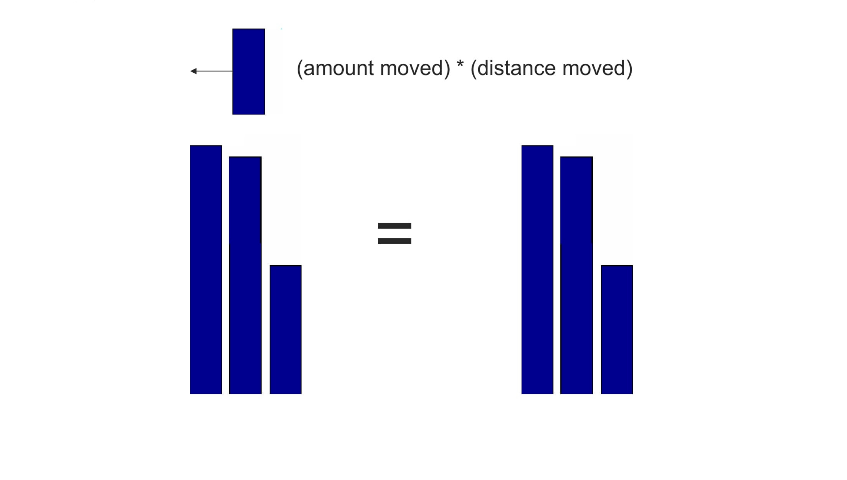
Synthetic data f (left) and observed data g (right)

Optimal transport



Synthetic data f (left) and observed data g (right)

Optimal transport

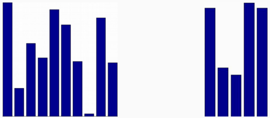


Synthetic data f (left) and observed data g (right)

Optimal transport: Wasserstein distance

Finally, for general functions f and g, the Wasserstein distance is

$$\min_{\text{All the map } T} \left(\sum_{\text{All movements of } T} \text{distance moved} \times \text{amount moved} \right)$$



Function f and g sharing the same mass by normalization

Different choice of distance in the application of FWI

- $|x-y|^2$: [Engquist and Froese, 2014, Engquist et al., 2016, Yang et al., 2016]
- $|x-y|$: [Métivier et al., 2016, Métivier et al., 2016]

Quadratic Wasserstein Distance (Earth Mover's Distance)

Definition of the Wasserstein distance

For $f : X \to \mathbb{R}^+$, $g : Y \to \mathbb{R}^+$, the distance can be formulated as

$$W_p(f,g) = \left(\inf_{T \in \mathcal{M}} \int |x - T(x)|^p f(x) dx \right)^{\frac{1}{p}} \tag{1}$$

$\mathcal{M}$ is the set of all maps that rearrange the distribution f into g.

Quadratic Wasserstein distance: $p = 2$

$$W_2^2(f,g) = \inf_{T \in \mathcal{M}} \int_X |x - T(x)|^2 f(x)\, dx \tag{2}$$

Misfit functions

Generalized least squares functional: from local to global

Oridinary least squares method

$$J_1(m) = \frac{1}{2}\sum_r \int |f(x_r, t; m) - g(x_r, t)|^2 \, dt, \tag{3}$$

Generalized least squares functional: from local to global

Oridinary least squares method

$$J_1(m) = \frac{1}{2} \sum_r \int |f(x_r, t; m) - g(x_r, t)|^2 \, dt, \tag{3}$$

The integral wavefields misfit functional [Huang et al., 2014]

$$J_2(m) = \frac{1}{2} \sum_r \int \left| \int_0^t f(x_r, \tau; m) d\tau - \int_0^t g(x_r, \tau) d\tau \right|^2 \, dt, \tag{4}$$

Generalized least squares functional: from local to global

Oridinary least squares method

$$J_1(m) = \frac{1}{2} \sum_r \int |f(x_r, t; m) - g(x_r, t)|^2 \, dt, \tag{3}$$

The integral wavefields misfit functional [Huang et al., 2014]

$$J_2(m) = \frac{1}{2} \sum_r \int \left| \int_0^t f(x_r, \tau; m) d\tau - \int_0^t g(x_r, \tau) d\tau \right|^2 dt, \tag{4}$$

Normalized Integration Method (NIM) [Liu et al., 2012]

$$J_3(m) = \frac{1}{2} \sum_r \int |W(f(x_r, t; m)) - W(g(x_r, t))|^2 \, dt, \quad W(u)(x_r, t) = \frac{\int_0^t P(u)(x_r, \tau) d\tau}{\int_0^T P(u)(x_r, \tau) d\tau}. \tag{5}$$

The operator P is included to make the data nonnegative: $|u|$, u^2 or $E(u)$.

Relations among misfit functions

L^2, Integral L^2 and NIM

L^2 norm

Compute the leasts-square difference

Integral wavefields misfit functional

1. Integrate data (the same as integrate source)
2. Compute the leasts-square difference

Normalized integration method (NIM)

1. Transform data to be positive
2. Integrate data
3. Compute the leasts-square difference

1D Optimal transport (trace by trace)

The explicit formulation for the 1D Wasserstein metric is:

$$W_2^2(f,g) = \int_0^1 |F^{-1}(x) - G^{-1}(x)|^2 dx. \tag{6}$$

where $F(t) = \int_{-\infty}^{t} \tilde{f}(\tau)d\tau$ and $G(t) = \int_{-\infty}^{t} \tilde{g}(\tau)d\tau$. $\tilde{f}$ and $\tilde{g}$ are normalized signals that have positivity and conservation of mass.

1D Optimal transport (trace by trace)

The explicit formulation for the 1D Wasserstein metric is:

$$W_2^2(f,g) = \int_0^1 |F^{-1}(x) - G^{-1}(x)|^2 dx. \tag{6}$$

where $F(t) = \int_{-\infty}^{t} \tilde{f}(\tau)d\tau$ and $G(t) = \int_{-\infty}^{t} \tilde{g}(\tau)d\tau$. $\tilde{f}$ and $\tilde{g}$ are normalized signals that have positivity and conservation of mass.

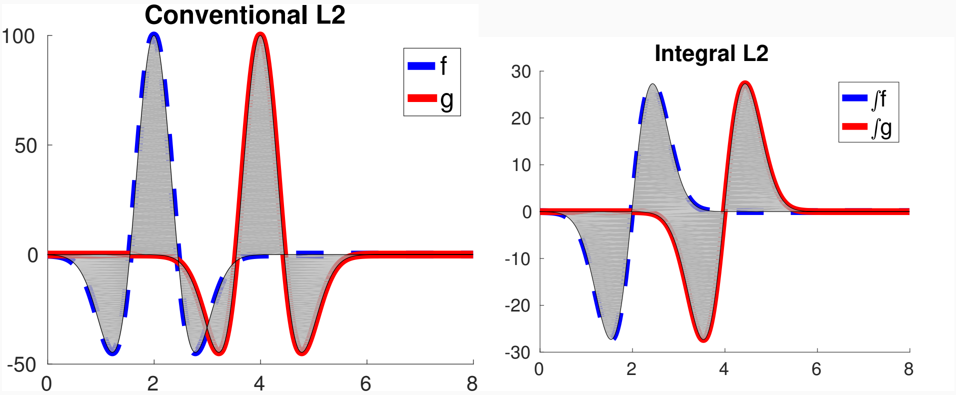
Normalized integration method

NIM has a similar objective function, which is also the norm of Sobolev space H^{-1} in functional analysis:

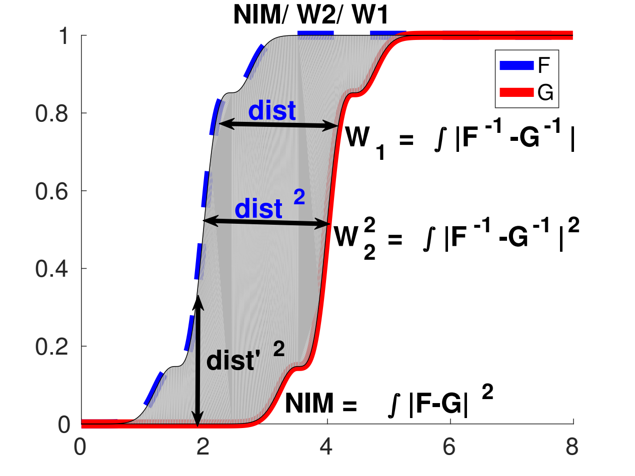
$$NIM(f,g) = \int_0^T |F(t) - G(t)|^2 dt. \tag{7}$$

Relations among misfit functions

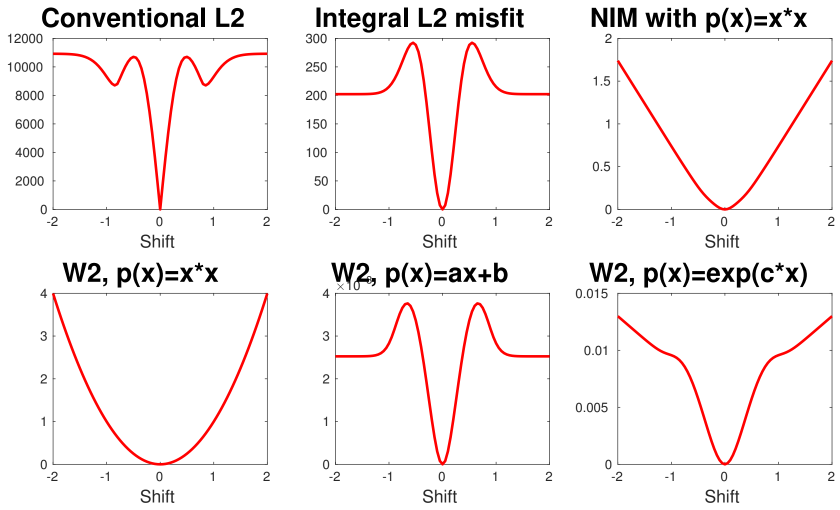
Signal g (red) is a shift of Ricker wavelet f (blue).



Both **positivity** and **integration** in time are important here:



Convexity of the misfit functions w.r.t. shift



The importance of normalization function

- $p(x) = |x|$ hurt the data regularity (smoothness)

The importance of normalization function

- $p(x) = |x|$ hurt the data regularity (smoothness)
- $p(x) = x^2$ gives perfect convexity, but there is an issue with the gradient at reflections.

The importance of normalization function

- $p(x) = |x|$ hurt the data regularity (smoothness)
- $p(x) = x^2$ gives perfect convexity, but there is an issue with the gradient at reflections.
  v changes $\Rightarrow f$ changes, and $\frac{\partial J(f^2, g^2)}{\partial v} \neq 0$, but $\frac{\partial J(f^2, g^2)}{\partial f} = 0$ whenever $f = 0$.

The importance of normalization function

- $p(x) = |x|$ hurt the data regularity (smoothness)
- $p(x) = x^2$ gives perfect convexity, but there is an issue with the gradient at reflections.
  v changes $\Rightarrow f$ changes, and $\frac{\partial J(f^2,g^2)}{\partial v} \neq 0$, but $\frac{\partial J(f^2,g^2)}{\partial f} = 0$ whenever $f = 0$.
  It is OK for transmission, source inversion, Camembert, etc.

The importance of normalization function

- $p(x) = |x|$ hurt the data regularity (smoothness)
- $p(x) = x^2$ gives perfect convexity, but there is an issue with the gradient at reflections.
  v changes $\Rightarrow f$ changes, and $\frac{\partial J(f^2,g^2)}{\partial v} \neq 0$, but $\frac{\partial J(f^2,g^2)}{\partial f} = 0$ whenever $f = 0$.
  It is OK for transmission, source inversion, Camembert, etc.
- $p(x) = ax + b$ no longer have the global convexity w.r.t. shifts.
  - averaging over several receivers may eliminate or reduce the local minima.
  - linear scaling does not distort the shared events in datasets if seismic data has mean zero property. Good for reflections.

The importance of normalization function

- $p(x) = |x|$ hurt the data regularity (smoothness)
- $p(x) = x^2$ gives perfect convexity, but there is an issue with the gradient at reflections.
  v changes $\Rightarrow f$ changes, and $\frac{\partial J(f^2, g^2)}{\partial v} \neq 0$, but $\frac{\partial J(f^2, g^2)}{\partial f} = 0$ whenever $f = 0$.
  It is OK for transmission, source inversion, Camembert, etc.
- $p(x) = ax + b$ no longer have the global convexity w.r.t. shifts.
  - averaging over several receivers may eliminate or reduce the local minima.
  - linear scaling does not distort the shared events in datasets if seismic data has mean zero property. Good for reflections.
- $p(x) = exp(cx)$ is very close to linear transformation when c is small from Taylor expansion, but it can be global convex if c is chosen carefully.

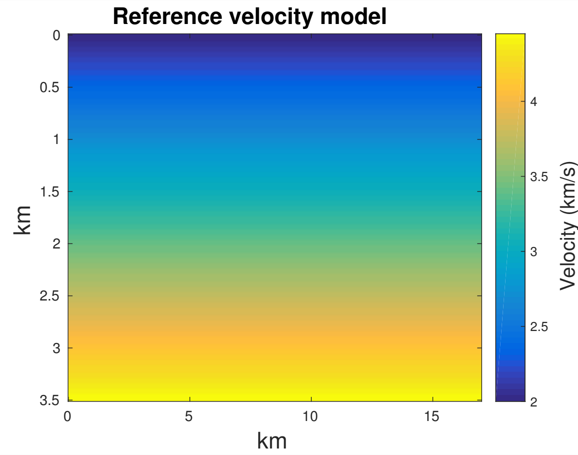
Convexity: data domain to model domain ($p(x) = ax + b$)

Here is a test for the convexity in model domain [Métivier et al., 2016].
The p-wave velocity model is assumed to vary linearly in depth such that

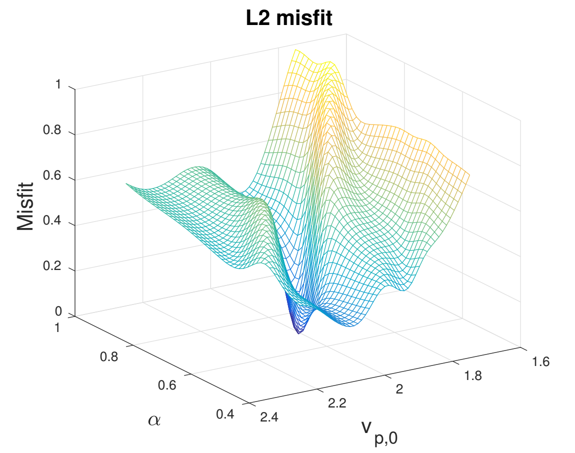
$$v_p(x,z) = v_{p,0} + \alpha z$$

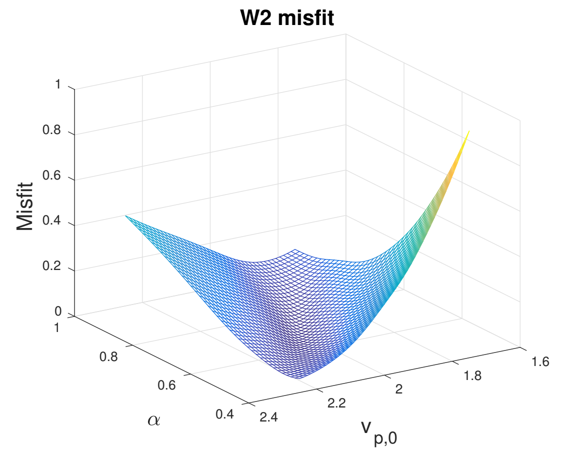
The reference is chosen so that $v_{p,0} = 2$ km/s and $\alpha = 0.7\ s^{-1}$. The L^2 and W_2 misfit functions are then evaluated on a grid of 41×45 points such that

$$v_{p,0} \in [1.75, 2.25], \quad \alpha \in [0.4, 1]$$



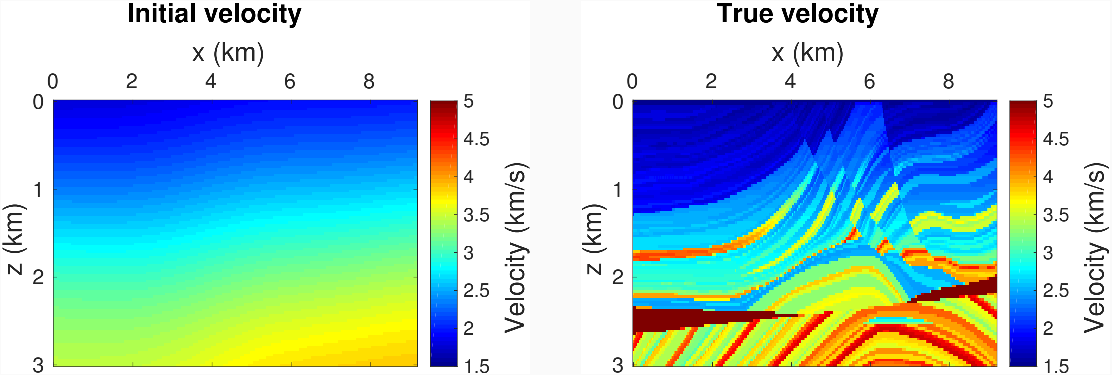
Convexity: data domain to model domain ($p(x) = ax + b$)



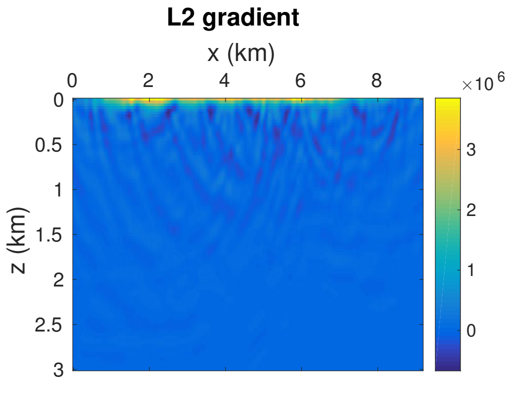


Numerical Results

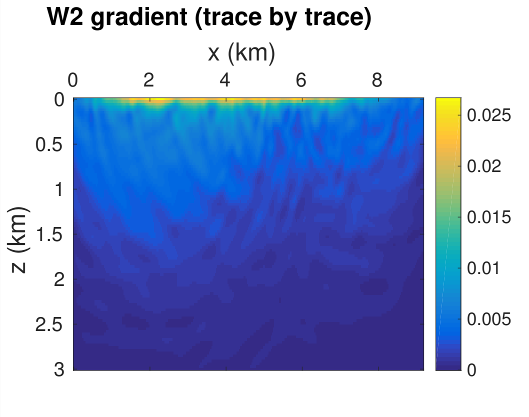
1. Marmousi model: trace-by-trace W_2 comparison



1. Marmousi model: gradient of the first iteration

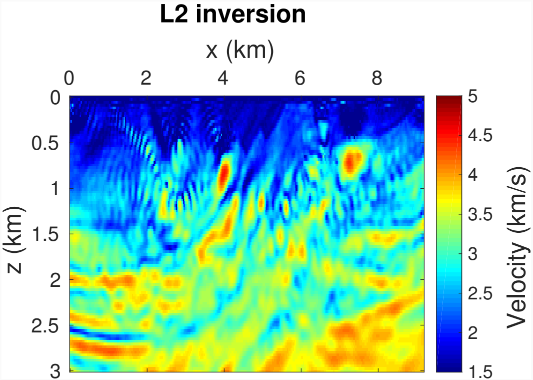


5Hz Ricker without 0-2 Hz

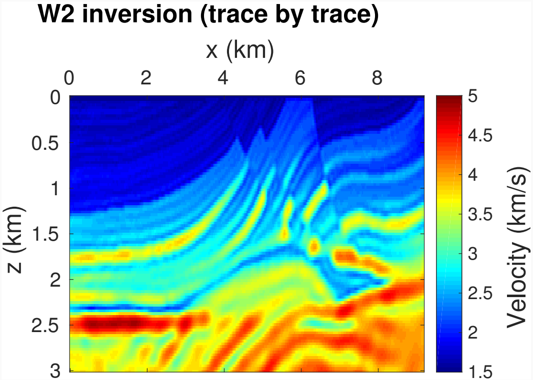


11 sources on top

1. Marmousi model: trace-by-trace W_2 comparison

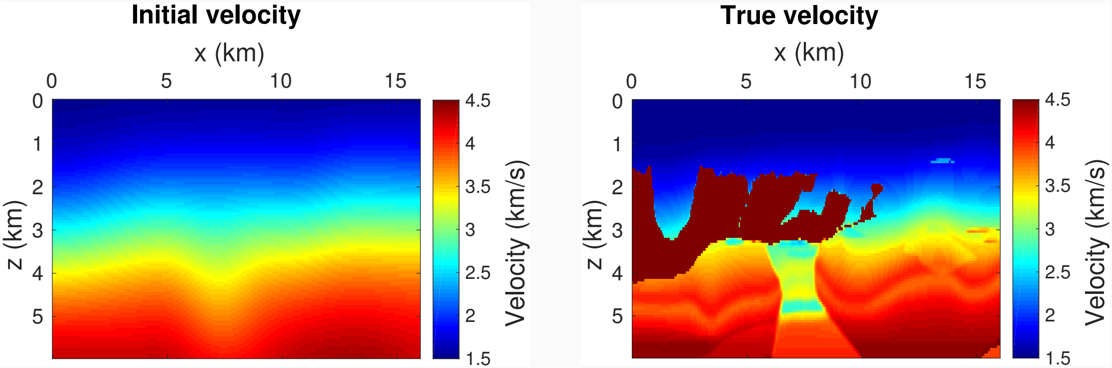


After 300 l-BFGS iterations

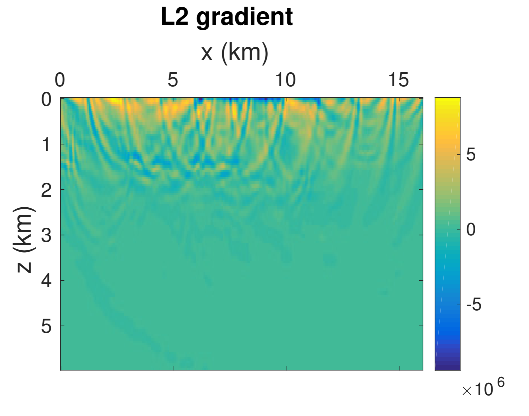


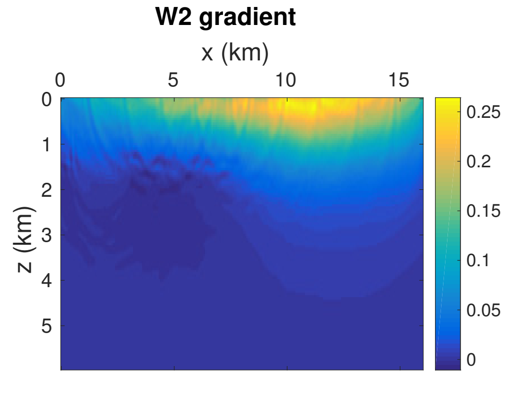
Computing time are the same.

2. Modified BP model: trace-by-trace W_2 comparison



2. Modified BP model: gradient of the first iteration

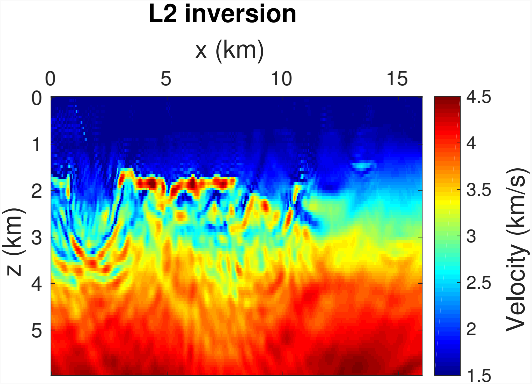




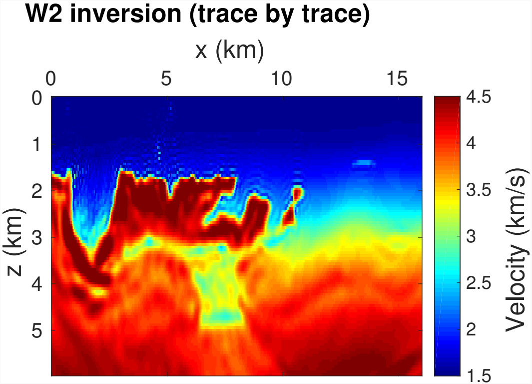
5Hz Ricker keeping 3-9 Hz

11 sources on top

2. Modified BP model: trace-by-trace W_2 comparison



After 300 l-BFGS iterations



Computing time are the same.

Conclusion

Conclusion

Integration in time

- Reduces high frequencies
- Help remove the noise by averaging
- May not avoid local minima

Positivity

- "Breaks" the oscillatory periodicity which further reduces the risk of cycle skipping.

Summary

The analysis brings additional insights into the importance of seismic data preconditioning

Remarks

Challenges

- Constraints of optimal transport are not natural in seismology:

$$\int_X f(x)dx = \int_Y g(y)dy, \quad f, g \geq 0, \quad \text{convex domain}$$

Remarks

Challenges

- Constraints of optimal transport are not natural in seismology:

$$\int_X f(x)dx = \int_Y g(y)dy, \quad f, g \geq 0, \quad \text{convex domain}$$

Need data normalization ($|f|$, f^2, $a(f+b)$, ...)

Remarks

Challenges

- Constraints of optimal transport are not natural in seismology:

$$\int_X f(x)dx = \int_Y g(y)dy, \quad f, g \geq 0, \quad \text{convex domain}$$

  Need data normalization ($|f|$, f^2, $a(f+b)$, ...)
- For W_2 in 2D, a faster Monge-Ampère solver is required (Regularity of g, error, cost, etc.)

Remarks

Challenges

- Constraints of optimal transport are not natural in seismology:

$$\int_X f(x)dx = \int_Y g(y)dy, \quad f, g \geq 0, \quad \text{convex domain}$$

  Need data normalization ($|f|$, f^2, $a(f+b)$, ...)
- For W_2 in 2D, a faster Monge-Ampère solver is required (Regularity of g, error, cost, etc.)
- Trace-by-trace comparison (W_2 in 1D) is successful [Yang et al., 2016]

References

- Engquist, B., and B. D. Froese, 2014, Application of the Wasserstein metric to seismic signals: Communications in Mathematical Sciences, **12**.
- Engquist, B., B. D. Froese, and Y. Yang, 2016, Optimal transport for seismic full waveform inversion: Communications in Mathematical Sciences, **14**
- Yang, Y., B. Engquist, J. Sun, and B. D. Froese, 2016, Application of optimal transport and the quadratic wasserstein metric to full-waveform inversion: arXiv preprint arXiv:1612.05075.
- Yang, Y., and Engquist, B., 2017, Analysis of optimal transport and related misfit functions in FWI: submitted.