Optimal transport for Seismic Imaging

Bjorn Engquist

In collaboration with Brittany Froese
and Yunan Yang

ICERM Workshop - Recent Advances in Seismic Modeling and Inversion:
From Analysis to Applications, Brown University, November 6-10, 2017

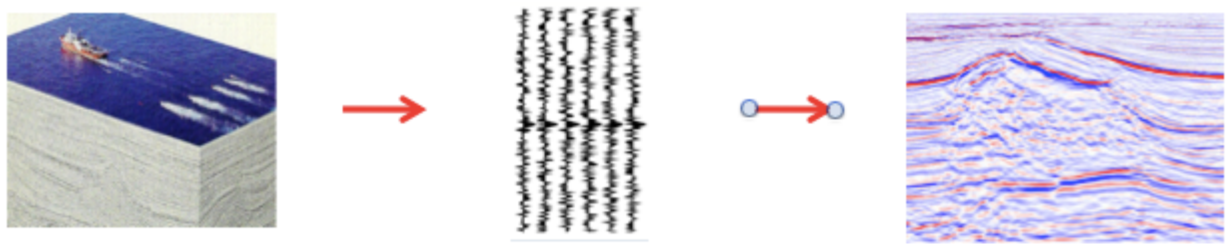
Outline

1. Remarks on Full Waveform Inversion (FWI)
2. Measures of mismatch
3. Optimal transport and Wasserstein metric
4. Monge-Ampère equation and its numerical approximation
5. Applications to full waveform inversion
6. Conclusions

“Background: matching problem and technique”

1. Remarks on Full Waveform Inversion (FWI)

- Full Waveform Inversion is an increasingly important technique in the inverse seismic imaging process



- It is a PDE constrained optimization formulation
- Model parameters *v* are determined to fit data

$$\min_{m(x)} \left(\left\| u_{comp}(m) - u_{data} \right\|_A + \lambda \left\| Lv \right\|_B \right)$$

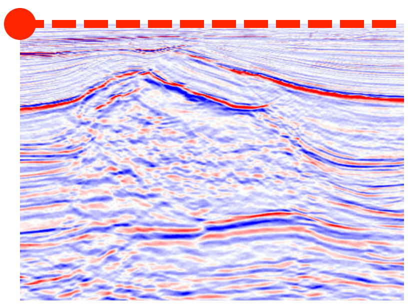
FWI: PDE constrained optimization

- FWI: Measured and processed data is compared to a computed wave field based on model parameters *v* to be determined (for example, *P*-wave velocity)

$$\min_{m(x)}\left(\left\|u_{comp}(m)-u_{data}\right\|_A+\lambda\left\|Lv\right\|_B\right)$$

- $||\,.\,||_A$ measure of mismatch
  - L_2 the standard choice
- $||\,Lc\,||_B$ potential regularization term, which we will omit for this presentation
- at the surface

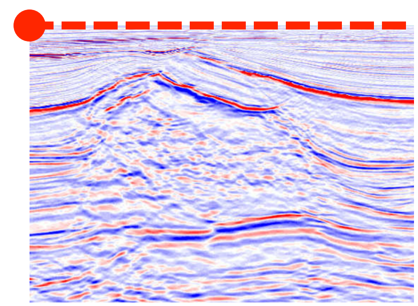
Over determined boundary conditions at the surface



Mathematical and computational challenges

- Important computational steps
- Relevant measure of mismatch (✔)
- Fast wave field solver
  - In our case scalar wave equation in time or frequency domain, for example

$$u_{tt} = m(x)^2 \Delta u,$$



- Efficient optimization
  - Adjoint state method for gradient computation

2. Measures of mismatch

- We will denote the computed wave field by f(x,t;v) and the data by $g(x,t)$,

$$u_{comp}(x,t;m) = f(x,t;m),\ \ u_{data}(x,t) = g(x,t)$$

- The common and original measure of mismatch between the computed signal f and the measured data g is L_2, [Tarantola, 1984, 1986]

$$\min_m \left\| f_m - g \right\|_{L_2}$$

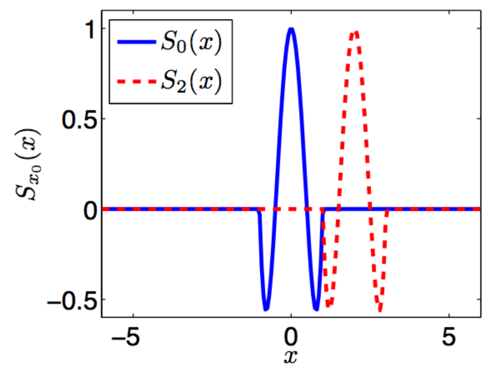
- We will make some remarks on different Measures of mismatch starting with local estimates to more global

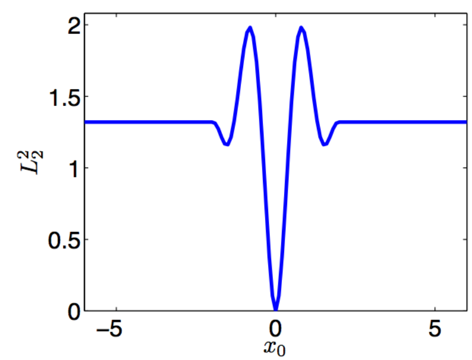
Global minimum

- It can be expected that the mismatch functional will have local minima that complicates minimization algorithms
- Ideally, local minima different from the global min should be avoided for some natural parameterizations as “shift” and “dilations” ($f\,(t) = g(at - s)$)
- Shift as a function of *t*, dilation as a function of *x*

$$u_{tt} = m^2 u_{xx}, \quad x > 0, t > 0$$

$$u(0,t) = u_0(t) \rightarrow u = u_0(t - x/m)$$





Local measures

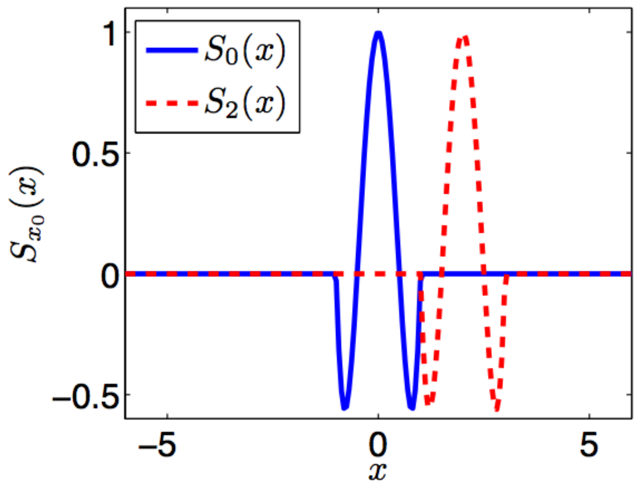
- In the L_2 local mismatch, estimators *f* and *g* are compared point wise,

$$J(v) = \|f - g\|_{L_2} = \left(\sum_{i,j} \left| f(x_i, t_j) - g(x_i, t_j) \right|^2 \right)^{1/2}$$

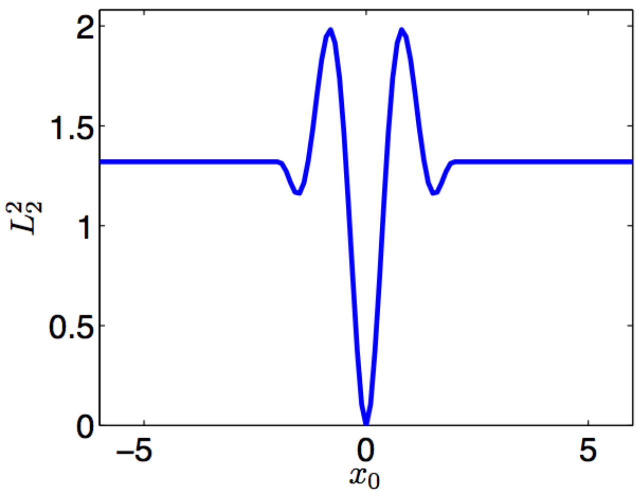
- This works well if the starting values for the model parameters are good otherwise there is risk for trapping in local minima "cycle skipping"

"Cycle skipping"

- The need for better mismatch functionals can be seen from a simple shift example – small basin of attraction
- For other examples, [Vireux et al 2009]



Shift or displacement



"Cycle skipping"
Local minima

$L_2{}^2$

Global measures

- Different measures have been introduced to to compare all of *f* and *g* – not just point wise.
- Integrated functions
  - NIM [Liu et al 2014]
  - [Donno et al 2014]
- Stationary marching filters
  - Example AWI, [Warner et al 2014]
- Non-stationary marching filters
  - Example [Fomel et al 2013]
- Measures based on optimal transport (✔)

Integrated functions

- f and g are integrated, typically in 1D-time, before L_2 comparison

$$J = \left\| F - G \right\|_{L_2} = \left(\sum\nolimits_{i,j} \left| F(x_i, t_j) - G(x_i, t_j) \right|^2 \right)^{1/2},$$

$$F_{i,j} = \sum\nolimits_{k=1}^{j} f(x_i, t_k), \quad G_{i,j} = \sum\nolimits_{k=1}^{j} g(x_i, t_k),$$

- In mathematical notation this is the H^{-1} semi-norm
- Slight increase in wave length for short signals (Ricker wavelet)
- Often applied to modified signals like squaring scaling or envelope to have f and g positive and with equal integral

Matching filters

- The filter based measures typically has two steps
  - Computing filter coefficients *K*

$$K = \operatorname{argmin} \left\| K * f - g \right\|_{L_2}$$

  - Estimation of difference between computed filter and the identity map.

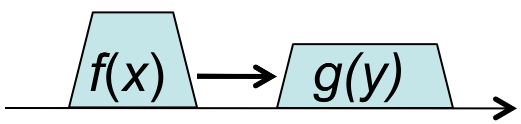
$$\left\| K - I \right\|$$

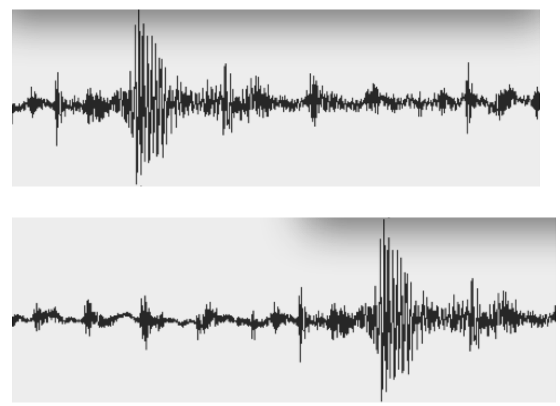
- The filter can be stationary or non-stationary
- The optimal transport based techniques Does this in one step
  - Minimization is of a measure of transform *K* or as it is called transport

3. Optimal transport and Wasserstein metric

- Wasserstein metric measures the "cost" for optimally transport one measure (signal) *f* to the other: *g* – Monge-Kantorivich optimal transport measure

"earth movers distance"
in computer science

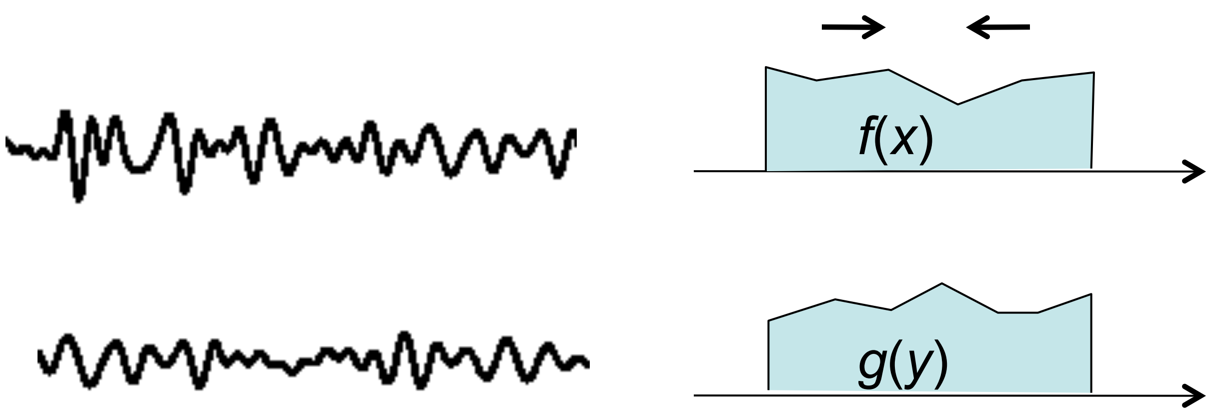




Compare travel time distance
Classic in seismology

Optimal transport and Wasserstein metric

- The Wasserstein metric is directly based on one cost function
- Signals in exploration seismology are not as clean as above and a robust functional combining features of L_2 and travel time is desirable
- Extensive mathematical foundation



Wasserstein distance

$$W_p(f,g) = \left(\inf_{\gamma} \int_{X \times Y} d(x,y)^p \, d\gamma(x,y) \right)^{1/p}$$

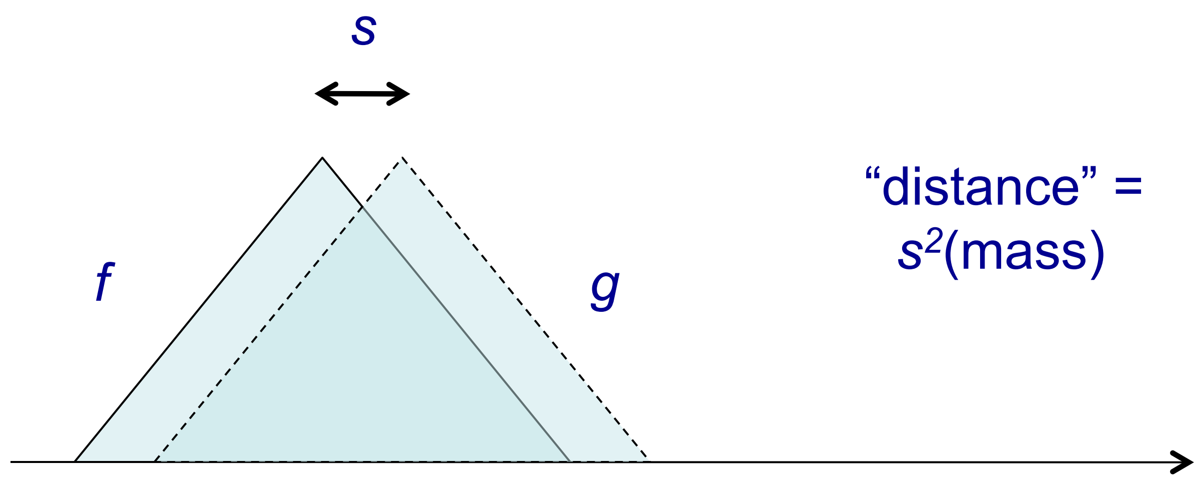
$$\gamma \in \Gamma \subset X \times Y, \textit{the set of product measure}: f \textit{ and } g$$

$$\int_X f(x)\,dx = \int_Y g(y)\,dy, \quad f,\ g \ge 0$$

$$W_2(f,g) = \left(\inf_{T_{f,g}} \int_X \left\| x - T_{f,g}(x) \right\|_2^2 f(x)\,dx \right)^{1/2}$$

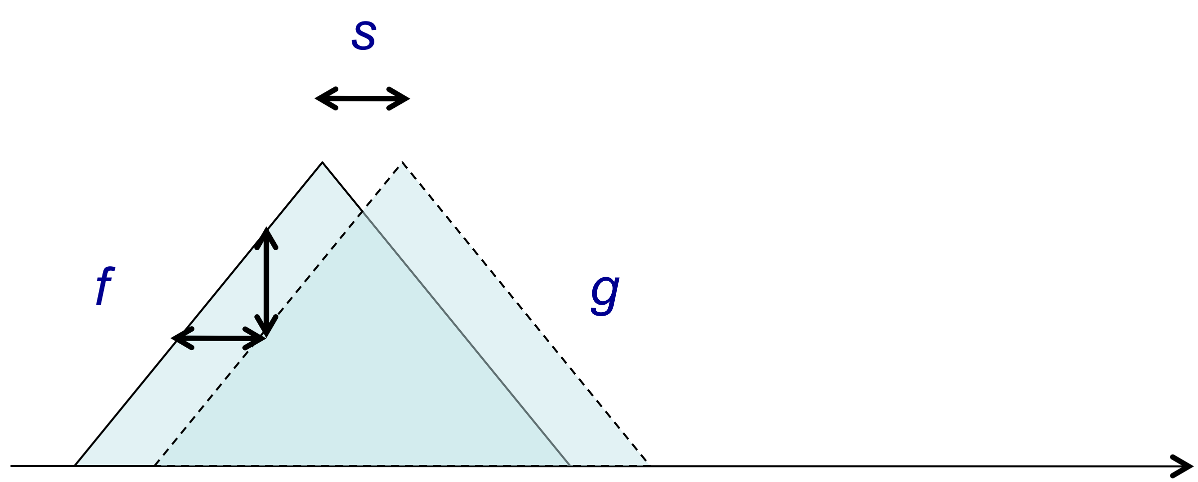
- Here the "plan" *T* is the optimal transport map from positive Borel measures *f* to g of equal mass
- Well developed mathematical theory, [Villani, 2003, 2009]

Wasserstein distance



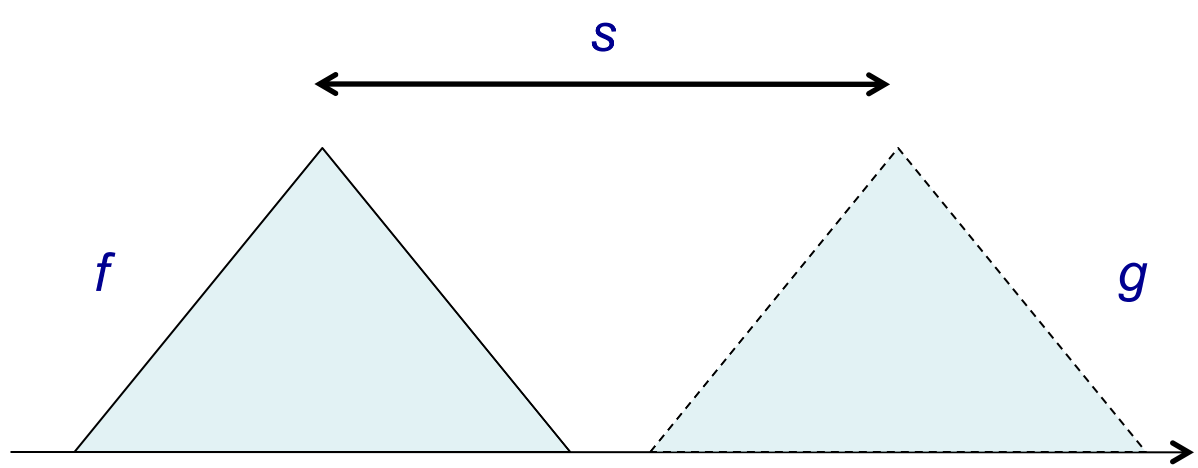
"distance" =
s^2(mass)

Wasserstein distance



- In this model example W_2 and L_2 is equal (modulo a constant) to leading order when separation distance *s* is small. Recall L_2 is the standard measure

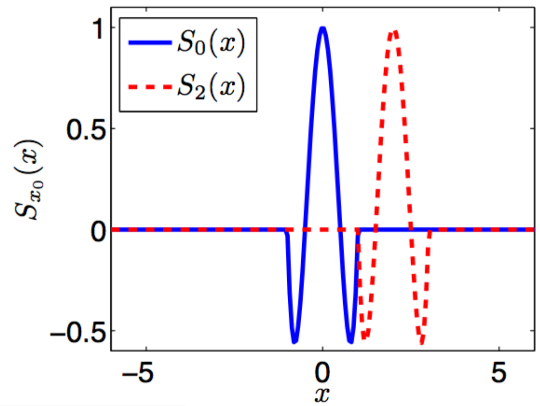
Wasserstein distance

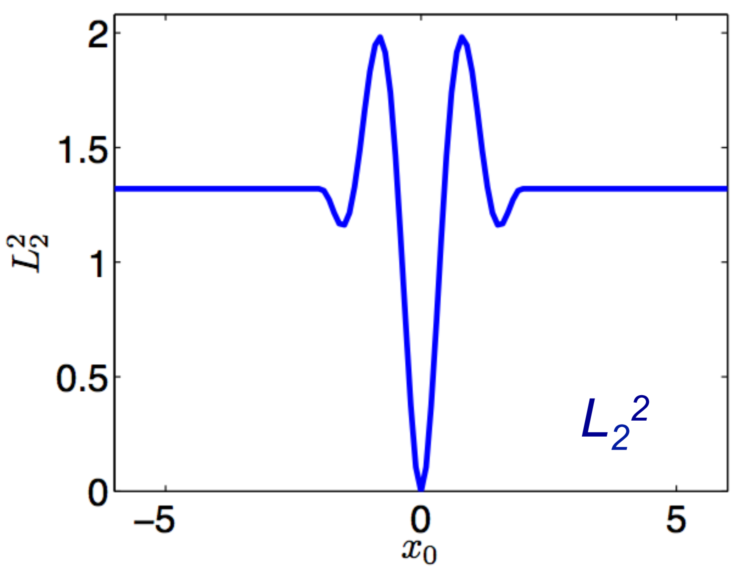


- When *s* is large W_2 = *s* = travel distance (time), ("higher frequency"), L_2 independent of s
- Potential for avoiding cycle skipping

Wasserstein distance vs L_2

- Fidelity measure



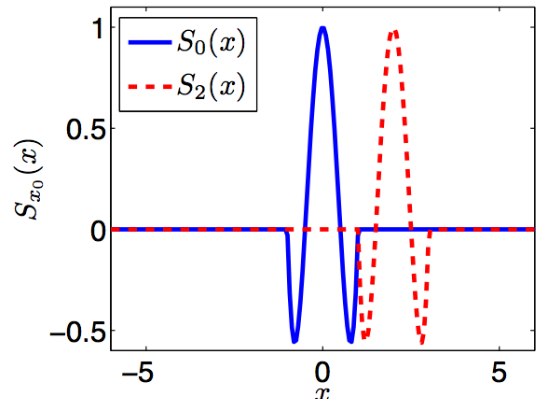


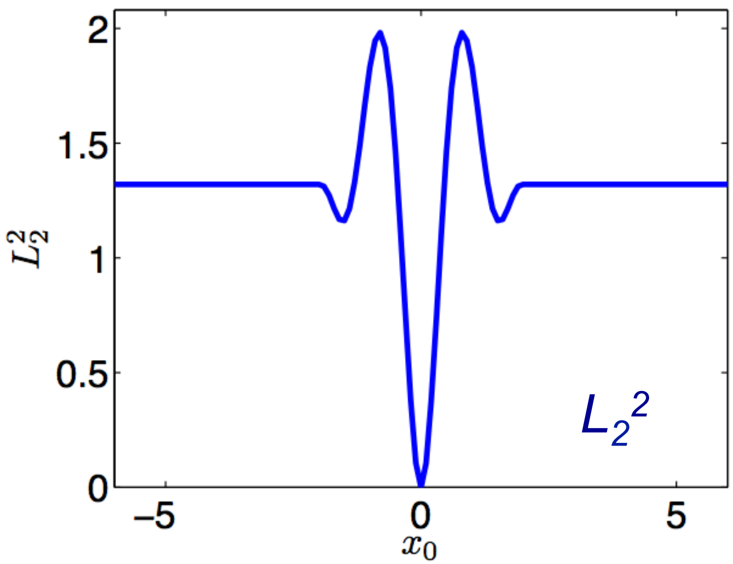
"Cycle skipping"
Local minima

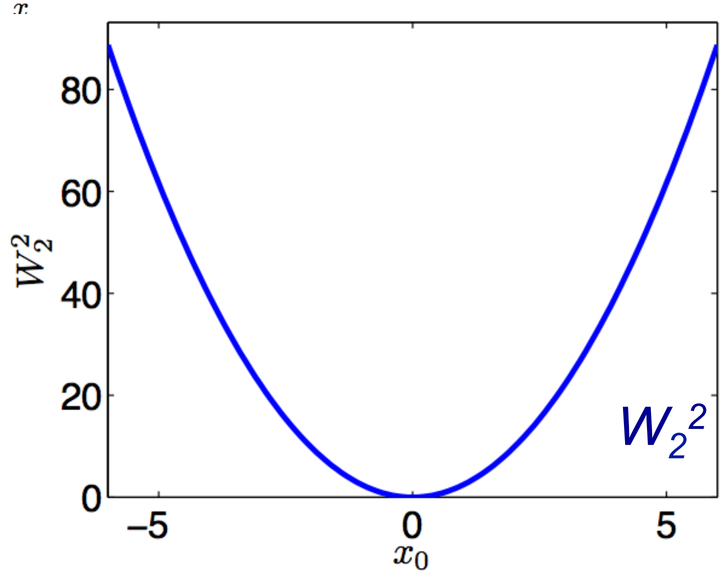
Function of displacement

Wasserstein distance vs L_2

- Fidelity measure

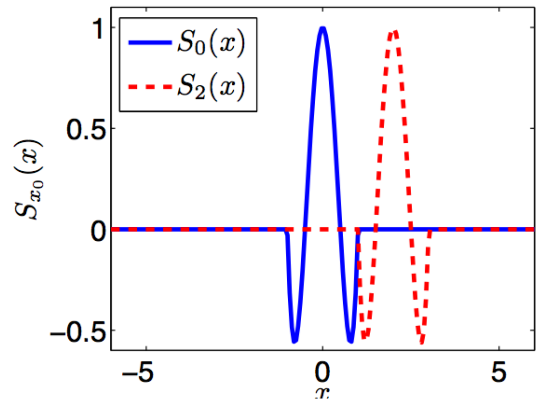


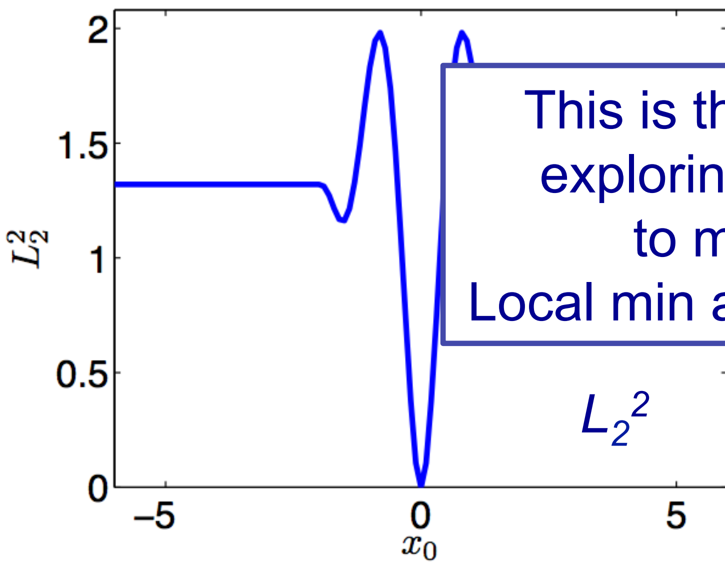


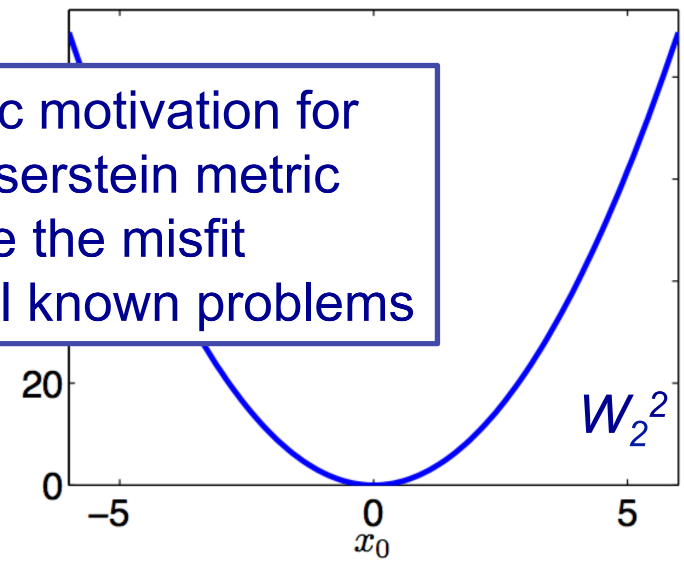


Wasserstein distance vs L_2

- Fidelity measure





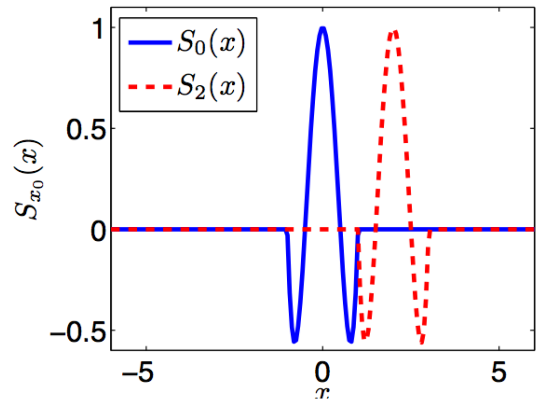


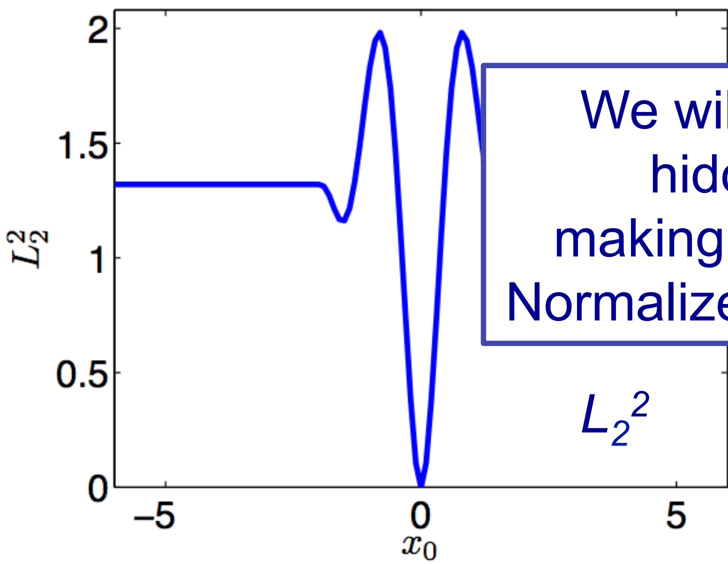
This is the basic motivation for exploring Wasserstein metric to measure the misfit

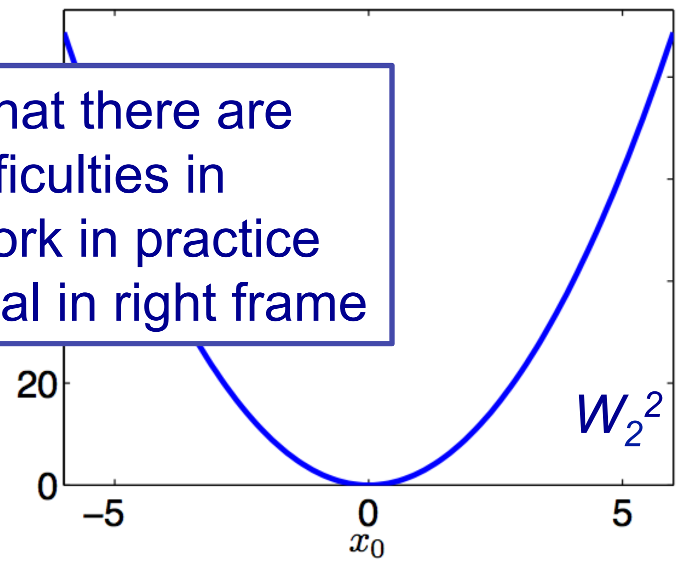
Local min are well known problems

Wasserstein distance vs L_2

- Fidelity measure







We will see that there are
hidden difficulties in
making this work in practice
Normalized signal in right frame

Analysis

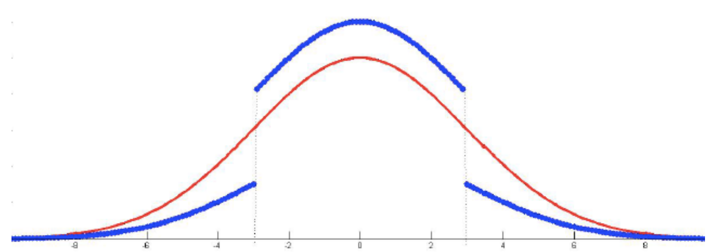
- Theorem 1: $W_2^{\,2}$ is convex with respect to translation, *s* and dilation, *a,*

$$W_2^2(f,g)[\alpha,s],\ \ f(x)=g(ax-s)\alpha^d,\ \ a>0,\, x,\, s\in R^d$$

- Theorem 2: $W_2^{\,2}$ is convex with respect to local amplitude change*, λ*

$$W_2^2(f,g)[\beta],\ \ f(x)=\begin{cases} g(x)\lambda, & x\in\Omega_1 \\ \beta g(x)\lambda, & x\in\Omega_2 \end{cases}\quad \beta\in R,\ \ \Omega=\Omega_1\cup\Omega_2$$

$$\lambda=\int_\Omega g\,dx \Big/ \left(\int_{\Omega_1} g\,dx+\beta\int_{\Omega_2} g\,dx\right)$$



- (L_2 only satisfies 2nd theorem)

Remarks

- The scalar dilation *ax* can be generalized to *Ax* where *A* is a positive definite matrix. Convexity is then in terms of the eigenvalues
- The proof of theorem 1 is based on c-cyclic monotonicity

$$\{(x_j, x_j)\} \in \Gamma \to \sum_j c(x_j, x_j) \leq \sum_j c(x_j, x_{\sigma(j)})$$

- The proof of theorem two is based on the inequality

$$W_2^2(sf_1 + (1-s)f_2, g) \leq sW_2^2(f_1, g) + (1-s)W_2^2(f_2, g)$$

Illustration: discrete proof (theorem 1)

- Equal point masses then weak limit for generl theorem alternative to using the c-cyclic propery

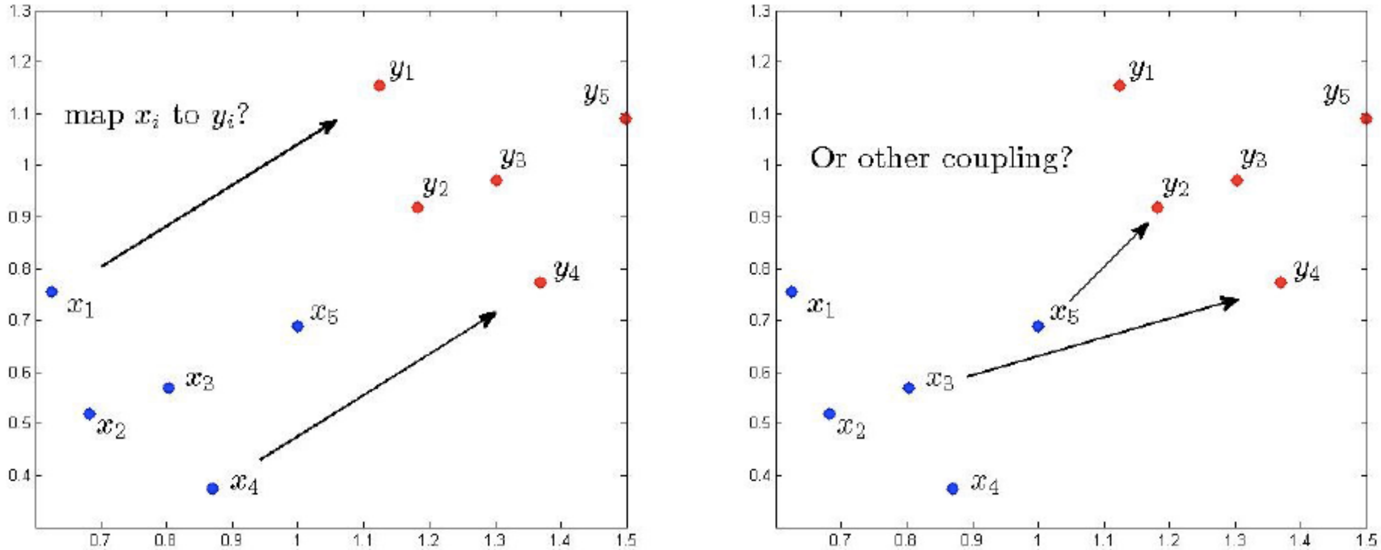


Illustration: discrete proof

$$W_2^2 = \min_{\sigma} \sum_{j=1}^{J} \left| x_{o_j} - (x_j - s\xi) \right|^2 = \quad (\sigma : \; permutation)$$

$$\min_{\sigma} \left(\sum_{j=1}^{J} \left| x_{o_j} - x_j \right|^2 - 2s \sum_{j=1}^{J} \left(x_{o_j} - x_j \right) \cdot \xi + J \left| s\xi \right|^2 \right) =$$

$$\min_{\sigma} \left(\sum_{j=1}^{J} \left| x_{o_j} - x_j \right|^2 + J \left| s\xi \right|^2 \right), \quad from \;\; \sum_{j=1}^{J} x_{o_j} = \sum_{j=1}^{J} x_j$$

$$\rightarrow \;\; x_{o_j} = x_j \rightarrow \sigma_j = j$$

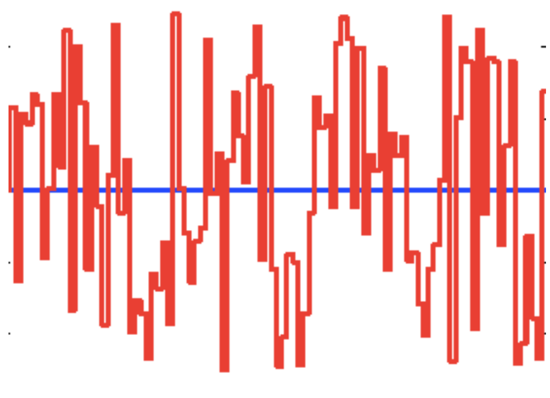
Noise

- W_2^2 less sensitive to noise than L_2
- Theorem 3: $f = g + \delta$, δ uniformly distributed uncorrelated random noise, ($f > 0$), discrete i.e. piecewise constant: N intervals

$$\|f-g\|_{L_2}^2 = O\ (1), \quad W_2^2\left(f-g\right) = O(N^{-1})$$

$$f = \left(f_1, f_2, \ldots, f_J\right)$$

- Proof by "domain decomposition" dimension by dimension and standard deviation estimates using closed 1D formula



Computing the optimal transport

- In 1D, optimal transport is equivalent to sorting with efficient numerical algorithms O(Nlog(N)) complexity, N data points

$$W_2(f,g) = \int \left(F^{-1}(y) - G^{-1}(y) \right) dy$$

$$F(x) = \int^x f(\xi)\,d\xi, \quad g(x) = \int^x g(\xi)\,d\xi$$

- In higher dimensions such combinatorial methods as the Hungarian algorithm are very costly O(N^3), Alternatives: linear programming, sliced Wasserstein, ADMM

Computing of optimal transport

- For higher dimensions fortunately the optimal transport related to W_2 can be solved via a Monge-Ampère equation [Brenier 1991, 1998]

$$W_2(f,g) = \left(\int_X \| x - \nabla u(x) \|_2^2 \, f(x)\, dx \right)^{1/2}$$

$$\det\left(D^2(u) \right) = f(x) / g(\nabla u(x))$$

$$Brenier\ map\ T(x) = \nabla u(x)$$

- Recently there are now alternative PDF formulations

4. Monge-Ampère equation and its numerical approximation

- Nonlinear equation with potential loss of regularity
- Weak viscosity solution *u* if *u* is both a sub and super solution

$$\det\left(D^2(u)\right) - f(x) = 0,\ u\ convex,\ f \in C^0(\Omega)$$

- Sub solution (super analogous)

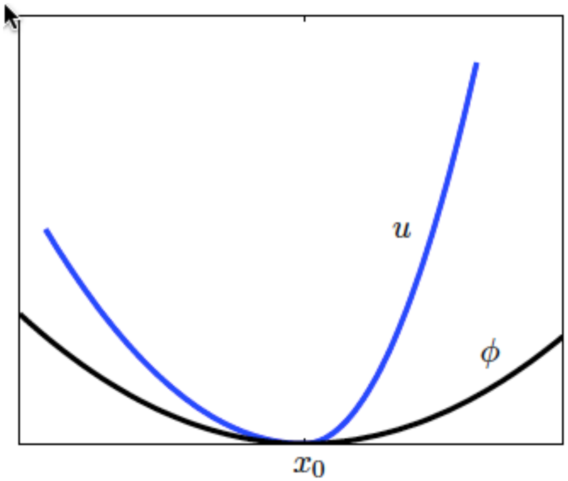
$$x_0 \in \Omega,\ if\ local\ \max\ of\ u - \phi,\ then$$

$$\det\left(D^2\phi\right) \le f(x_0)$$

- 1D

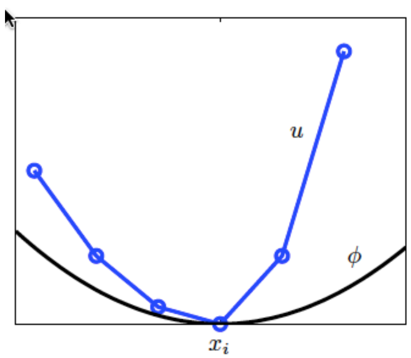
$$u_{xx} = f,\ \phi(x_0) = u(x_0), \phi(x_0) = u(x_0),$$

$$\phi(x) \le u(x) \rightarrow \phi_{xx} \le f$$



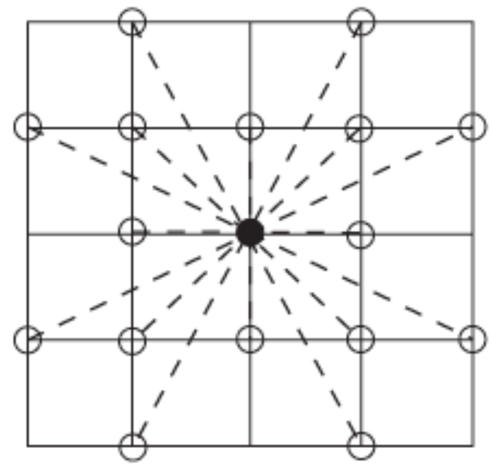
Numerical approximation

- Consistent, stable and monotone finite difference approximations will converge to Monge-Ampère viscosity solutions [Barles, Souganidis, 1991]



$$\det\left(D^2 u\right) = \prod_{j=1}^{d}\left(u_{v_j v_j}\right)^{+}$$

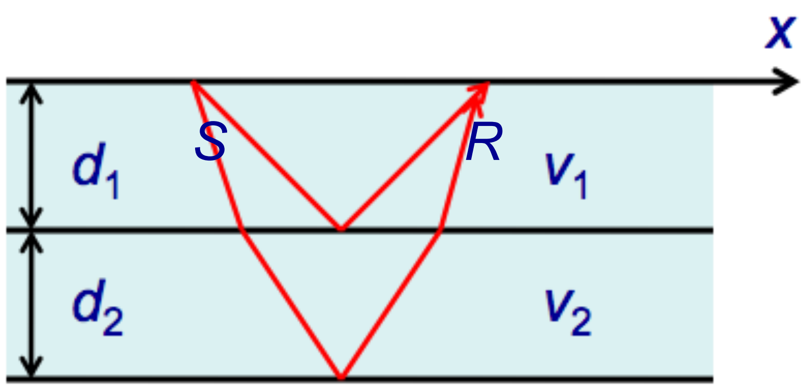
$$\left\{v_j\right\} : eigenvectors\ of\ D^2 u$$

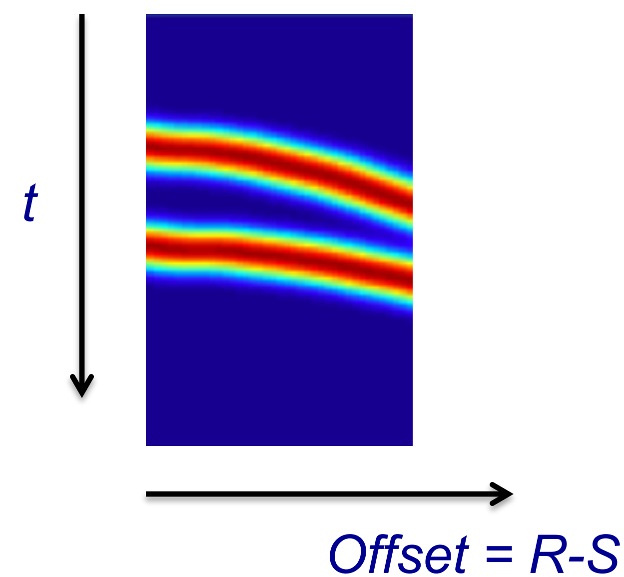


[Benamou, Froese, Oberman, 2014]

5. Applications to full waveform inversion

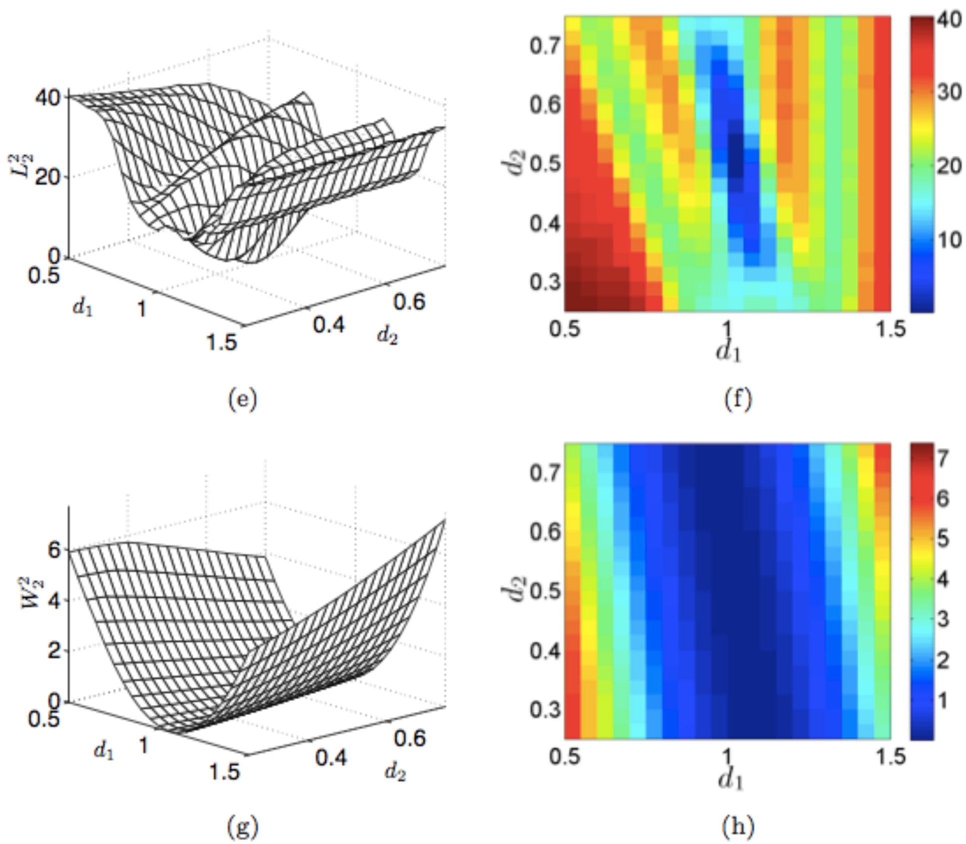
- First example: Problem with reflection from two layers – dependence on parameters, with Froese.





- Robust convergence with direct minimization algorithm (Simplex) geometrical optics and positive *f, g*

Reflections and inversion example



L_2

W_2

Gradient for optimization

- For large scale optimization, gradient of $J(f) = W_2^2(f,g)$ with respect to wave velocity is required in a quasi Newton method in the PDE constrained optimization step
- Based on linearization of J and Monge-Ampère equation resulting in linear elliptic PDE (adjoint source)

$$J + \delta J = \int (f + \delta f) \left\| x - \nabla(u_f + \delta u) \right\|^2 dx$$

$$f + \delta f = g(\nabla(u_f + \delta u)) \det(D^2(u_f + \delta u))$$

$$L(v) = g(\nabla u_f) tr((D^2 u_f)^{\bullet} D^2(v)) + \det(D^2 u_f) g(\nabla u_f) \cdot \nabla v = \delta f$$

W2 Remarks

+ Captures important features of distance in both travel time and L_2, Convexity with respect to natural parameters
+ There exists fast algorithms and technique robust vs. noise
- Constraints that are not natural for seismology

$$\int_X f(x)\,dx = \int_Y g(y)\,dy, \quad f,\ g \geq 0, \quad g > 0, M - A$$

- Normalize: transform *f, g* to be positive and with the same integral

Large scale applications

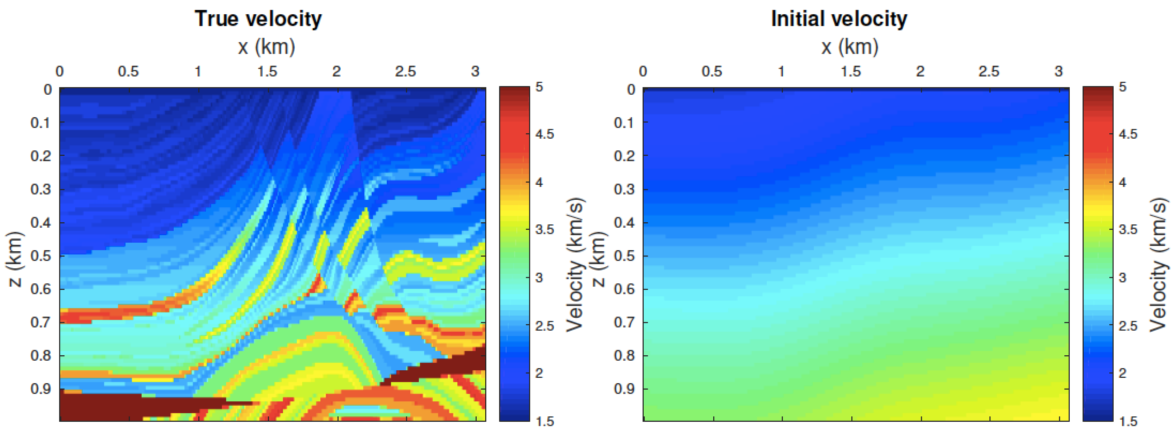
- Early normalizations: squaring, consider positive and negative parts of *f* and *g* separately – not appropriate for adjoint state technique
- Successful normalization – linear

$$\tilde{f}(x) = (f(x)+c) / \int \left(f(x)+c\right) dx, \quad \tilde{g}(x) = \ldots$$

- Efficient alternative to W_2 (2D): trace by trace W_2 (1D) coupled to L_2 [Yang et al 2016]
- Other alternatives, W_1, unbalanced transport [Chizat et al 2015], Dual formulation of optimal transport
- Normalized $W_2 + \lambda L_2$ is an unbalanced transport measure, $\lambda > 0$

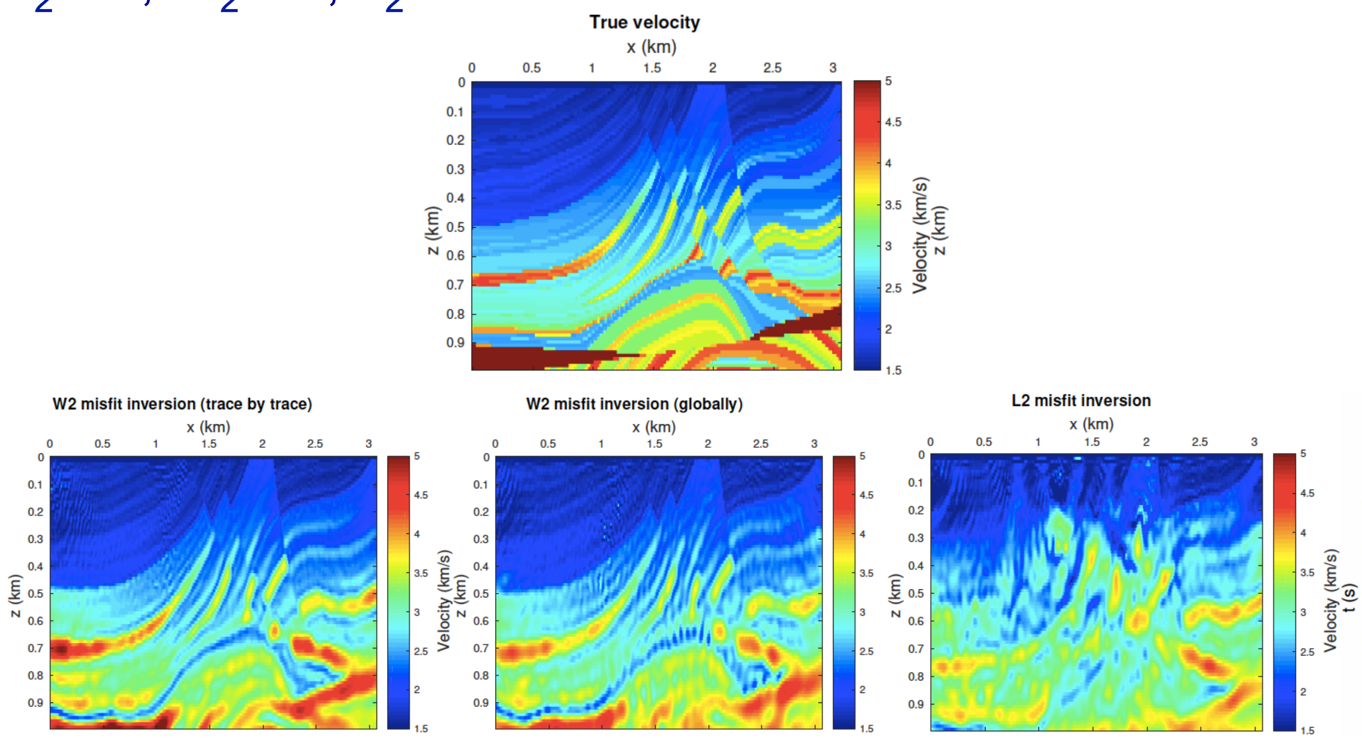
Applications Seismic test cases

- Marmousi model (velocity field)
- Original model and initial velocity field to start optimization



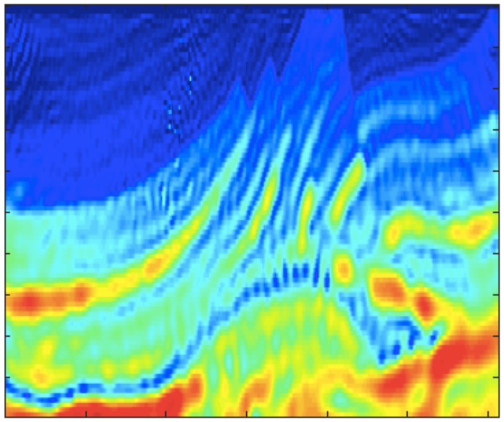
Marmousi model

- Original and FWI reconstruction with different initializations: W_2-1D, W_2-2D, L_2



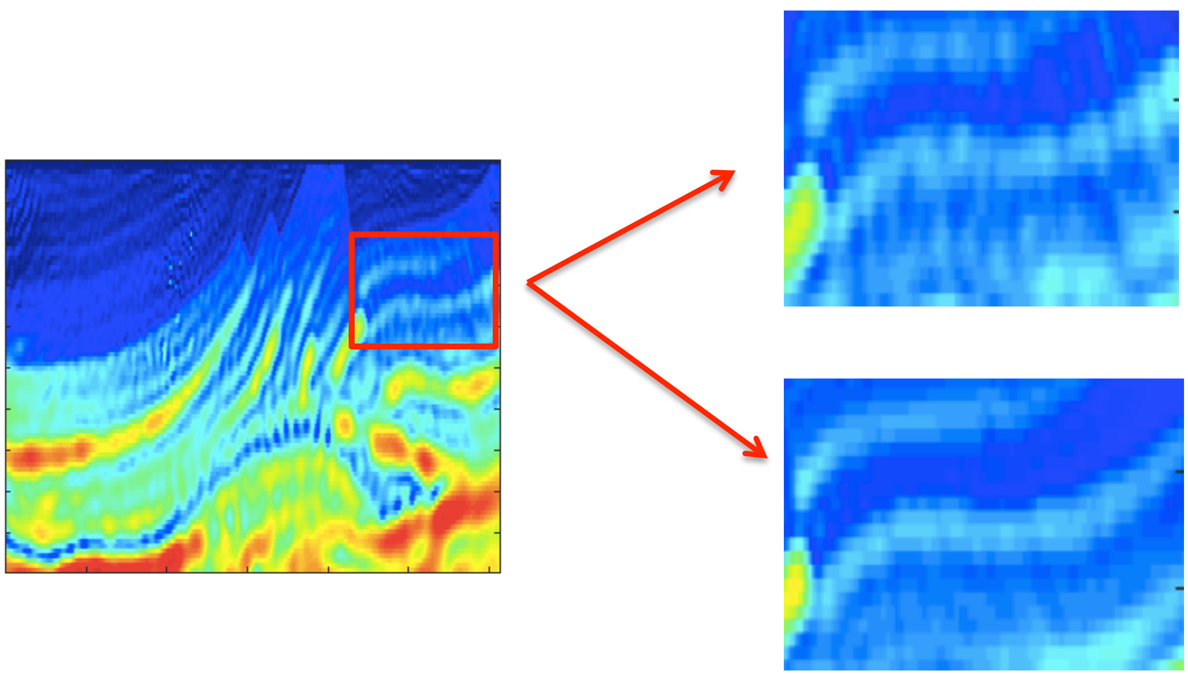
Marmousi model

- Robustness to noise: good for data but allows for oscillations in "optimal" computed velocity, numerical M – A errors
- Remedy: trace by trace, TV - regularization



Marmousi model

- Robustness to noise: good for data but allows for oscillations in “optimal” computed velocity, numerical M – A errors
- Remedy: trace by trace, TV - regularization

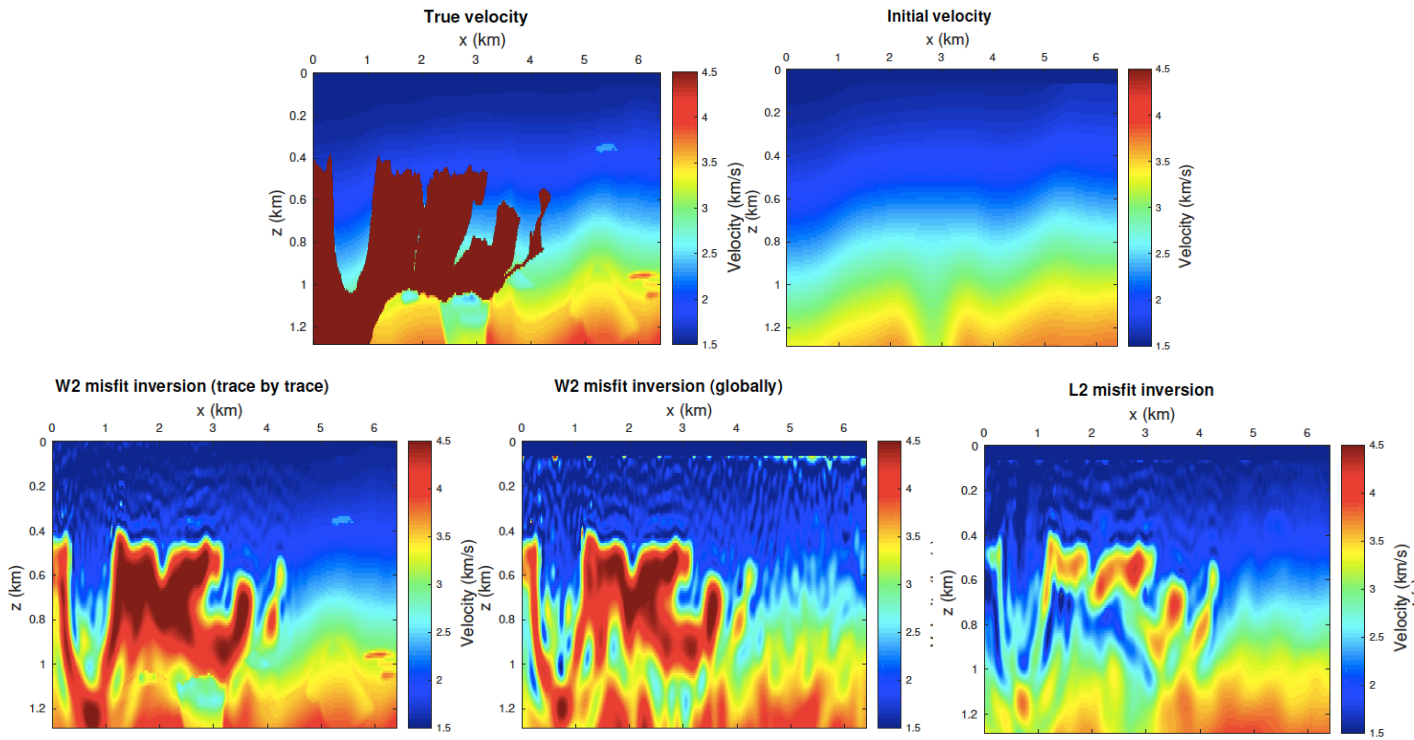


W_2

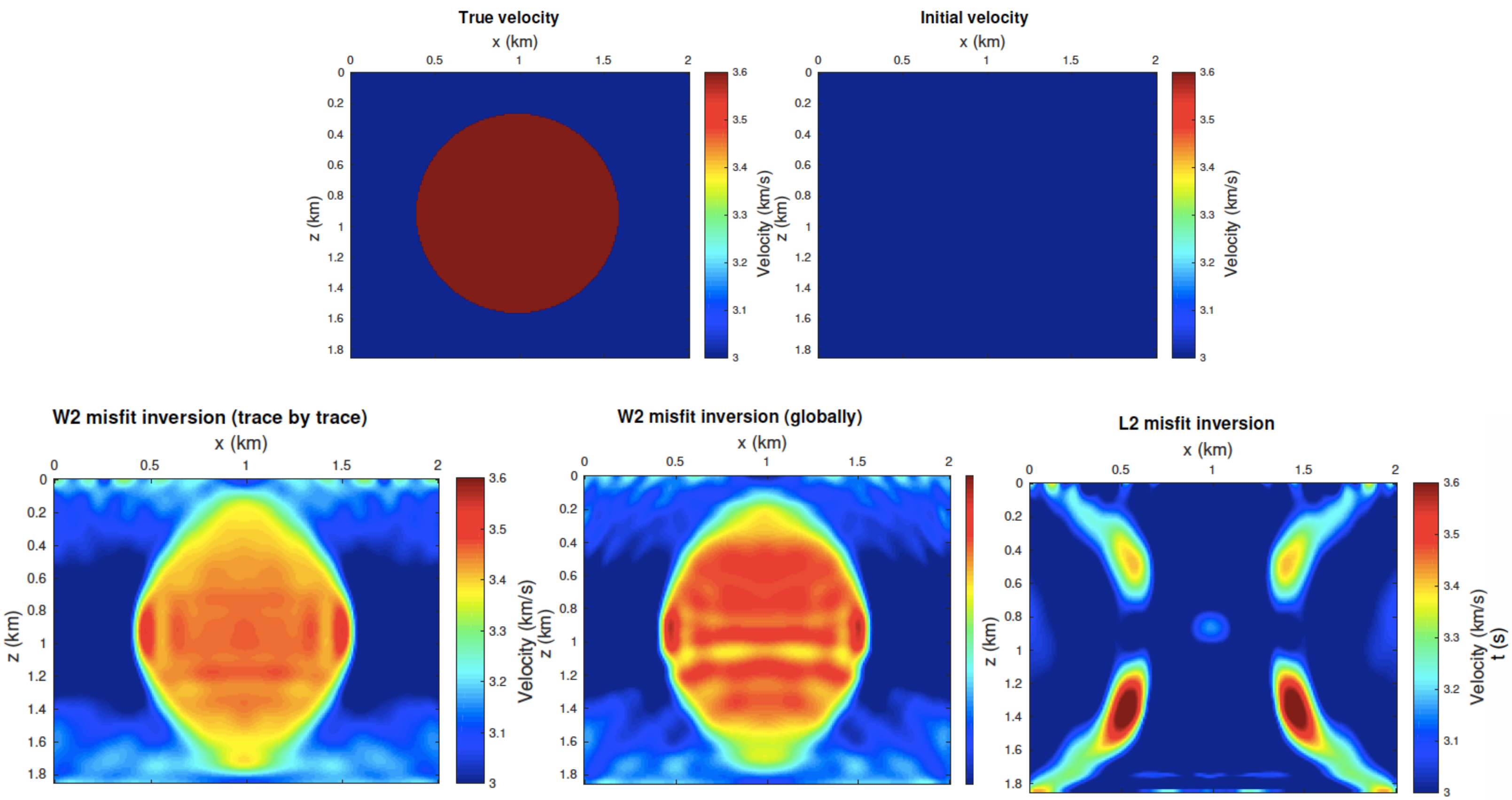
W_1

BP 2004 model

- High contrast salt deposit, W_2 - 1D, W_2 - 2D, L^2



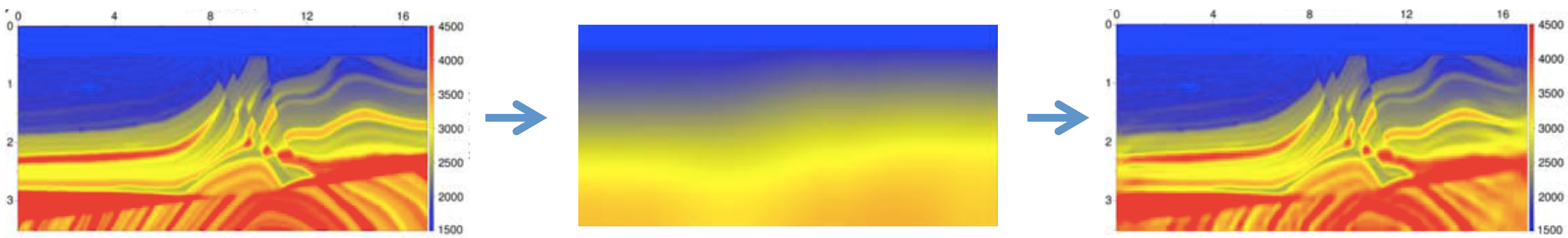
Camembert



True velocity
Initial velocity
W2 misfit inversion (trace by trace)
W2 misfit inversion (globally)
L2 misfit inversion
x (km)
z (km)
Velocity (km/s)
t (s)

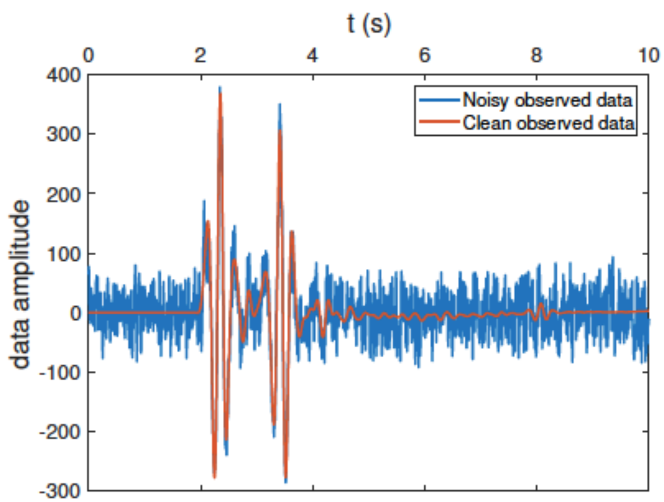
W_1 example

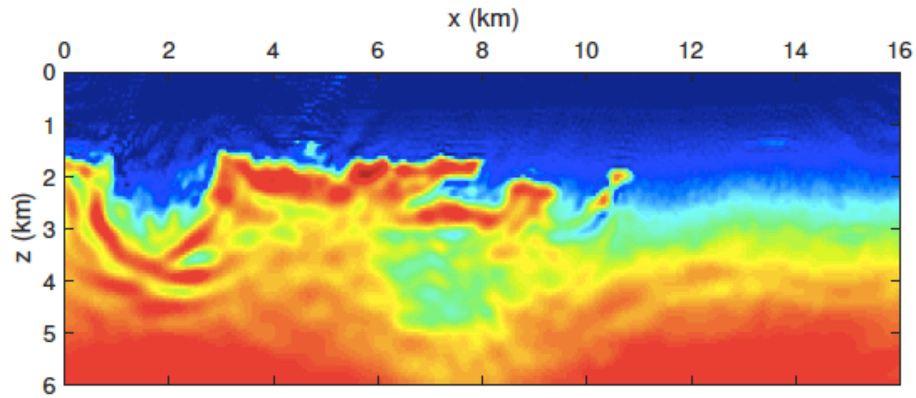
- Example below: W_1 measure and Marmousi p-velocity model [Metivier et. al, 2016]
- Similar quality but more sensitive to noise and $\neq L_2$ when $f \approx g$
- Solver with better 2D performance



W_2 with noise

- Slightly temporally correlated uniformly distributed noise



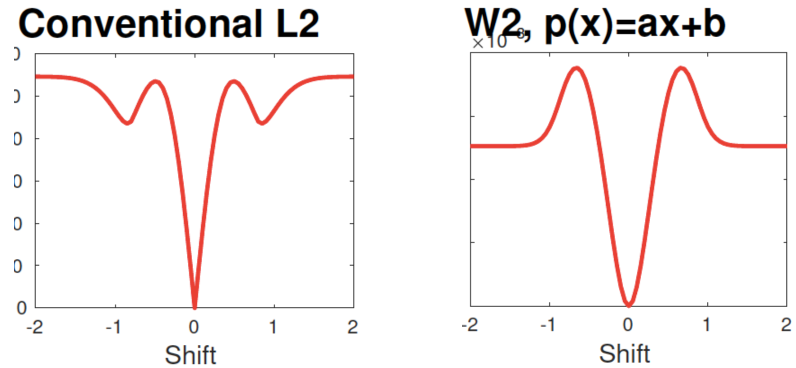


Remarks

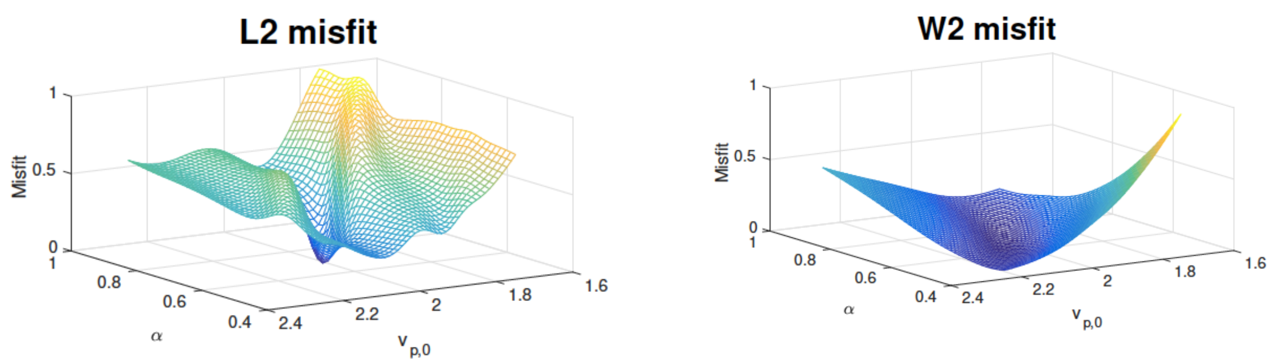
- Troubling issues
  - Theoretical results of convexity based on normalized signals of the form squaring etc. but not practically useful (squaring: not sign sensitive, requires compact support and problem with adjoint state method)
  - The practically working normalization based on linear normalization does not satisfy convexity with respect to shifts

Remarks

- Linear scaling: misfit as function of shift, Ricker wavelet



- Function of velocity parameters: $v = v_p + \alpha z$ [Metivier et. al, 2016]



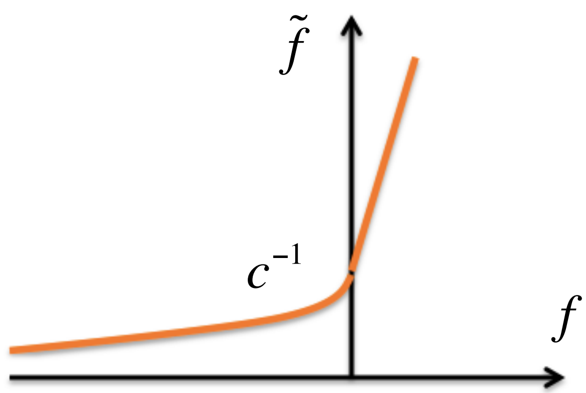
Remarks

- New and currently best normalization

$$\tilde{f}(x) = \hat{f}(x) / \int \hat{f}(x)\,dx, \quad \hat{f}(x) = \begin{cases} f(x) + c^{-1}, & x \geq 0 \\ c^{-1} \exp(c f(x), & x < 0 \end{cases}$$

- Good in practice – allows for less accurate initial mode than linear scaling
- Satisfies our theorems

  for c large enough



6. Conclusions

- Optimal transport and the Wasserstein metric are promising tools in seismic imaging
- Theory and basic algorithms need to be modified to handle realistic seismic data
- Ready for field data, [PGS, SEG2017, North Sea]