Mitigating cycle skipping issue in FWI: an optimal transport investigation

By L. Métivier$^{1,2}$, R. Brossier$^2$, Q. Mérigot$^3$, E. Oudet$^1$ and J. Virieux$^2$

SEISCOPE II [http://seiscope2.osug.fr](http://seiscope2.osug.fr)
Agenda

• Seismic imaging
• Full waveform inversion (FWI)
• Optimal transport (OT)
• Examples
• Conclusion
Seismic imaging
Seismic imaging of the Earth

What part of the data are we really using?
SEISMIC IMAGING: BIG DATA CHALLENGE

Drastic data increase ..

Drastic crunching capacities ...

Extract the most from seismograms (full database) with the best physics (full modeling) using the most efficient approach
Answer: FWI ...

“FWI”-related publications from SEG, from http://library.seg.org (Wellington, 2016)
Full waveform inversion
What is Full Waveform Inversion

Cost function

\[ C(m) = \| \Delta d \| \]
Fundamentals of FWI (b)
Fundamentals of FWI (c)
FWI – Classical workflow

Initial velocity model

\[ m_0 \]

Forward modeling

Wave equation

\[ \text{ie. finite differences} \]

Synthetic data

\[ d_{cal}(m) \]

Observed data

\[ d_{obs} \]

Calculate misfit

\[ C(m) = \frac{1}{2} (d_{obs}(x, t) - d_{cal}(x, t, m))^2 \]

Minimize L2 norm of difference

Updated velocity model

\[ m_{i+1} = m_i + \Delta m \]

Iterative approach

Calculate model update \( \Delta m \)

\[ C'(m) = \gamma \text{ and } C''(m) = \mathcal{H} \text{ (or approx.)} \]

Newton Equation

\[ \mathcal{H} \Delta m = -\gamma \]

HPC issues: computing \( \gamma \) and (sometimes) \( \mathcal{H} \nu \)

where \( \nu \) is a vector of model space

Minimal tasks

a) Wave modeling

b) Gradient \( \gamma \)

c) Hessian \( \mathcal{H} \)?
A successful story of FWI on real datasets by many different groups (in the last 5-10 years)
The Valhall reservoir: LoFS

3D sampling: 49,954 shots, 2300 hydrophones
Frequency range 3.5 - 10 Hz
Target area 145 km2; depth 4.5 km

Prieux et al. (2013a,b) & Gholami et al. (2013) & Operto et al. (2013, 2015)

2D sampling line with 4C components

(Operto et al., 2015)
The Valhall reservoir: LoFS

Similar to results from Sirgue et al. (2010) ... but anisotropy included in the forward modeling.
Scrapes left by drifting icebergs on the paleo seafloor ($z = 500m$)
FWI: a data-driven inversion

\[ C(m) = \frac{1}{2} \Delta d^\dagger \Delta d \] to be minimum

Data hierarchy strategy making FWI work

- Low frequencies to high frequencies
- Short offsets to long offsets

FWI: a purely data-driven technique

Data information enough for stable imaging
Linear Algebra

\[ C(m) = \frac{1}{2} (d_{obs}(x_{rec}, t) - d_{cal}(x_{rec}, t, m))^2 \] to be minimized

At the new model \( m_j + \Delta m_j \),

\[ \frac{\partial C}{\partial m}(m_j + \Delta m_j) = 0 \]

\[ \frac{\partial^2 C}{\partial m^2}(m_j)\Delta m_j + \frac{\partial C}{\partial m}(m_j) = 0 \]

\[ \frac{\partial^2 C}{\partial m^2}(m_j)\Delta m_j = -\frac{\partial C}{\partial m}(m_j) \]

The full Newton approach

\[ \mathcal{H} \Delta m_j = -\hat{\mathcal{g}} \]

10^5 to 10^7 unknowns in 2D; 10^6 to 10^{10} unknowns in 3D

Optimization under constraints

\[
\min_m \frac{1}{2} \left( d_{obs}(x_{rec}, t) - d_{cal}(x_{rec}, t, m) \right)^2
\]

s.t.

\[
\frac{1}{m(x)^2} \partial_{tt} p(x, t) - \Delta p(x, t) = f_s(x_s, t) \quad \text{with} \quad (x, t) \in \Omega \times [0, T]
\]

with \(d_{cal}(x_{rec}, t, m) = R(p(x, t))\).

We define the extractor operator \(R\) at receivers location such that

\[
R : p(x, t) \rightarrow [p(x_1, t), \ldots, p(x_r, t)]
\]
Augmented Lagrangian approach

We define the lagrangian by

\[ \mathcal{L}(m, p, a) = \frac{1}{2} \langle d_{obs} - R[m] | d_{obs} - R[m] \rangle + \left\langle \frac{1}{m^2} \partial_{tt} p - \Delta p - f_s | a \right\rangle \]

The field \( a \) is the adjoint state field verifying the adjoint equation

\[ \frac{1}{m(x)^2} \partial_{tt} a - \Delta a = \frac{\partial C(m)}{\partial p} = R^T (d_{obs} - R[m]) \]

The misfit function minimum is the saddle point of the Lagrangian
Adjoint state method: efficient method

\[ \frac{\partial L}{\partial a} = 0 \iff \frac{1}{V_p^2(x)} \partial_{tt} p(x, t) - \Delta p(x, t) = f_s(x, t) \quad \text{state equation} \]

\[ \frac{\partial L}{\partial p} = 0 \iff \frac{1}{V_p^2(x)} \partial_{tt} a(x, t) - \Delta a(x, t) = R^T \frac{\partial C}{\partial p} \quad \text{adjoint equation} \]

\[ \frac{\partial L}{\partial V_p} = 0 \iff \frac{\partial C}{\partial V_p} = \left\langle \frac{2}{V_p^3} \partial_{tt} p(x, t) \middle| a(x, t) \right\rangle \quad \text{gradient equation} \]

- For computing the gradient, we have two PDEs to solve: it is used for updating the model.

- The definition of distance impacts only terms in red: any other distance definition should modify these terms.

- The Hessian product could be approximated through various strategies (quasi-Newton, truncated Newton, full Newton ...)

6/9/2016
FWI – Classical workflow

Initial velocity model

$m_0$

Forward modeling

Wave equation

ie. finite differences

Synthetic data

$\mathbf{d}_{cal}(m)$

Observed data

$\mathbf{d}_{obs}$

Calculate misfit

$$ \mathcal{C}(m) = \frac{1}{2} (\mathbf{d}_{obs}(x,t) - \mathbf{d}_{cal}(x,t,m))^2 $$

Minimize L2 norm of difference

Iterative approach

$m = m_i + \Delta m$

Updated velocity model

Calculate model update

$\mathcal{C}'(m) = \gamma$ and $\mathcal{C}''(m) = \mathcal{H}$ (or approx.)

Newton Equation

$$ \mathcal{H} \Delta m = - \gamma $$

HPC issues: computing $\gamma$ and (sometimes) $\mathcal{H} \nu$

where $\nu$ is a vector of model space

Minimal tasks

a) Wave modeling

b) Gradient $\gamma$

c) Hessian $\mathcal{H}$
No prior scale separation

Single-scattering formulation

Pixel structure (# from a blocky structure)

The model wavenumber spectrum is described through this pixel strategy

\[ k = 2\pi f q = \frac{4\pi f}{c} \cos \left( \frac{\theta}{2} \right) n \]

\[ k = \frac{4\pi}{\lambda} \cos \left( \frac{\theta}{2} \right) n \]

The frequency is f and the angle \( \theta \) is the aperture or illumination angle: they are the controlling parameters of the model velocity spectrum.
FWI of body waves

Considering diving waves and reflected waves information has led to the success of the FWI.

no seismic phase identification and, therefore no scale separation

Incident wave modeling
FWI of body waves

Gradient field connected to model update

\[ \gamma = f(\Delta d) \]
Residuals are sent back and not data (#LSM)

Direct and reflected waves are interpreted the same way through first-order scattering (no scale separation)
Stacking over sources and receivers

Gradient shape

Two zones:

A – both weak and strong interactions between waves and matter

B – only strong interaction between waves and matter
Good news: gradient is often enough

BP 2004 model

Metivier et al. (2013)
Initial model design

We need an initial model not too far away for the final model to succeed (secondary minima issue)

“How far could we start?” is the big question

- Seismic data is an oscillatory signal
- Observed phases in the data are not all predicted at each iteration (non conservation)
Cycle skipping issue

Initial model design

Prediction of a phase of an arrival with a modeling error less than half the period

\[
\frac{\Delta T}{T_L} < \frac{1}{2N_\lambda}
\]

\(N_\lambda\): number of propagated wavelengths
\(T_L\): travel time

50 \(\lambda\): 1% of relative time errors -> at long offsets, cycle-skipping difficulties increase

Data-oriented mitigation of this non-linearity
Low frequencies could be the easy answer as reducing \(N_\lambda\) ... Hierarchical data feeding (shorter apertures and slowly increasing recording times) ...
Simple example on Ricker wavelet

Computing the L2 misfit between seismic traces with respect to time-shifts yields a multi-modal misfit function ...

Point-to-point comparison is not optimal
Other ways: cross-correlation, convolution, wiener filter ...
How to compare two datasets?

What we compare in these gathers?

Definition of the distance?

Observed  Calculated
Distance through optimal transport

Measuring the distance using an optimal transport distance, a proposition from Engquist and Froese (2014)

For 2 Ricker signals $r_1(t)$ and $r_2(t)$, the optimal transport distance is

$$\min_{M} \int_{t \in [0,T]} c(t, M(t)) r_1(t) dt$$

where

- $M : t \mapsto M(t) = t'$ is a transport map from $r_1(t)$ to $r_2(t)$
- For any subset $A \subset [0, T]$, we have $\int_{t \in A} r_2(t) dt = \int_{M(t) \in A} r_1(t) dt$
- Function $c(t, t')$ measures the « price » for moving the signal (« mass ») from $t$ to $t'$: $c(t, t') = |t - t'|$
Computing the optimal transport misfit between seismic traces with respect to time-shifts yields a convex misfit function (Engquist and Froese, 2014).
Two main difficulties ... and a third one

1) No « mass » preservation (some phases are not predicted...)

2) Oscillatory signals

3) Large scale problem, especially in 3D: the total data cube is 5D
Summary

- Subsurface reconstruction requires a correct interpretation of time shifts between observed and predicted phases.
- L2-norm is unable to do so: non-convex misfit function.
- OT-norm does provide a convex misfit function.
- Standard OT assumes positive signal and « mass » conservation.
- Efficient OT algorithms are required.
Optimal transport: from Monge problem to Kantorovitch-Rubinstein distance
Proposition

Change the distance in the FWI problem by the Kantorovitch-Rubinstein distance, leading to

$$\min_{m} \| d_{cal}[m] - d_{obs} \|_{KR}$$

where this distance is defined by

$$\max_{\varphi \in BL_1} \int_0^T \int_{\partial \Omega} \varphi(x, t)(d_{cal}[m](x, t) - d_{obs}(x, t)) dx dt$$

We consider bounded 1-Lipschitz functions $\varphi$
What we need to do …

\[
\frac{\partial \mathcal{L}}{\partial a} = 0 \iff \frac{1}{V_p^2(x)} \partial_{tt} p(x, t) - \Delta p(x, t) = f_s(x, t) \quad \text{state equation}
\]

\[
\frac{\partial \mathcal{L}}{\partial p} = 0 \iff \frac{1}{V_p^2(x)} \partial_{tt} a(x, t) - \Delta a(x, t) = R^T \frac{\partial C_{KR}}{\partial p} \quad \text{adjoint equation}
\]

\[
\frac{\partial \mathcal{L}}{\partial V_p} = 0 \iff \frac{\partial c}{\partial V_p} = \left( \frac{2}{V_p^3} \partial_{tt} p(x, t) \right|_{a(x, t)}
\]

\[
\text{Computing the new adjoint, thanks to the adjoint source term } \quad \frac{\partial C_{KR}}{\partial p} = \overline{\phi} = \arg \max_{\phi \in BLip_{1,c}} \int_0^T \int_{\partial \Omega_r} \phi \delta d \ dx dt
\]

\[
\text{... and the gradient}
\]

\[
\text{Computing the new misfit function}
\]

Intrusion is minimal into FWI structure
An new look at an old problem

Monge problem (théorie des déblais et des remblais, 1781)

Move a pile of sand (rubble) to an excavation (fill) through a transport $T$ while minimizing the workers energy defined by

$$C(T) = \int_{X} |T(x) - x| dx$$

The transport of positive weights should preserve the total mass.

Very non-linear problem: generalization by Leonid Kantorovitch (‘42, ‘48) with mass splitting from the original point to final destinations.
Another wording of the same pb

Monge problem (théorie des déblais et des remblais, 1781)

Moving a quantity from position $x$ to position $y$ is defined by a function $c(x, y)$ such that

$$
C(T) = \int_{\Omega} c(x, T(x)) \, d\mu(x) \quad \text{with} \quad T\#\mu = \nu
$$

is minimum. The quantity $\mu$ is an initial measure of the sand while the quantity $\nu$ is the final one.
Kantorovitch relaxation of the Monge problem

Let $X$ and $Y$ denote two separable metric spaces with $\mathcal{P}(X)$ and $\mathcal{P}(Y)$ the two spaces of probability distributions on $X$ and $Y$. Consider two measures $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ and a cost function $c(x, y)$ defined by

$$
\begin{cases}
  c: (x, y) \rightarrow c(x, y) \\
  X \times Y \rightarrow \mathbb{R}_+
\end{cases}
$$

Kantorovitch relaxation is

$$
\inf_{\gamma \in \Pi(\mu, \nu)} \left\{ \int_{X \times Y} c(x, y) d\gamma(x, y) \right\}
$$

with transport plans $\Pi(\mu, \nu)$ defined by

$$
\Pi(\mu, \nu) = \{ \gamma \in \mathcal{P}(X \times Y), (\pi_X)_\#\gamma = \mu, (\pi_Y)_\#\gamma = \nu \}
$$

noting $\pi_X$ and $\pi_Y$, projectors on $X$ and $Y$ respectively.
By introducing the indicator function of $\Pi(\mu, \nu)$, we can formulate the dual expression

$$\inf_{\gamma \in \Pi(\mu, \nu)} \left\{ \int_{X \times Y} c \, d\gamma \right\} = \max_{\varphi, \psi} \left\{ \int_X \varphi \, d\mu + \int_Y \psi \, d\nu \right\}$$

$$\forall (x, y) \in X \times Y, \varphi(x) + \psi(y) \leq c(x, y)$$

with $\varphi \in C_b(X), \psi \in C_b(Y)$

Functions $\varphi$ and $\psi$ realizing the maximization are Kantorovitch potentials.
If $Y = X$ and the cost function $c(x, y)$ is a distance, the dual formulation can be simplified into

$$\max_{\varphi \in \text{Lip}_{1,c}} \left\{ \int_X \varphi d(\mu - \nu) \right\}$$

(KDF)

where

$$\text{Lip}_{1,c} = \{ \varphi : x \in X \to \mathbb{R}, \forall (x, x') \in X \times X, |\varphi(x) - \varphi(x')| \leq c(x, x') \}$$

(Santambrogio, 2015, 3.1.1)
« Mass » conservation needed!

Consider the case where \( \mu \) and \( \nu \) have different « mass »

\[
\int_X d(\mu - \nu) \neq 0
\]

Consider the constant function \( \varphi_\alpha(x) = \alpha \) for \( \alpha \in \mathbb{R}^+ \). We have \( \varphi_\alpha \in Lip_{1,c} \) and

\[
\int_X \varphi_\alpha d(\mu - \nu) = \alpha \int_X d(\mu - \nu) \neq 0
\]

The problem (KDF) has no solution as, for any \( \varphi \in Lip_{1,c} \), one can find \( \varphi_\alpha \in Lip_{1,c} \), such that, for \( \alpha \) sufficiently large,

\[
\int_X \varphi_\alpha d(\mu - \nu) > \int_X \varphi d(\mu - \nu)
\]
Generalization of Kantorovitch dual formulation

A natural generalization of the KDF consists in complementing the 1-Lipschitz constraint with a bound constraint, giving

$$\max_{\phi \in BLip_{1,c}} \left\{ \int_X \phi d(\mu - \nu) \right\}$$  \hspace{1cm} (KDF)

where $\phi \in Lip_{1,c}$ and $\|\phi\|_\infty \leq \lambda$

In the particular case where $c(x, x') = \|x - x'\|_1$, the definition corresponds to the Kantorovitch-Rubinstein norm (Bogachev, 2007; Lellmann et al., 2014)
Finding $\varphi$ Or Computing KR distance efficiently!

$\varphi$ is bounded and 1-Lipschitz

$\varphi_{\text{max}}$

$\varphi_{\text{min}}$

$\text{Max/Min}(\Delta \varphi_{ij})$

$\text{Max/Min}(\Delta \varphi)$

three states

$\text{Max/Min}(\Delta \varphi_{ij})$

1-Lipschitz

Bookshelf ordering ...

- moving the $n$ books one-book width: $n$ small displacements?

- moving the left-most book to the right: 1 large displacement?
Discrete notations

For a source (CSG), we consider a 3D Cartesian mesh \((x_i, y_j, t_k) \in \mathbb{R}^3\) such that
\[
\begin{align*}
  x_i &= a_x + (i - 1)h_x, \\
  y_j &= a_y + (j - 1)h_y, \\
  t_k &= a_t + (k - 1)h_t
\end{align*}
\]

We use the standard discretization notations such that
\[
\forall \varphi, \varphi(x_i, y_j, t_k) = \varphi_{i,j,k}
\]

We introduce the set of indexes on the mesh
\[
\mathcal{A} = \{(i, j, k) \in N^3, 1 \leq i \leq N_x, 1 \leq j \leq N_y, 1 \leq k \leq N_t\}
\]
\[
\text{card} \mathcal{A} = N = N_x \times N_y \times N_t
\]

We also introduce the set of indexes \(\mathcal{A}_x, \mathcal{A}_y, \mathcal{A}_z\) such that
\[
\mathcal{A}_x = \{(i, j, k) \in N^3, 1 \leq i \leq N_x - 1, 1 \leq j \leq N_y, 1 \leq k \leq N_t\},
\]
\[
\mathcal{A}_y = \{(i, j, k) \in N^3, 1 \leq i \leq N_x, 1 \leq j \leq N_y - 1, 1 \leq k \leq N_t\},
\]
\[
\mathcal{A}_t = \{(i, j, k) \in N^3, 1 \leq i \leq N_x, 1 \leq j \leq N_y, 1 \leq k \leq N_t - 1\}
\]
Discrete KDF

Kantorovich-Rubinstein distance for one shot

$$\max_{\varphi_{ijk}} \sum_{i,j,k} \varphi_{ijk} (\mu_{ijk} - \nu_{ijk}),$$

$$\forall (i, j, k), (l, m, n) \in \mathcal{A}^2, |\varphi_{ijk} - \varphi_{lmn}| < |x_i - x_l| + |y_j - y_m| + |t_k - t_n|$$

$$\forall (i, j, k) \in \mathcal{A}$$

$$|\varphi_{ijk}| \leq \lambda$$

$\mathcal{O}(N^2)$ global linear constraints:

too high complexity for efficient numerical algorithms
Equivalent discrete local KDF

Equivalent (to be shown in next 2 slides) distance for one shot

\[
\max_{\varphi_{ijk}} \sum_{i,j,k} \varphi_{ijk} (\mu_{ijk} - \nu_{ijk}),
\]

\[
\begin{align*}
\forall (i, j, k) \in \mathcal{A}_x & \quad |\varphi_{i+1jk} - \varphi_{ijk}| < |x_{i+1} - x_i| \\
\forall (i, j, k) \in \mathcal{A}_y & \quad |\varphi_{ij+1k} - \varphi_{ijk}| < |y_{j+1} - y_j| \\
\forall (i, j, k) \in \mathcal{A}_t & \quad |\varphi_{ijk+1} - \varphi_{ijk}| < |t_{k+1} - t_k| \\
\forall (i, j, k) \in \mathcal{A} & \quad |\varphi_{ijk}| < \lambda
\end{align*}
\]

\(\mathcal{O}(N)\) local linear constraints:

far better complexity for efficient numerical algorithms
Global inequalities imply local inequalities

✓ Consider a pair of points on the mesh denoted by \((u, v)\), such that

\[
\begin{align*}
    u &= (x_i, y_j, t_k), \\
    v &= (x_l, y_m, t_n)
\end{align*}
\]

✓ For the \(L_1\) norm, a sequence of points \(w_q = (x_{i_q}, y_{j_q}, t_{k_q})\), \(q = 1, \ldots, M\) can be selected to form a path from \(u\) to \(v\) such that \(w_1 = u\) and \(w_M = v\), and \(w_q\) are all adjacent on the grid, with monotonically varying coordinates

\[
\|v - u\|_1 = \sum_{q=1}^{M} \|w_{q+1} - w_q\|_1
\]
Local inequalities imply global inequalities

✓ Consider a function \( \varphi \) satisfying

\[
\begin{align*}
\forall (i, j, k) \in \mathcal{A}_x & : |\varphi_{i+1jk} - \varphi_{ijk}| < |x_{i+1} - x_i| \\
\forall (i, j, k) \in \mathcal{A}_y & : |\varphi_{ij+1k} - \varphi_{ijk}| < |y_{j+1} - y_j| \\
\forall (i, j, k) \in \mathcal{A}_t & : |\varphi_{ijk+1} - \varphi_{ijk}| < |t_{k+1} - t_k|
\end{align*}
\]

✓ The triangle inequality yields

\[
\|\varphi(v) - \varphi(u)\|_1 \leq \sum_{q=1}^{M} \|\varphi(w_{q+1}) - \varphi(w_q)\|_1
\]

✓ As points \( w_q \) are adjacent, the local inequalities satisfied by \( \varphi \) yield

\[
\sum_{q=1}^{M} \|\varphi(w_{q+1}) - \varphi(w_q)\|_1 \leq \sum_{q=1}^{M} \|w_{q+1} - w_q\|_1
\]

✓ Putting these two inequalities yields

\[
\|\varphi(v) - \varphi(u)\|_1 \leq \sum_{q=1}^{M} \|\varphi(w_{q+1}) - \varphi(w_q)\|_1 = \|v - u\|_1
\]
Local KDF as a convex problem

Problem (LKDF) is reformulated as a non-smooth convex problem

\[
\max_\varphi f_1(\varphi) + f_2(\varphi),
\]

where \( f_1(\varphi) = \sum_{(i,j,k) \in \mathcal{A}} \varphi_{ijk}(\mu_{ijk} - \nu_{ijk}), \quad f_2(\varphi) = i_K(A\varphi) \)

✓ \( K \) the unit hypercube \( \{x \in \mathbb{R}^p, |x_i| \leq 1, i = 1, \ldots, p\} \)
✓ \( i_K \) the indicator function of \( K \) \( i_K(x) = \begin{cases} 0 & \text{if } x \in K \\ +\infty & \text{if } x \notin K \end{cases} \)
✓ \( A \in \mathcal{M}_{p \times N}(\mathbb{R}) \), rectangular matrix with \( P \) rows and \( N \) columns

\[
A^T = \begin{bmatrix} D_x & D_y & D_t & \frac{1}{\lambda} I_N \end{bmatrix}^T
\]

where \( I_N \) is the real identity matrix of size \( N \) and \( D_x \), \( D_y \), \( D_t \) are the forward FD operators: for example

\[
(D_x \varphi)_{ijk} = \frac{\varphi_{i+1,j,k} - \varphi_{ijk}}{h_x}
\]
SDMM algorithm (Combettes & Pesquet, 2011)

Solving this **non-smooth convex problem** with SDMM algorithm

\( y_1^0 = 0, y_2^0 = 0, z_1^0 = 0, z_2^0 = 0 \)

For \( n=0,1, \ldots \),

\( \varphi^n = (I_N + A^T A)^{-1}[(y_1^n - z_1^n) + A^T (y_2^n - z_2^n)] \)

\( y_1^{n+1} = \text{prox}_{f_1}(\varphi^n + z_1^n) \)

\( z_1^{n+1} = z_1^n + \varphi^n - y_1^{n+1} \)

\( y_2^{n+1} = \text{prox}_{i_K}(A\varphi^n + z_2^n) \)

\( z_2^{n+1} = z_2^n + A\varphi^n - y_2^{n+1} \)

End for

Closed-form expressions for proximal operators: \( \mathcal{O}(N) \) complexity

- \( \text{prox}_{f_1}(\varphi) = \varphi - \mu + \nu \)

- \( \text{prox}_{i_K}(x_i) = \begin{cases} x_i & \text{if } -1 \leq x_i \leq 1 \\ 1 & \text{if } x_i > 1 \\ -1 & \text{if } x_i < -1 \end{cases} \)
Inversion of the matrix $I_N + A^T A$

The discrete matrix $A^T A \in \mathcal{M}_{N \times N}(\mathbb{R})$ is defined such that

$$A^T A = \Delta + I$$

where the matrix $\Delta$ correspond to the second-order FD discretization of the Laplacian operator with Neumann homogeneous boundary conditions (tedious demonstration through recurrence).

(Métivier et al, 2016b).

The SDMM algorithm requires the solution at each iteration of a Poisson problem discretized with second-order FD schemes.

FFT solver (Swarztrauber, 1974): complexity $O(N \log N)$

Multigrid algorithms (Brandt, 1977; Adams, 1989): complexity $O(N)$
Summary

• The Kantorovich-Rubinstein norm can be seen as a generalization of the dual formulation of the 1-Wasserstein distance
• Thanks to a property of $L_1$ norm, the number of constraints can be decreased from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$
• The linear constraint minimization problem is recast as a convex non-smooth optimization problem
• This convex non-smooth optimization problem is solved with the proximal splitting techniques SDMM
• Each iteration of the SDMM algorithm requires the solution of a Poisson's problem for which fast solvers are available

For a given number of SDMM iterations, we end up with a numerical strategy such that a linear computational complexity to compute the Kantorovich-Rubinstein norm allows efficiency.
Numerical Examples: 1D time-shift Ricker signals

No local minima
Same resolution

$L_2$ distance wrt time shift
KR distance wrt time shift

We recover a single minimum but we loose the convexity
Numerical Examples: 2 parameters case study

Velocity model with a depth gradient $V_p(z) = V_0 + \alpha z$

Surface acquisition with one source and 192 receivers

$L_2$ and $KR$ misfit functions are evaluated over a grid defined by

$V_0 \in [1750,2250]$ and $\alpha \in [0.4,0.9]$

with discretization steps $12.5 \text{ m.s}^{-1}$ and $0.015 \text{s}^{-1}$

Mulder and Plessix (2008)
Numerical Examples:
2 parameters case study

Mitigating second minima
Numerical Examples:  
Marmousi example  

(Lailly et al, 1991; Martin et al, 2006)

Surface acquisition with 128 sources each 125 m and 168 receivers each 100 m

Acoustic modeling engine for generating seismograms
Inversion using $L_2$ distance
Inversion using KR distance
Residues

Representation L2 of residues computed using KR distance
Influence of the initial model

![Exact model](image1)

![Initial model 1](image2)

![Initial model 2](image3)

![L2 results](image4)

![KR results](image5)
Computational costs

• Computation overhead per gradient (FISHPACK and 50 SDMM iterations): 3.8 s (19%)
  – L2 gradient computation time 20.6 s
  – KR gradient computation time 24.4 s

• Number of iterations
  – L2 inversion number of iterations: 83
  – KR inversion number of iterations: 439 (502 if noise)
Conclusion (1)

- Reconstructing subsurface wave velocities from seismic record requires to correctly account for **time-shifts**

- L2 misfit function fails to measure correctly these time-shifts and OT misfit function appears more robust to cycle skipping (still there!)

- We propose an implementation based on the **Kantorovich-Rubinstein norm** with
  - non conservation of mass between compared quantities
  - oscillatory (positive/negative) signals
  - linear complexity algorithms
Conclusion (2)

• The linear complexity is obtained through
  – the reduction to a linear number of constraints in the KR problem
  – the use of the SDMM proximal splitting technique

• Each iteration of the SDMM algorithm requires the solution of a Poisson problem which can be performed efficiently thanks to the use of FFT or multigrid based algorithms

• Other aspects: behaviour wrt noise; data dimensionality (1D -> 5D)

(Métivier et al, 2016a; Métivier et al, 2016b)
Thank you for your attention

IDRIS & TGCC, French national computing centers
CIMENT, Grenoble computing center
SEISCOPE sponsors: http://seiscope2.osug.fr

Questions?
Bibliography


KR residues

Through SDMM iterations 5, 10, 25, 50
Noise influence
Influence of the dataset

In 2D world, we may consider trace/trace (1D) or source/source (2D) or dataset/dataset (3D) for comparing data values ....
BP 2004 Benchmark
BP 2004 Benchmark
Chevron 2014 dataset. Common shot-gather for the source situated at $x = 0$ km for different frequency bands.
Chevron 2014 benchmark

Chevron 2014 starting P-wave velocity model
Chevron 2014 benchmark

Estimated P-wave velocity model at 4 Hz
Chevron 2014 benchmark

Estimated P-wave velocity model at 10 Hz
Chevron 2014 benchmark

Distance (km)

Depth (km)

*Estimated P-wave velocity model at 16 Hz*
Chevron 2014 benchmark

Estimated P-wave velocity model at 25 Hz
Chevron 2014 benchmark (L2)
Chevron 2014 benchmark (L2)

Estimated P-wave velocity model at 4 Hz
Chevron 2014 benchmark (L2)

Estimated P-wave velocity model at 10 Hz
Exact common shot-gather for the left most source at 25 Hz, compared to the corresponding synthetic in the final model at 25 Hz (orange panels). The synthetic data is mirrored and placed on both sides of the real data to better compare the match of the different phases.