Mitigating *local* minima in full-waveform inversion by *expanding* the *search* space with the *penalty* method

Felix J. Herrmann

SLIM
University of British Columbia

Submitted for publication & supporting theory submitted in "A penalty method for PDE-constrained optimization".

A penalty method for waveform inversion

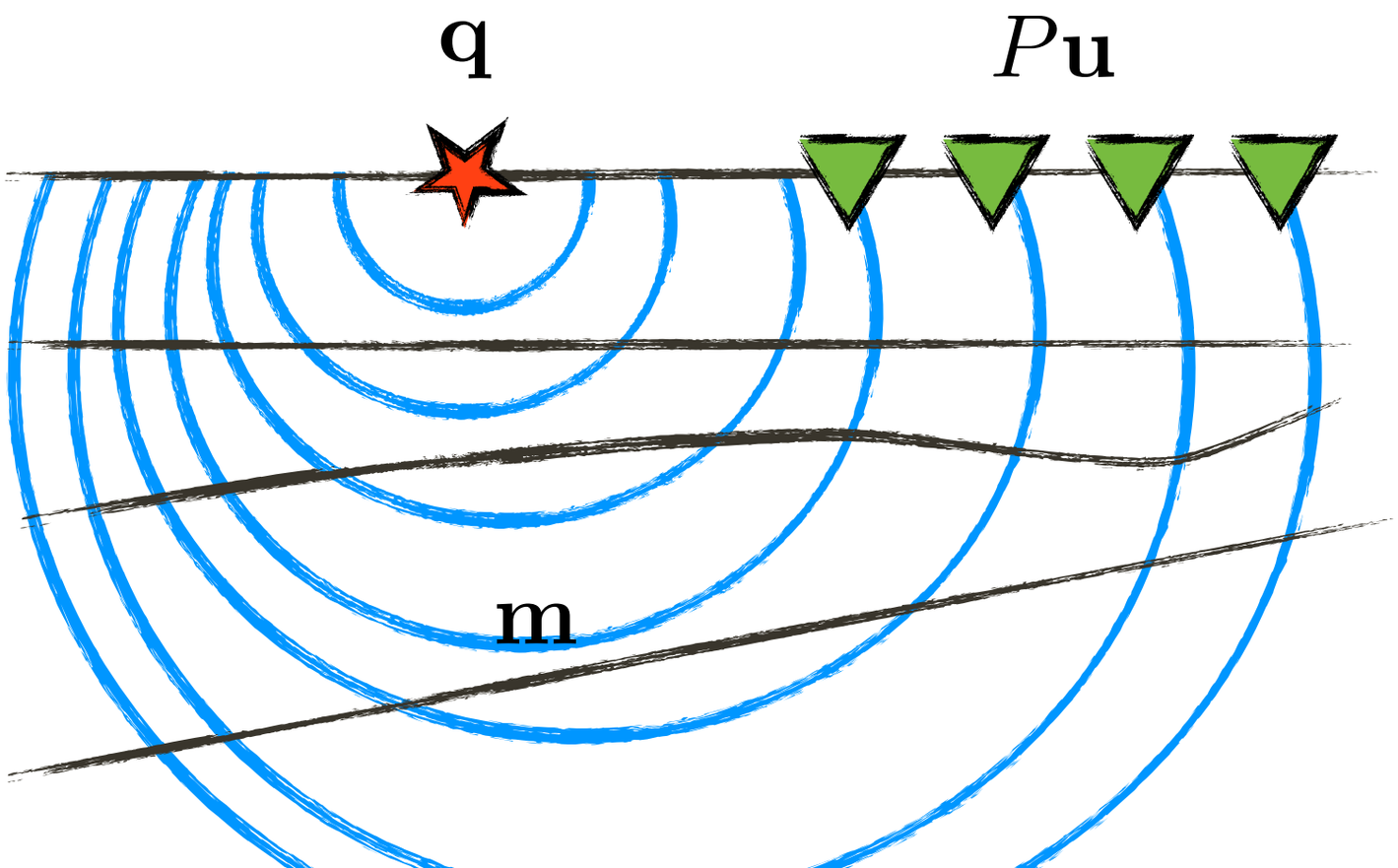
Tristan van Leeuwen & Felix J. Herrmann

SLIM
University of British Columbia

US Provisional Patent Application No. 61/815,533
Title: A Penalty Method for PDE-Constrained Optimization with Applications to Wave-Equation Based Seismic Inversion

Seismic waveform inversion

Given *observed* data, $\mathbf{d}$, find *model* parameters, $\mathbf{m}$
and a *wavefield*, $\mathbf{u}$, such that $P\mathbf{u} \approx \mathbf{d}$ and $A(\mathbf{m})\mathbf{u} \approx \mathbf{q}$



Observation

If we measure the wavefield *everywhere,* we can *trivially* solve $P\mathbf{u} = \mathbf{d}$.

We can then *find* the *medium* parameters by *solving*

$$A(\mathbf{m})\mathbf{u} = \mathbf{q}$$

for $\mathbf{m}$ *knowing* $\mathbf{q}$.

This would lead to $\min_{\mathbf{m}} ||A(\mathbf{m})P^{-1}\mathbf{d} - \mathbf{q}||_2^2$

as opposed to $\min_{\mathbf{m}} ||PA(\mathbf{m})^{-1}\mathbf{q} - \mathbf{d}||_2^2$

Question

When P is *not* invertible, can we *reconstruct* a *wavefield* from the *data* that is *useful*?

For a given $\mathbf{m}$, we want the wavefield to
i) *fit* the *data*, $P\mathbf{u} \approx \mathbf{d}$, and
ii) *obey* the *wave*-equation, $A(\mathbf{m})\mathbf{u} \approx \mathbf{q}$

Why *not* use the least-squares solution?!

$$\begin{pmatrix} P \\ A(\mathbf{m}) \end{pmatrix} \mathbf{u} \approx \begin{pmatrix} \mathbf{d} \\ \mathbf{q} \end{pmatrix}$$

Naive algorithm

1. Reconstruct the *wavefield* by solving

$$\mathbf{u}_k = \arg\min_{\mathbf{u}} \left\| \begin{pmatrix} P \\ A(\mathbf{m}_k) \end{pmatrix} \mathbf{u} - \begin{pmatrix} \mathbf{d} \\ \mathbf{q} \end{pmatrix} \right\|_2^2$$

2. Reconstruct the *model* by solving

$$\mathbf{m}_{k+1} = \arg\min_{\mathbf{m}} \|A(\mathbf{m})\mathbf{u}_k - \mathbf{q}\|_2^2$$

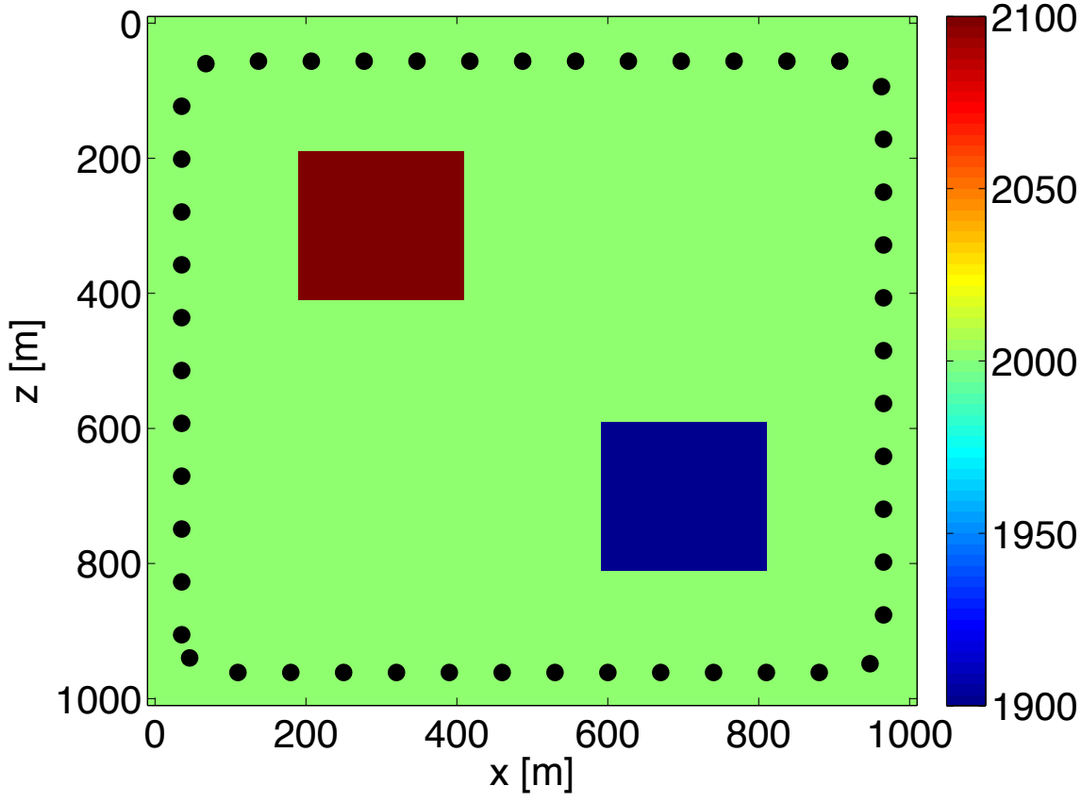
3. Repeat

Example

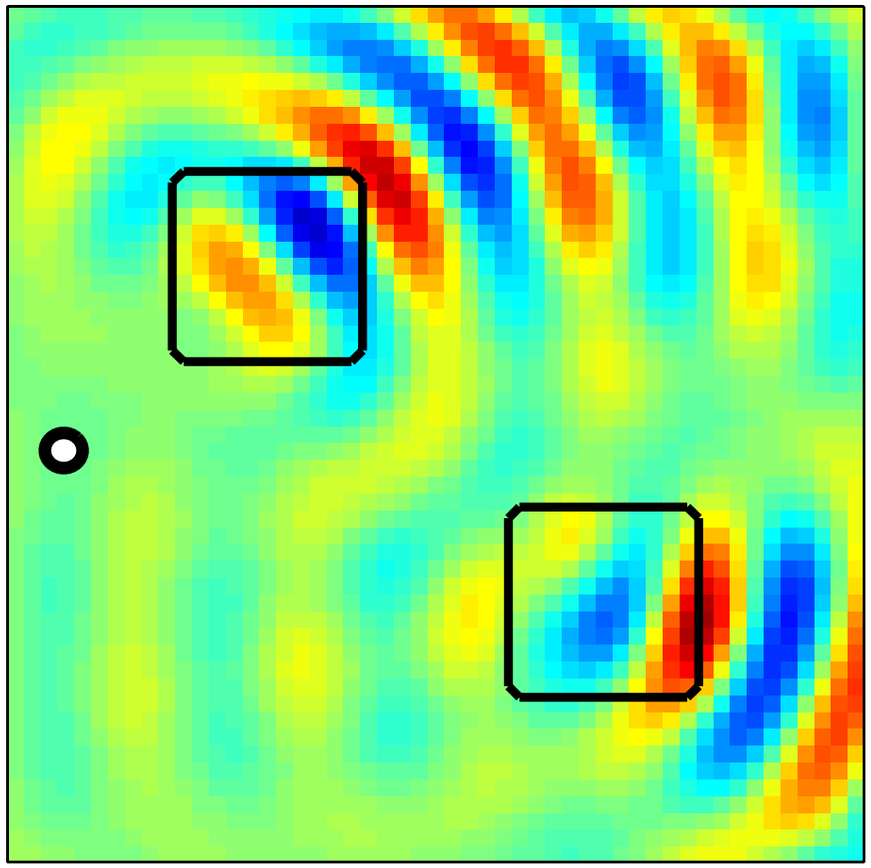
- Invert *monochromatic* data (10 Hz), transducers all around.
- Initial model is *constant*.
- 5-point discretization of Helmholtz operator:

$$A(\mathbf{m}) = \omega^2 \text{diag}(\mathbf{m}) + L$$

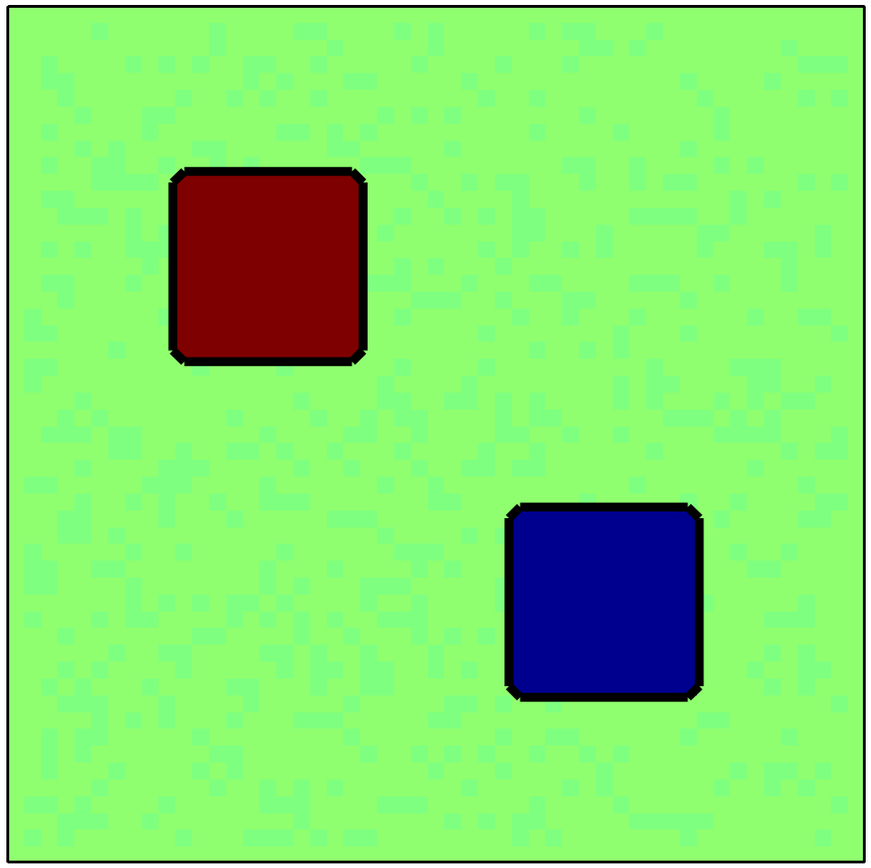
- *subproblems* are both *linear*!



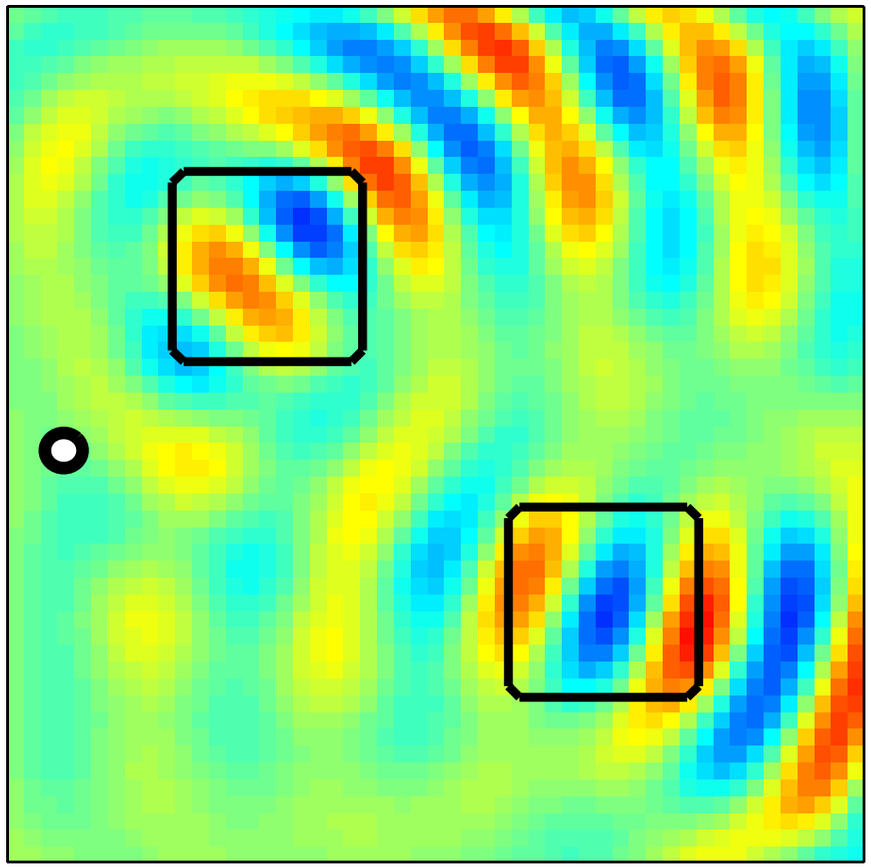
true wavefield

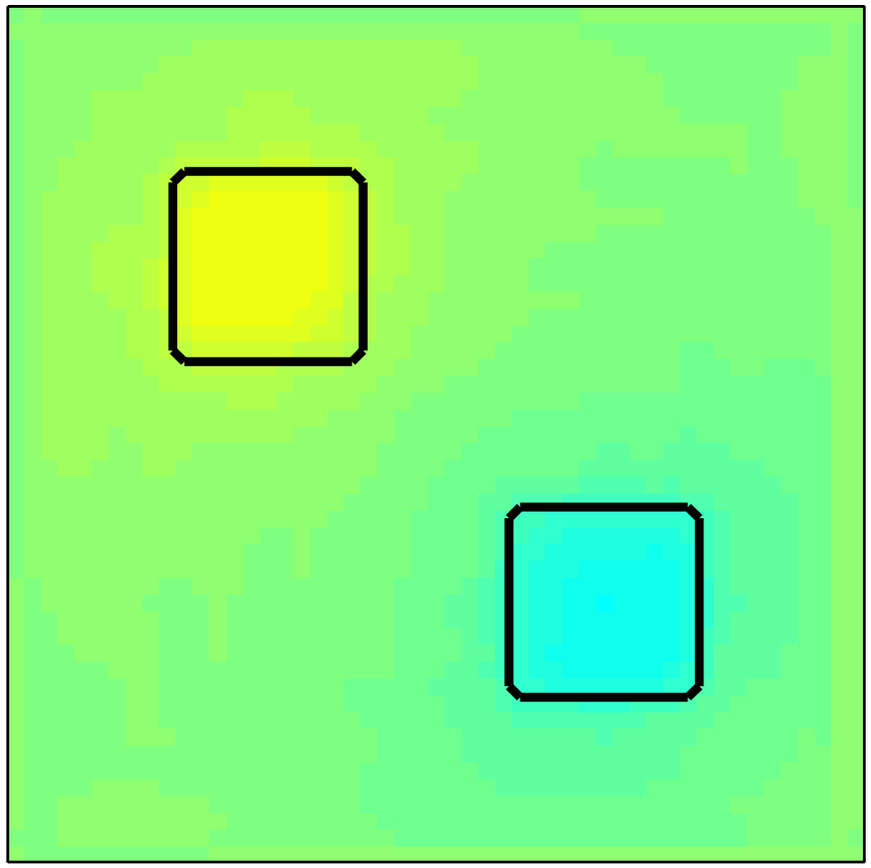


true model



iteration 1

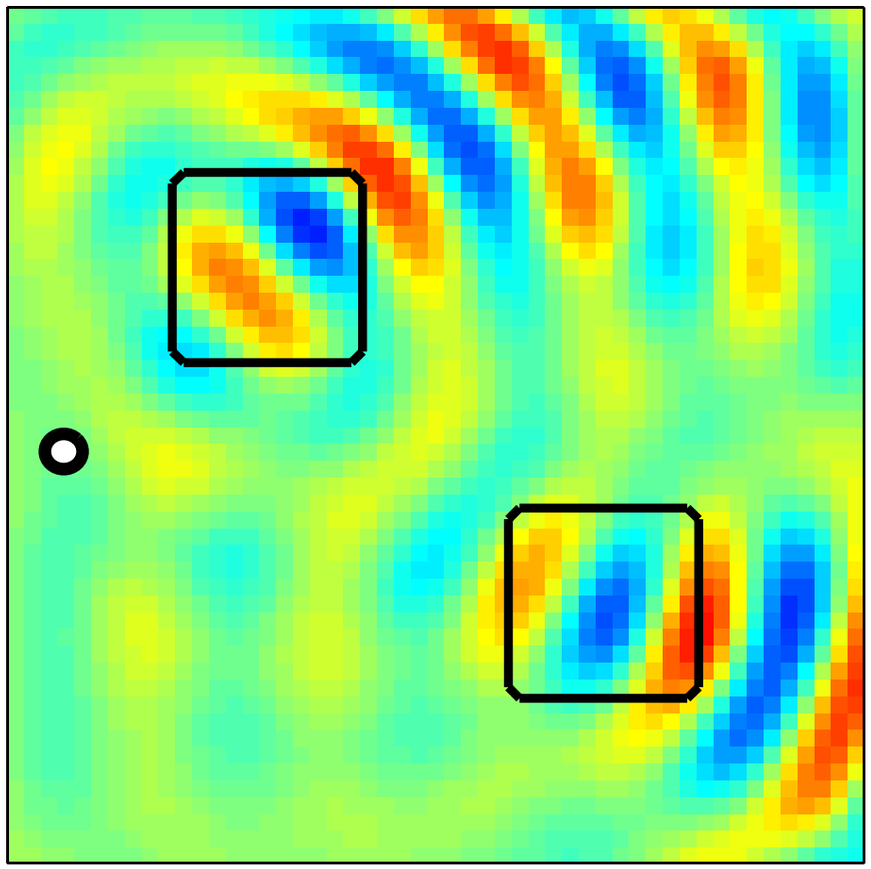


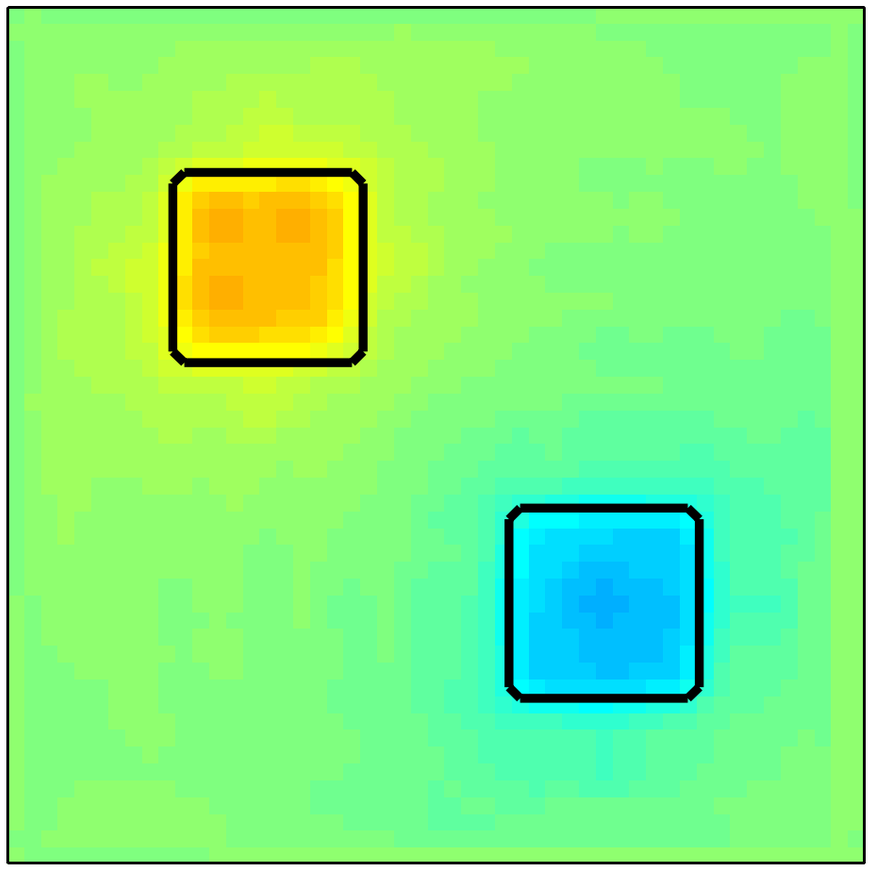


$$\mathbf{u}_0 = (P^*P + A_0^*A_0)^{-1}\,(P^*\mathbf{d} + A_0^*\mathbf{q})$$

$$\mathbf{m}_1 = \mathsf{diag}(|\mathbf{u}_0|^2)^{-1}\mathsf{diag}(\mathbf{u}_0)^*\,(\mathbf{q} - L\mathbf{u}_0)$$

iteration 2

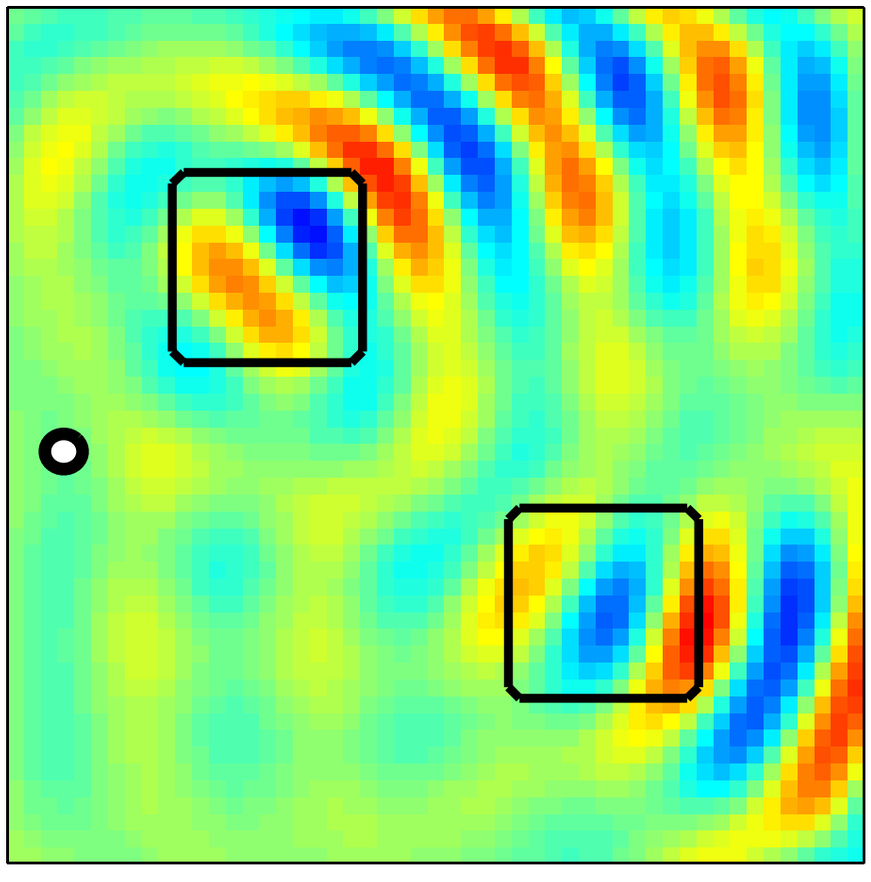


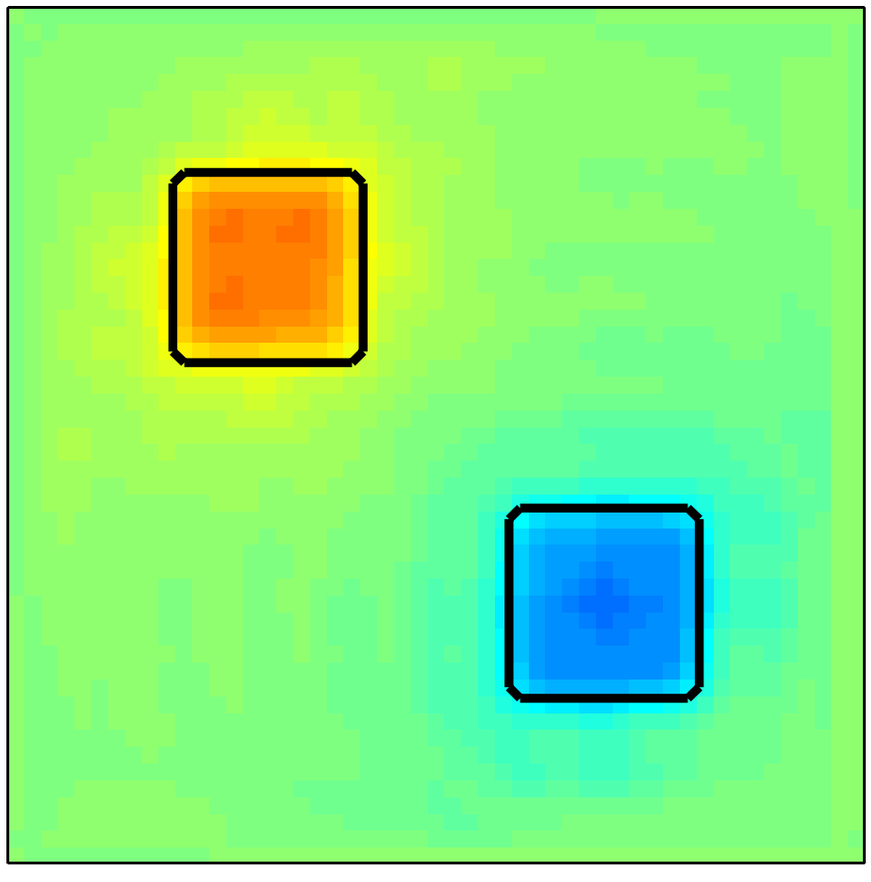


$$\mathbf{u}_1 = (P^*P + A_1^*A_1)^{-1}\,(P^*\mathbf{d} + A_1^*\mathbf{q})$$

$$\mathbf{m}_2 = \mathsf{diag}(|\mathbf{u}_1|^2)^{-1}\mathsf{diag}(\mathbf{u}_1)^*\,(\mathbf{q} - L\mathbf{u}_1)$$

iteration 3

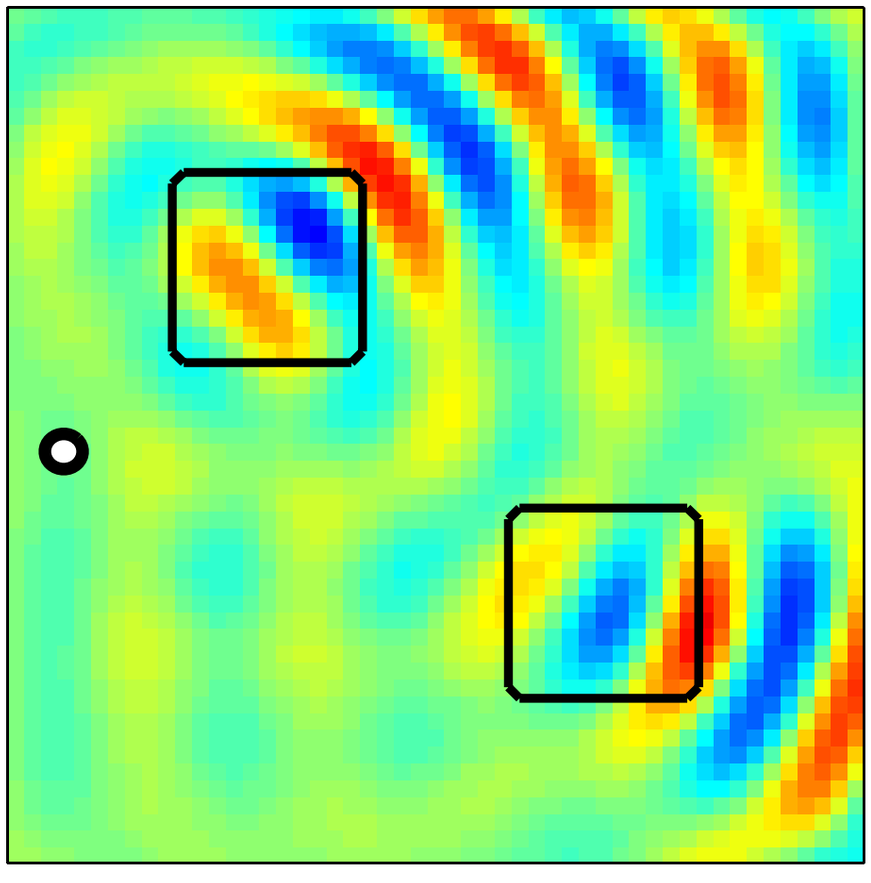


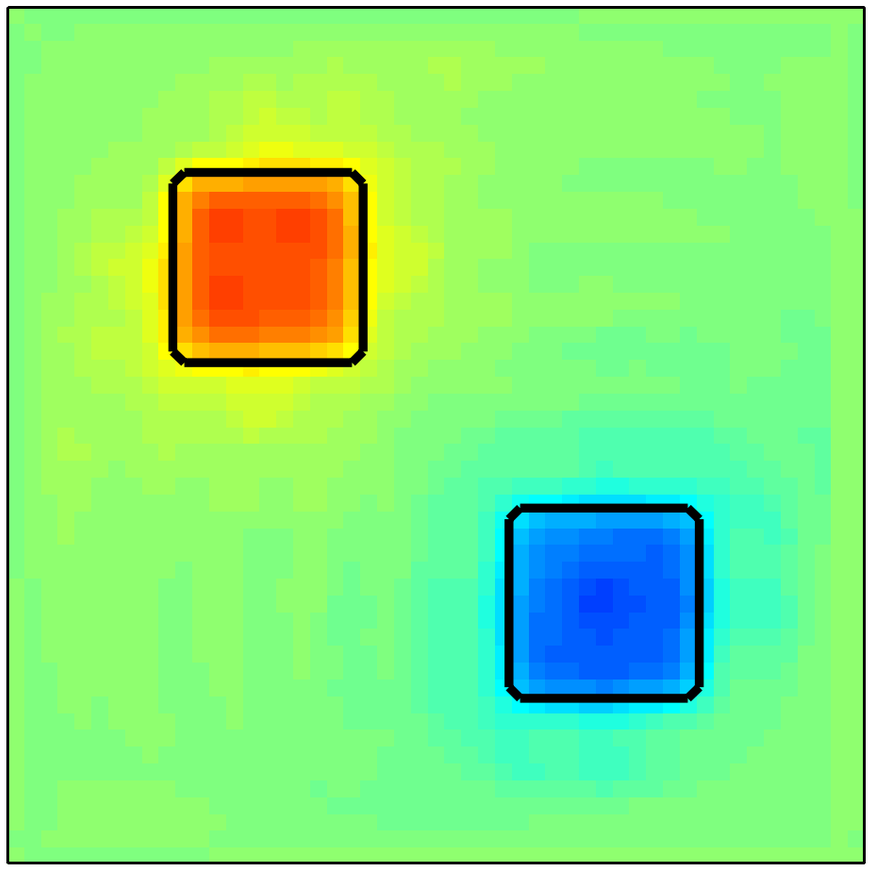


$$\mathbf{u}_2 = (P^*P + A_2^*A_2)^{-1}\,(P^*\mathbf{d} + A_2^*\mathbf{q})$$

$$\mathbf{m}_3 = \mathsf{diag}(|\mathbf{u}_2|^2)^{-1}\mathsf{diag}(\mathbf{u}_2)^*\,(\mathbf{q} - L\mathbf{u}_2)$$

iteration 4

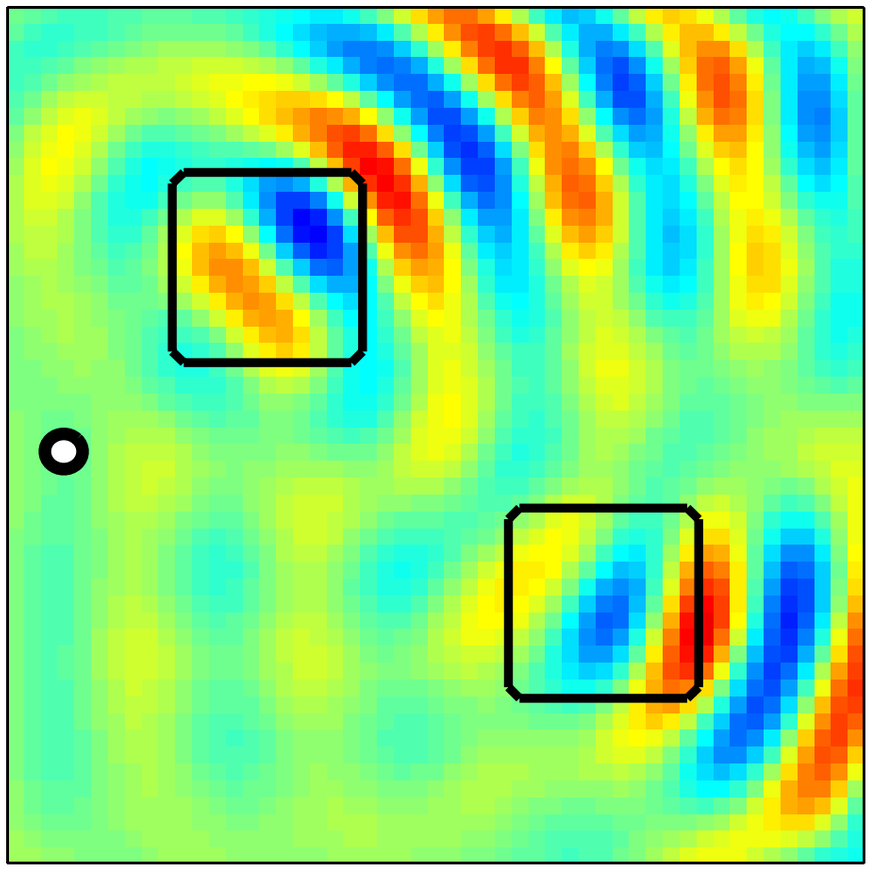


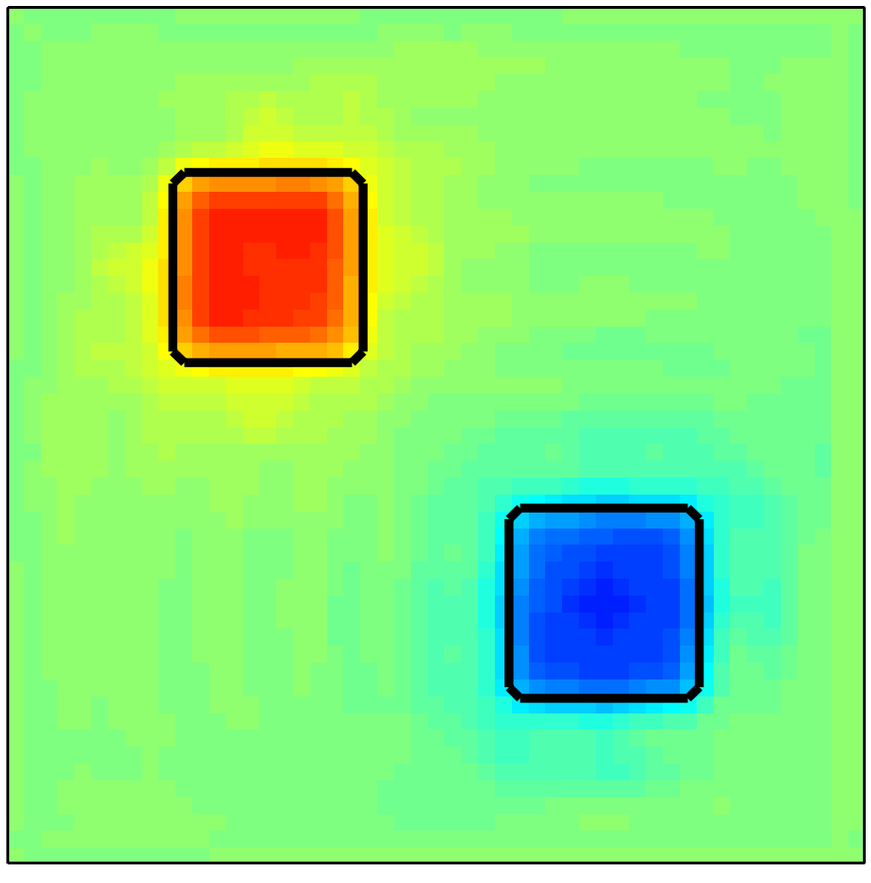


$$\mathbf{u}_3 = \left(P^*P + A_3^*A_3\right)^{-1}\left(P^*\mathbf{d} + A_3^*\mathbf{q}\right)$$

$$\mathbf{m}_4 = \mathsf{diag}(|\mathbf{u}_3|^2)^{-1}\mathsf{diag}(\mathbf{u}_3)^*\left(\mathbf{q} - L\mathbf{u}_3\right)$$

iteration 5

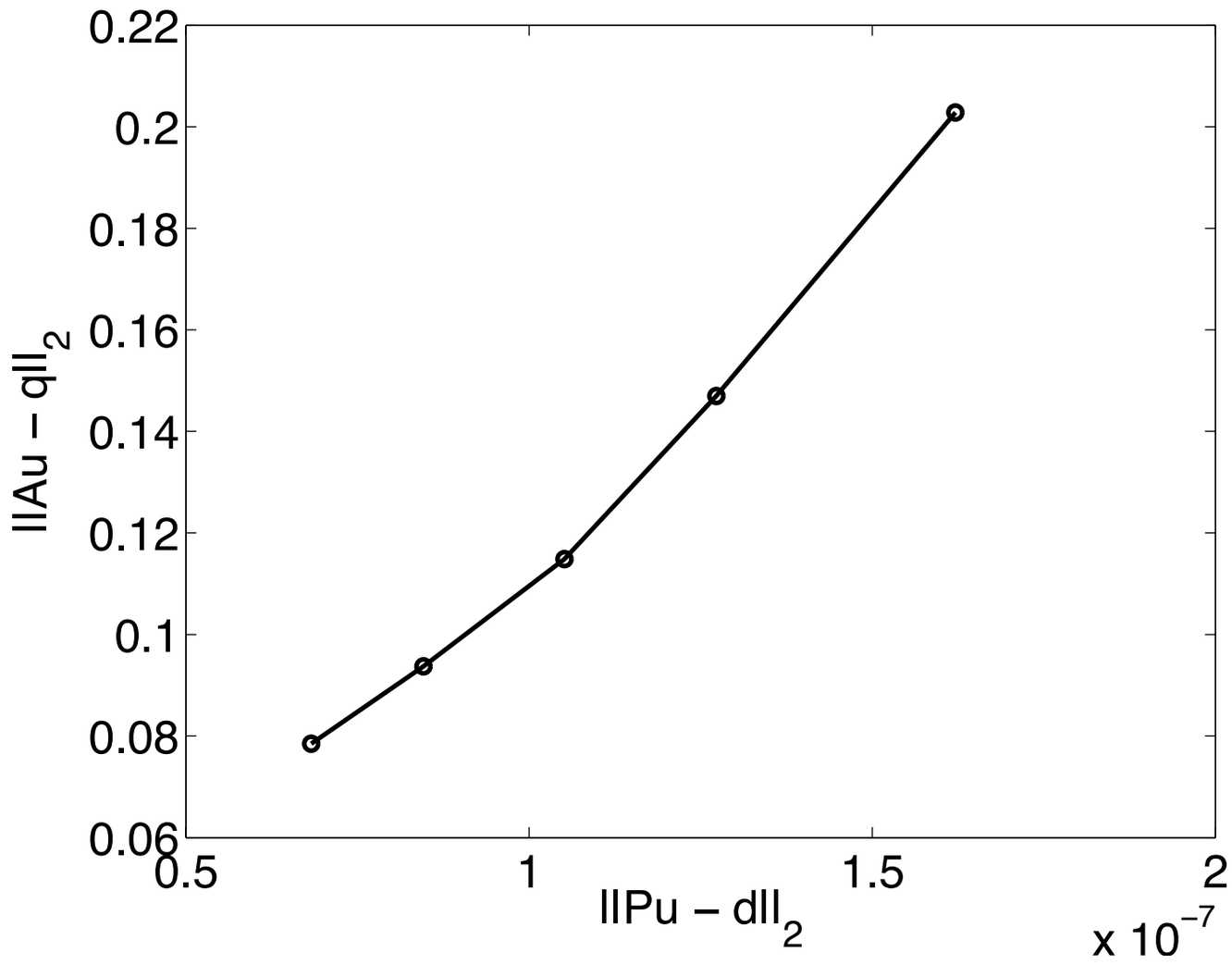




$$\mathbf{u}_4 = \left(P^*P + A_4^*A_4\right)^{-1}\left(P^*\mathbf{d} + A_4^*\mathbf{q}\right)$$

$$\mathbf{m}_5 = \mathsf{diag}(|\mathbf{u}_4|^2)^{-1}\mathsf{diag}(\mathbf{u}_4)^*\left(\mathbf{q} - L\mathbf{u}_4\right)$$

Convergence in terms of *data*-fit and *PDE*-fit:



Overview

- *Full* and *reduced*-space approach
- *Penalty* formulation
- Numerical examples
- Future work & Conclusions

PDE-*constrained* optimization

$$\min_{\mathbf{m},\mathbf{u}} ||P\mathbf{u} - \mathbf{d}||_2^2 \quad \text{s.t.} \quad A(\mathbf{m})\mathbf{u} = \mathbf{q}$$

- $\mathbf{u}$ is typically *much* bigger than $\mathbf{m}$
- Solving the PDE is computationally *expensive*

All-at-once approach

Cast into *Lagrangian* form:

$$\min_{\mathbf{m},\mathbf{u},\mathbf{v}} ||P\mathbf{u}-\mathbf{d}||_2^2 + \mathbf{v}^*(A(\mathbf{m})\mathbf{u}-\mathbf{q})$$

and *solve* by *iterating* on *sparse KKT* system:

$$\begin{pmatrix} * & * & G^* \\ * & P^*P+* & A \\ G & A & 0 \end{pmatrix} \begin{pmatrix} \delta\mathbf{m} \\ \delta\mathbf{u} \\ \delta\mathbf{v} \end{pmatrix} = - \begin{pmatrix} G^*\mathbf{v} \\ A^*\mathbf{v}+P^*(P\mathbf{u}-\mathbf{d}) \\ A\mathbf{u}-\mathbf{q} \end{pmatrix}$$

with A, P, G block diagonal matrices and $G_i(\mathbf{m},\mathbf{u}) = \frac{\partial A_i(\mathbf{m})\mathbf{u}_i}{\partial \mathbf{m}}$

*** higher order terms.**

All-at-once approach

Pros:

- *no* explicit *PDE* solves
- iterations w/ *sparse* (KKT) system
- *expands* the *search* space (avoids *local* minima)

Cons:

- introduces *extra* variable
- requires access to *forward* & *adjoint* wavefields for *all* sources... That's a show *stopper*...

Reduced approach

For *each* experiment, eliminate the PDE *constraint*–*i.e.*,

$$\min_{\mathbf{m}} \phi_{\text{red}}(\mathbf{m}) = \sum_{i=1}^{M} \frac{1}{2} \| P_i A_i(\mathbf{m})^{-1} \mathbf{q}_i - \mathbf{d}_i \|_2^2$$

Pros:

- *no* need to have access to *all* wavefields
- *accumulate* gradients & Hessians

Cons:

- need to solve PDEs *explicitly*
- *dense* Jacobian and (Gauss-Newton) Hessian
- smaller *search* space, *highly* nonlinear in **m**

Reduced approach
gradient & Hessian

Accumulate gradients:

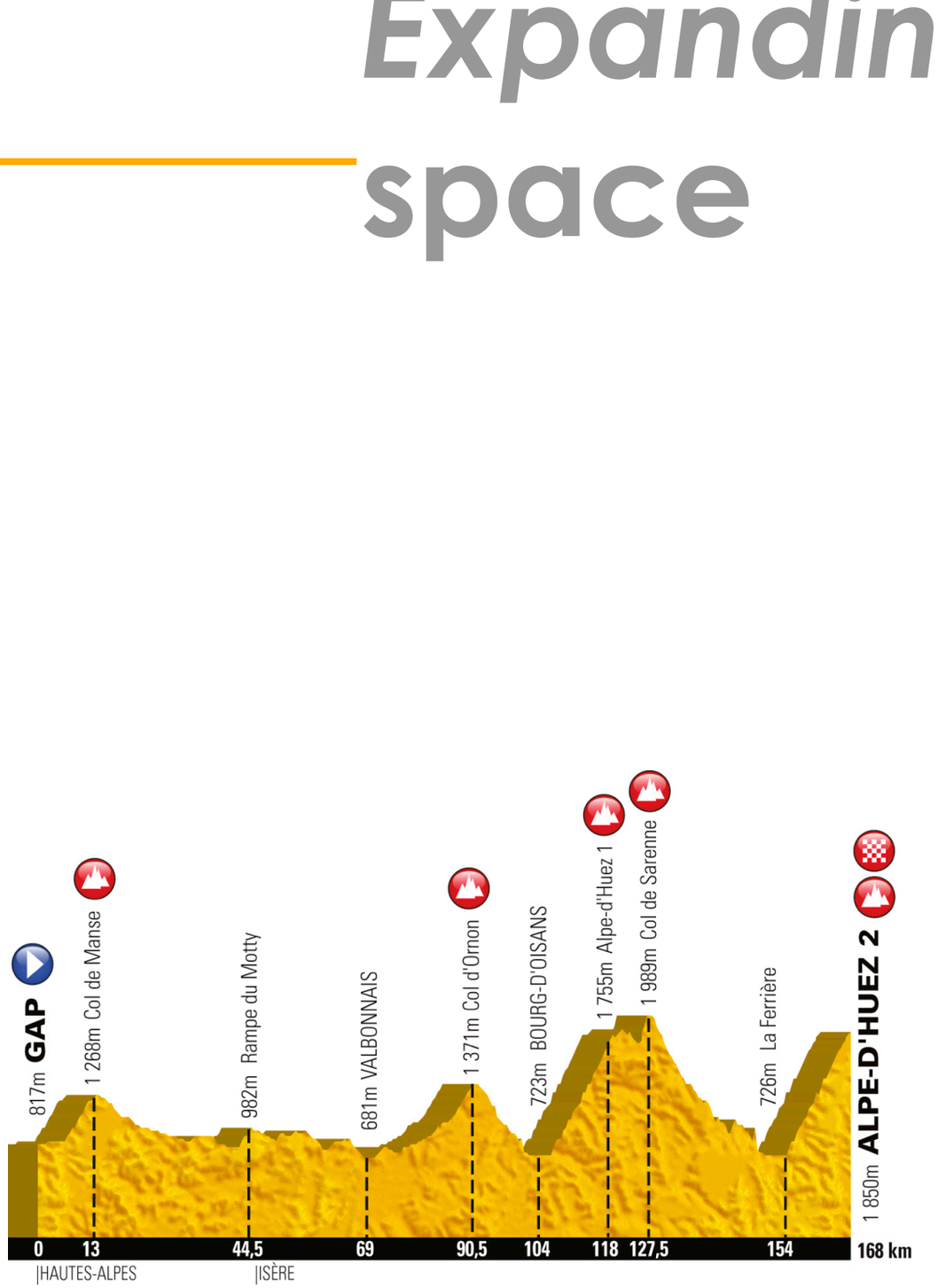
$$\nabla\phi_{\text{red}}(\mathbf{m}) = \sum_{i=1}^{M} G_i(\mathbf{m},\mathbf{u}_i)^*\mathbf{v}_i \quad \text{and} \quad G_i(\mathbf{m},\mathbf{u}) = \frac{\partial A(\mathbf{m})\mathbf{u}_i}{\partial \mathbf{m}},$$

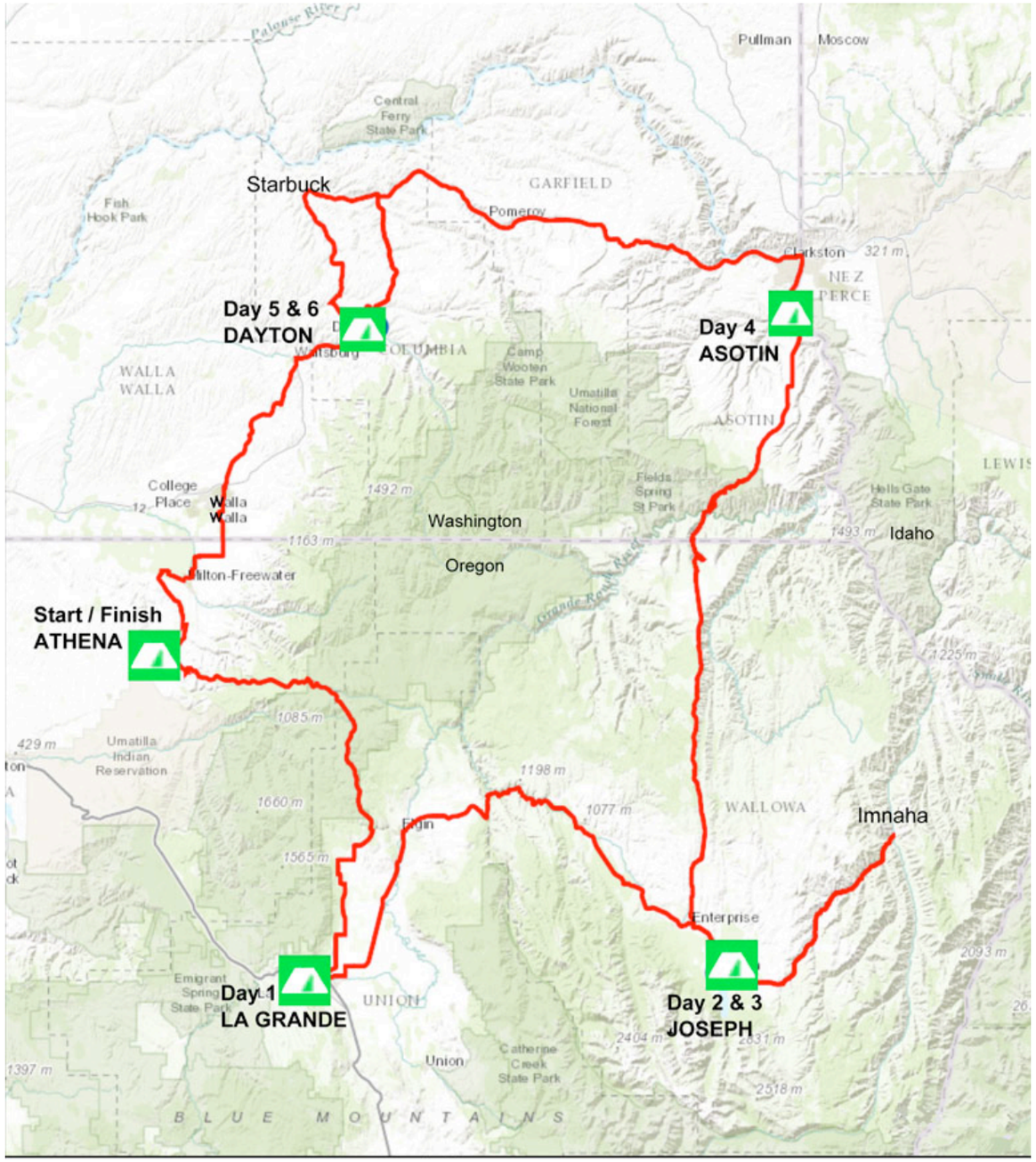
$$\begin{aligned} A_i(\mathbf{m})\mathbf{u}_i &= \mathbf{q}_i \\ A_i(\mathbf{m})^*\mathbf{v} &= -P_i^*(P_i\mathbf{u}_i - \mathbf{d}_i). \end{aligned}$$

Accumulate Gauss-Newton Hessians:

$$H_{\text{red}}^{GN} = \nabla^2\phi_{\text{red}}(\mathbf{m}) = \sum_{i=1}^{M} G_i^* A_i^{-*} P_i^T P_i A_i^{-1} G_i$$

Expanding search space





Challenge

Can we find an *alternative* approach that has *best* of *both* worlds?

- *expanded* search space as in *all-at-once* method w/o needing access to *all* wavefields for each *iteration*
- *sparse* matrices
- *mundane* nonlinearity
- *accumulation* of *gradient* & GN *Hessians*

Probabilistic argument

In Tarantola's formulation *least-squares* misfit ϕ_{reduced}

- allows for *errors* in the data *residual*
- considers *physics* in WE as *infallible*

Instead, we consider both *physics* & observed *data* as *fallible*...

- *least-squares* penalties for *model* & *data* errors

Penalty approach

Penalty formulation:

$$\min_{\mathbf{m},\mathbf{u}} \phi_\lambda(\mathbf{m},\mathbf{u}) = \sum_{i=1}^{M} \frac{1}{2}\|P_i\mathbf{u}_i - \mathbf{d}_i\|_2^2 + \frac{\lambda^2}{2}\|A_i(\mathbf{m})\mathbf{u}_i - \mathbf{q}_i\|_2^2$$

solution coincides with that of *original* problem as $\lambda \uparrow \infty$

Penalty approach

For *each* experiment, eliminate wavefield ($\nabla_{\mathbf{u}}\phi_\lambda(\mathbf{m},\mathbf{u})=0$) by solving data-augmented PDE:

$$\min_{\mathbf{u}_i}\left\|\begin{pmatrix} P_i \\ \lambda A_i(\mathbf{m}) \end{pmatrix}\mathbf{u}_i - \begin{pmatrix} \mathbf{d}_i \\ \lambda \mathbf{q}_i \end{pmatrix}\right\|_2^2$$

Define *new* objective $\phi_{\text{pen}}(\mathbf{m}) = \phi_\lambda(\mathbf{m},\overline{\mathbf{u}}(\mathbf{m}))$

Derivatives given by

$$\nabla\phi_{\text{pen}} = \nabla_{\mathbf{m}}\phi_\lambda$$

$$\nabla^2\phi_{\text{pen}} = \nabla^2_{\mathbf{m}}\phi_\lambda - \nabla^2_{\mathbf{m},\mathbf{u}}\phi_\lambda\left(\nabla^2_{\mathbf{u}}\phi_\lambda\right)^{-1}\nabla^2_{\mathbf{u},\mathbf{m}}\phi_\lambda$$

(all evaluated at the optimal $\overline{\mathbf{u}}_i$).

Penalty formulation

A *local* minimizer of $\phi_{\text{pen}}(\mathbf{m})$, together with the corresponding $\overline{\mathbf{u}}_i$ are also *local* minimizers of $\phi_\lambda(\mathbf{m}, \overline{\mathbf{u}}_i)$ and vice versa.

Penalty formulation

Accumulate gradients:

$$\nabla_{\mathbf{m}}\phi_{\text{pen}}(\mathbf{m},\overline{\mathbf{u}}) = \sum_{i=1}^{M} G_i(\mathbf{m},\overline{\mathbf{u}}_i)^*\left(A_i(\mathbf{m})\overline{\mathbf{u}}_i - \mathbf{q}_i\right)$$

Accumulate (approximate) Gauss-Newton Hessians:

$$\phi_{\text{pen}}(\mathbf{m}) = \sum_{i=1}^{M}(\lambda^2 - 1)G_i^* G_i.$$

Penalty formulation

Algorithm 1 FWI algorithm based on the penalty formulation

for $t = 0$ to N **do**
  $\mathbf{g}_t = 0$
  $H_t = 0$
  for $k = 1$ to N_f **do**
    for $l = 1$ to N_s **do**
      solve $\begin{pmatrix} \lambda_t A_k(\mathbf{m}_t) \\ P \end{pmatrix} \mathbf{u}_{kl} = \begin{pmatrix} \lambda_t \mathbf{q}_{kl} \\ \mathbf{d}_{kl} \end{pmatrix}$
      $\mathbf{g}_t = \mathbf{g}_t + \lambda_t^2 G_k(\mathbf{m}_t, \mathbf{u}_{kl})^* (A_k(\mathbf{m}_t)\mathbf{u}_{kl} - \mathbf{q}_{kl})$
      $H_t = H_t + (\lambda_t^2 - 1) G_k(\mathbf{m}_t, \mathbf{u}_{kl})^* G_k(\mathbf{m}_t, \mathbf{u}_{kl})$
    end for
  end for
  solve $H_t \Delta\mathbf{m}_t = -\mathbf{g}_t$
  update $\mathbf{m}_{t+1} = \mathbf{m}_t + \Delta\mathbf{m}_t$
  update λ_t.
end for

Penalty formulation

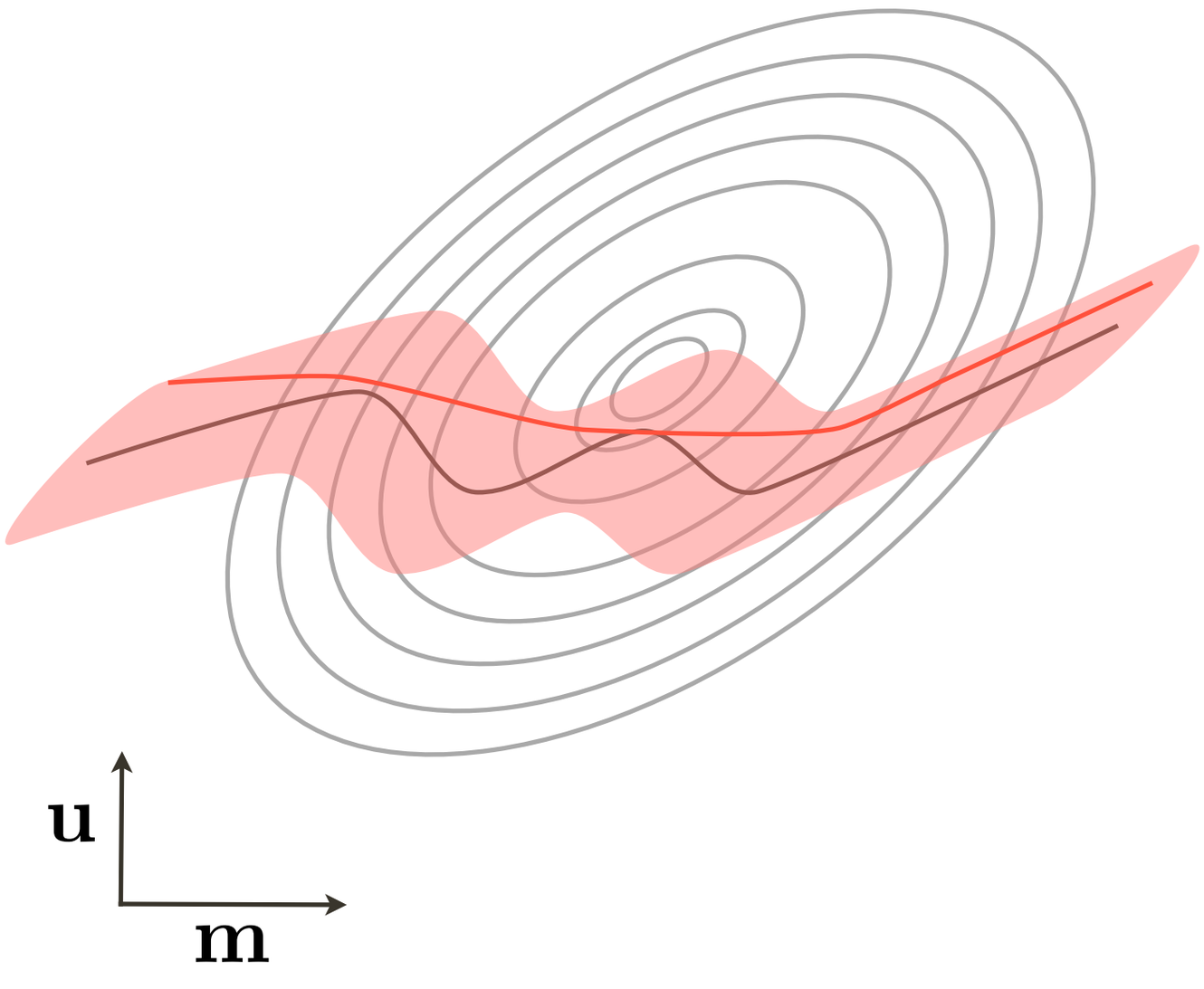
Pros:

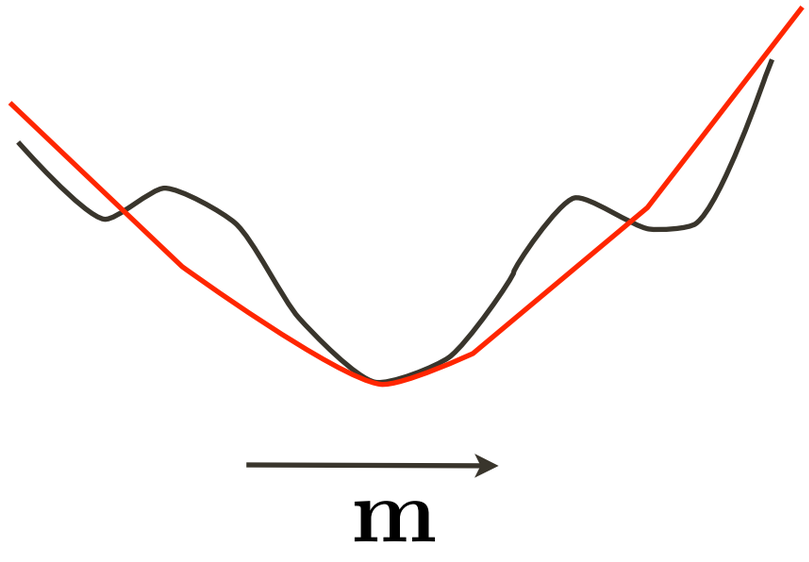
- *expansion* of *search* space
- *no* need to have access to *all* wavefields for *each* iteration
- *no* adjoint *wavefields*
- *sparse* approximation of *Hessian* at small λ
- less *non*-linear in **m**

Cons:

- solution of overdetermined *data*-augmented system
- "cooling" strategy for λ

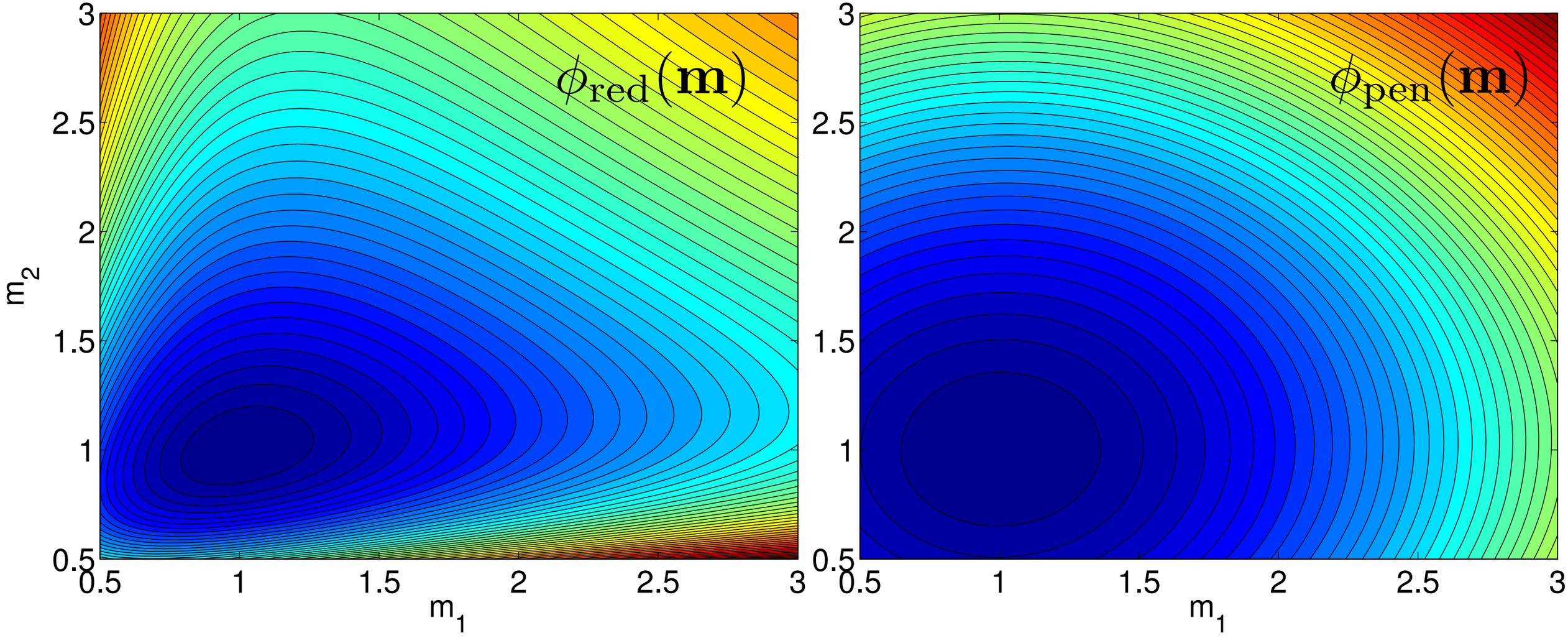
Penalty vs. reduced



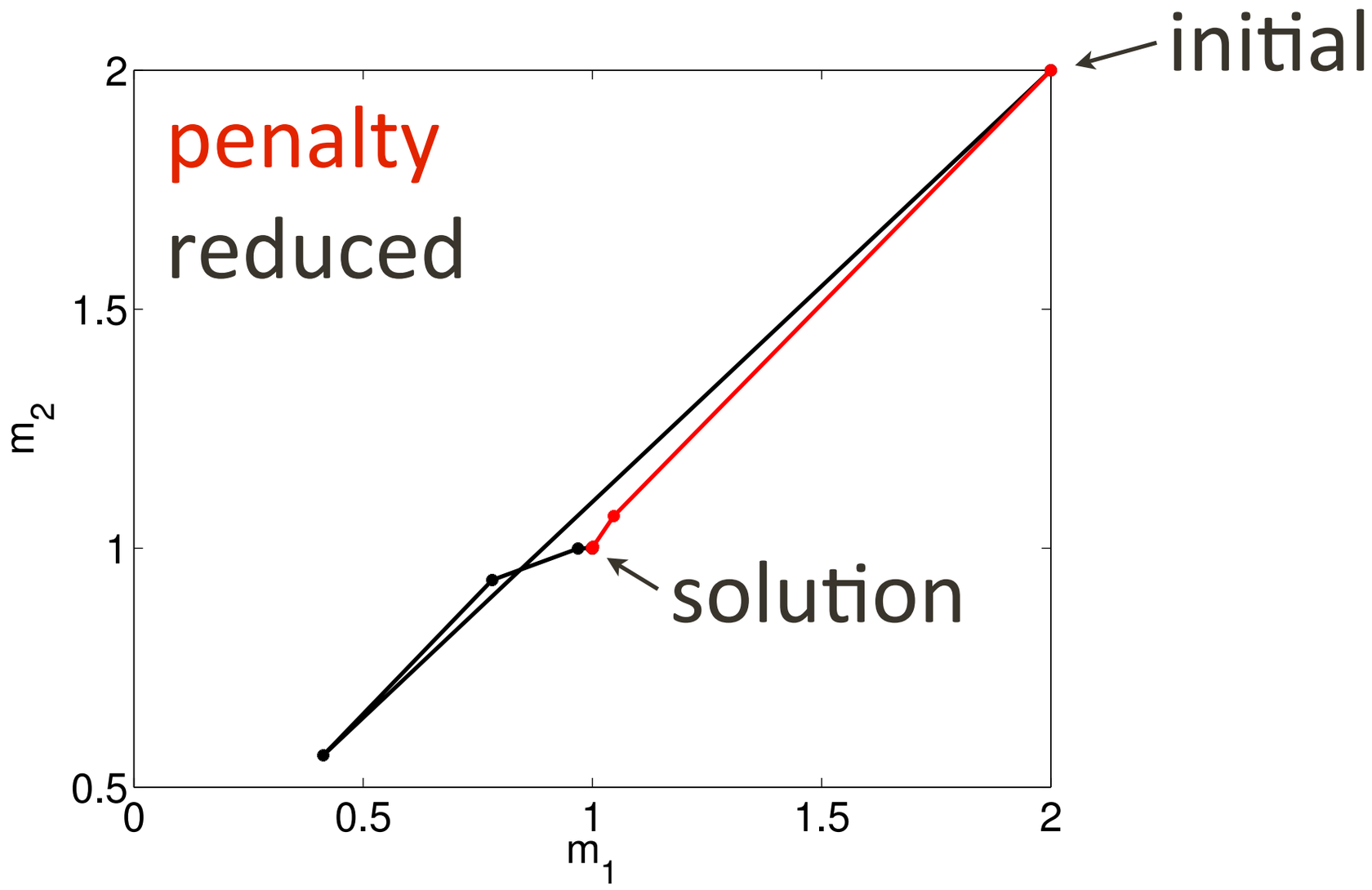


Penalty vs. reduced

$$A(\mathbf{m}) = \begin{pmatrix} m_1 + 0.5 & 0.25 \\ 0.25 & m_2 + 1 \end{pmatrix} \quad P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \mathbf{q} = \begin{pmatrix} 1.75 \\ 2.25 \end{pmatrix}$$



Penalty vs. Reduced



Penalty approach

	# PDE's	Storage	Gauss-Newton update
penalty	M	N	solve sparse SPSD system in N unknowns
reduced	$2M$	$2N$	solve matrix-free linear system in N unknowns, requires $3M$ PDE-solves per mat-vec
all-at-once	0	$(2M+1)N$	solve sparse symmetric, possibly indefinite system in $(2M+1)N$ unknowns

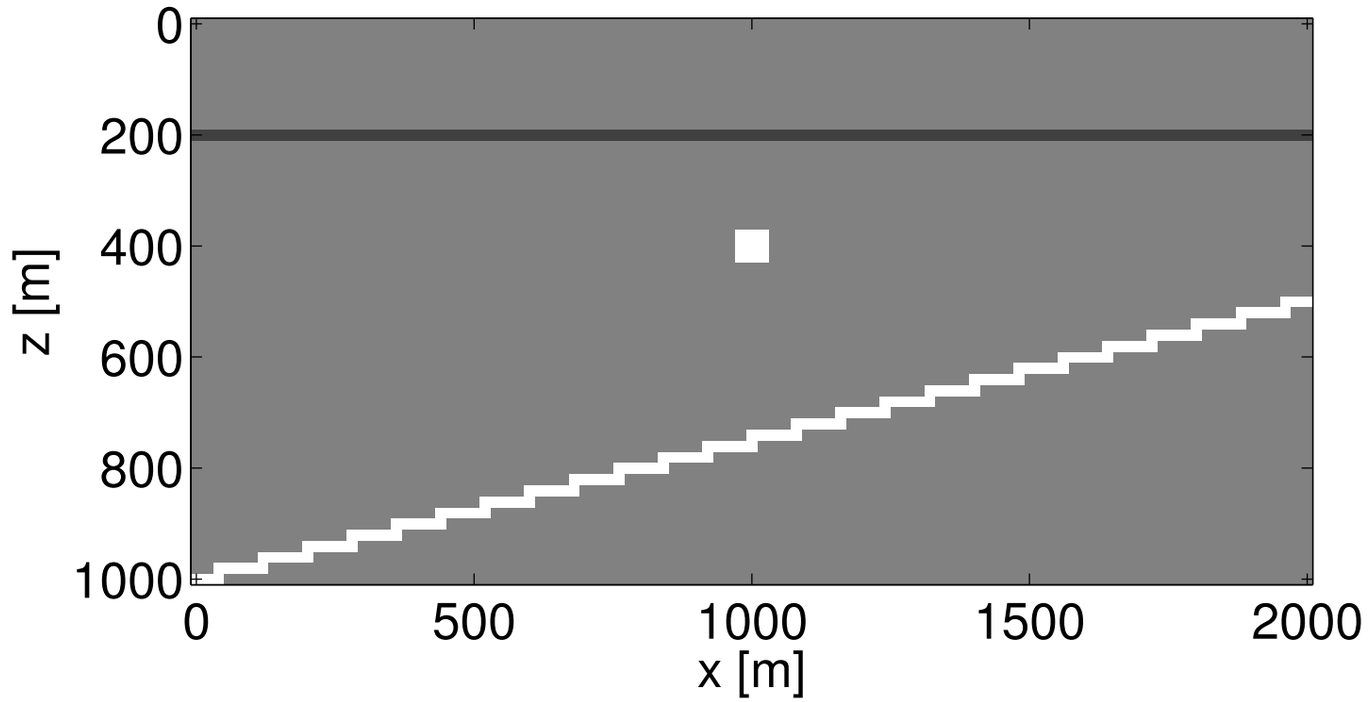
Table 1: Leading order computation and storage costs per iteration of different methods; M denotes the number of experiments and N denotes the number of gridpoints. For large-scale 3D seismic inverse problems we typically have $M = \mathcal{O}(10^4) - \mathcal{O}(10^7)$ and $N = \mathcal{O}(10^6) - \mathcal{O}(10^9)$

Numerical examples

- Imaging
- Local minima
- Full-waveform inversion
- DC resistivity
- Optical tomography

Imaging

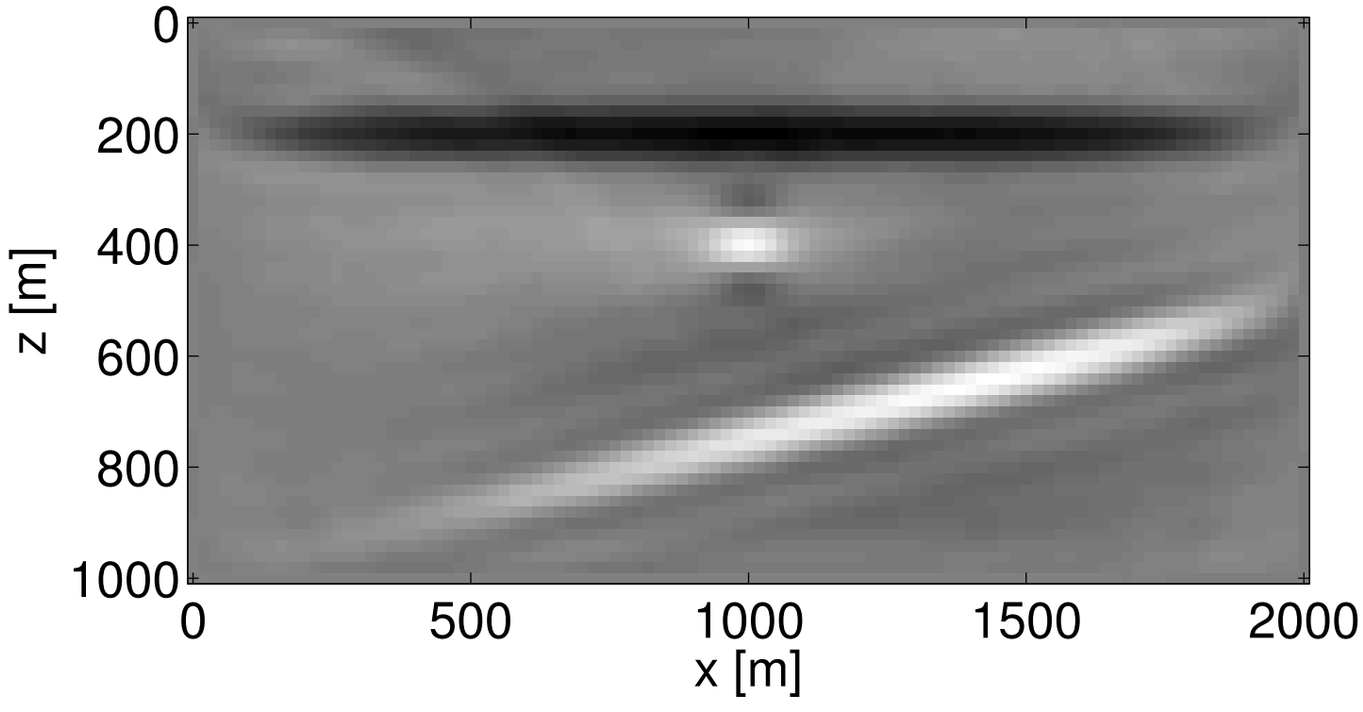
The *gradient* of the reduced objective yields an *image* of the subsurface..



Imaging

Conventional reverse-time migration:

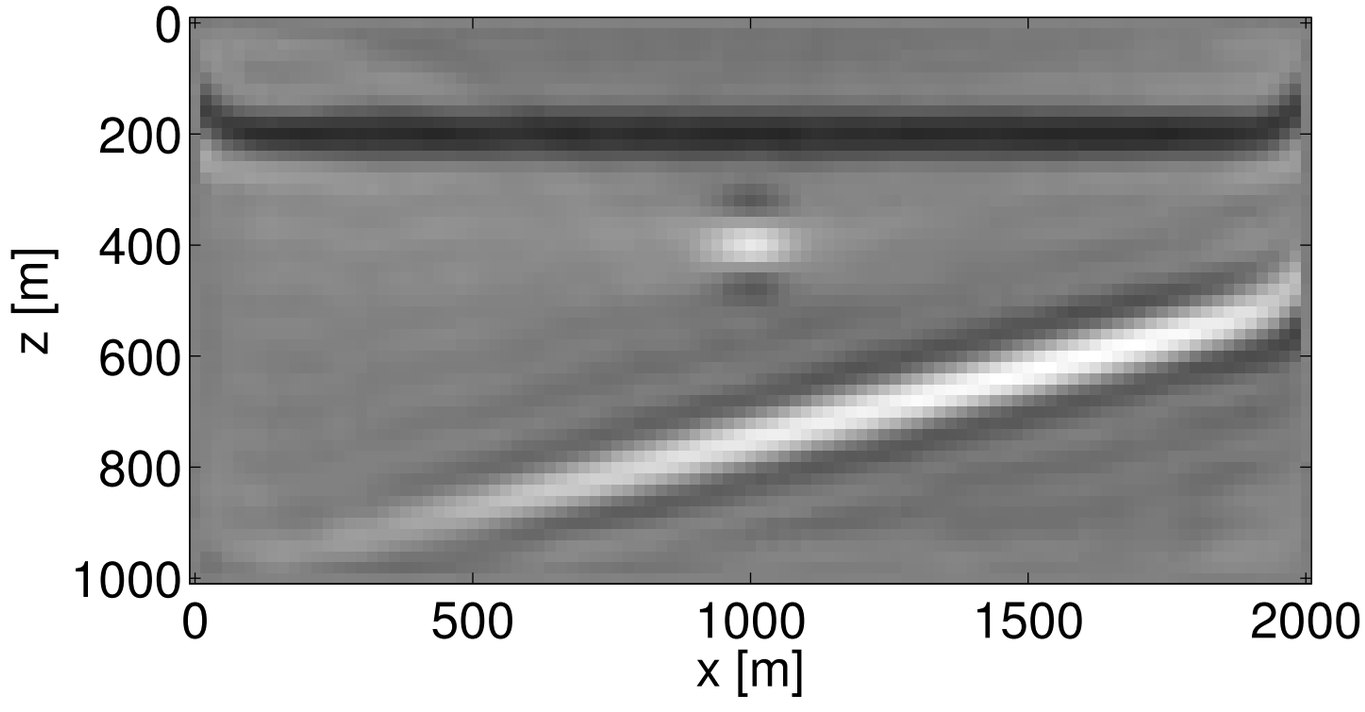
1. solve forward wave-equation
2. solve adjoint wave-equation
3. apply `imaging condition'



Imaging

Penalty-method reverse-time migration:

1. solve overdetermined wave-equation
2. go for lunch
3. apply `imaging condition'



Imaging

Reverse-time migration w/ *adjoint* state:

$$\delta\mathbf{m}_{\text{red}} = \sum_{i=1}^{M} \omega_i^2 \text{diag}(\mathbf{u}_i)^* \mathbf{v}_i \quad \text{with} \quad \mathbf{v}_i := \overbrace{A_i(\mathbf{m})^{-*}\left(P_i^*(\mathbf{d}_i - P_i\mathbf{u}_i)\right)}^{\text{back prop. data residue}}$$

"*Forward*"-time migration w/ *penalty* formulation:

$$\delta\mathbf{m}_{\text{pen}} = \lambda^2 \sum_{i=1}^{M} \omega_i^2 \text{diag}(\mathbf{u}_i)^* \delta\mathbf{u}_i \quad \text{with} \quad \delta\mathbf{u}_i := \overbrace{\left(A_i(\mathbf{m})\mathbf{u}_i - \mathbf{q}_i\right)}^{\text{PDE residue}}$$

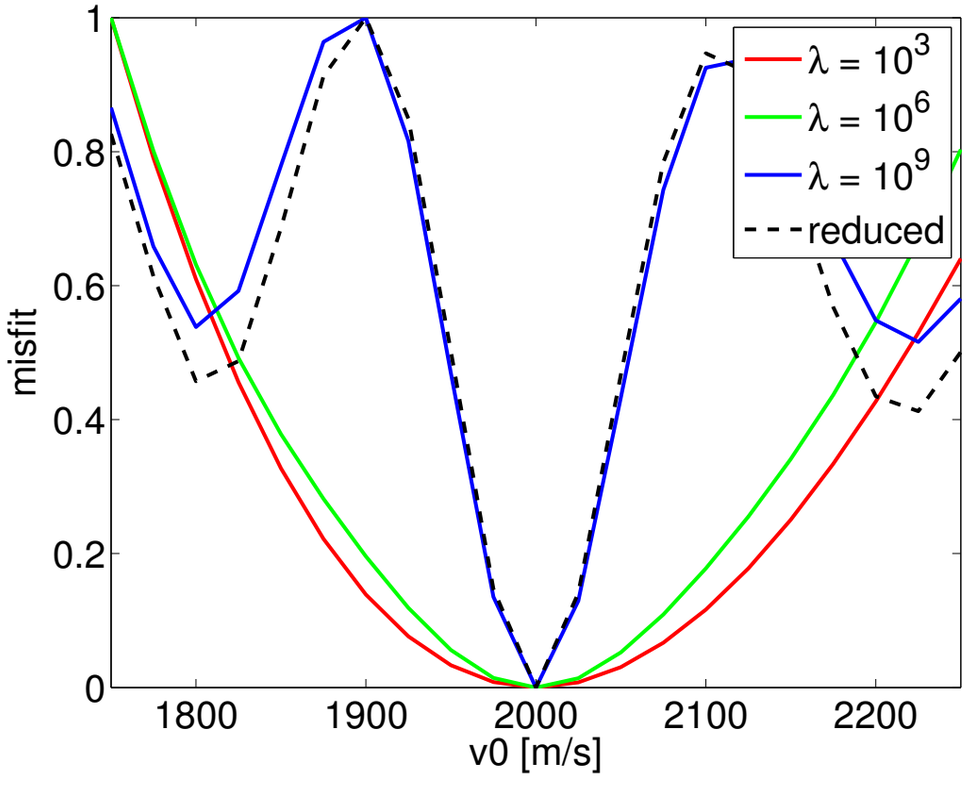
with *trivial* GN Hessian

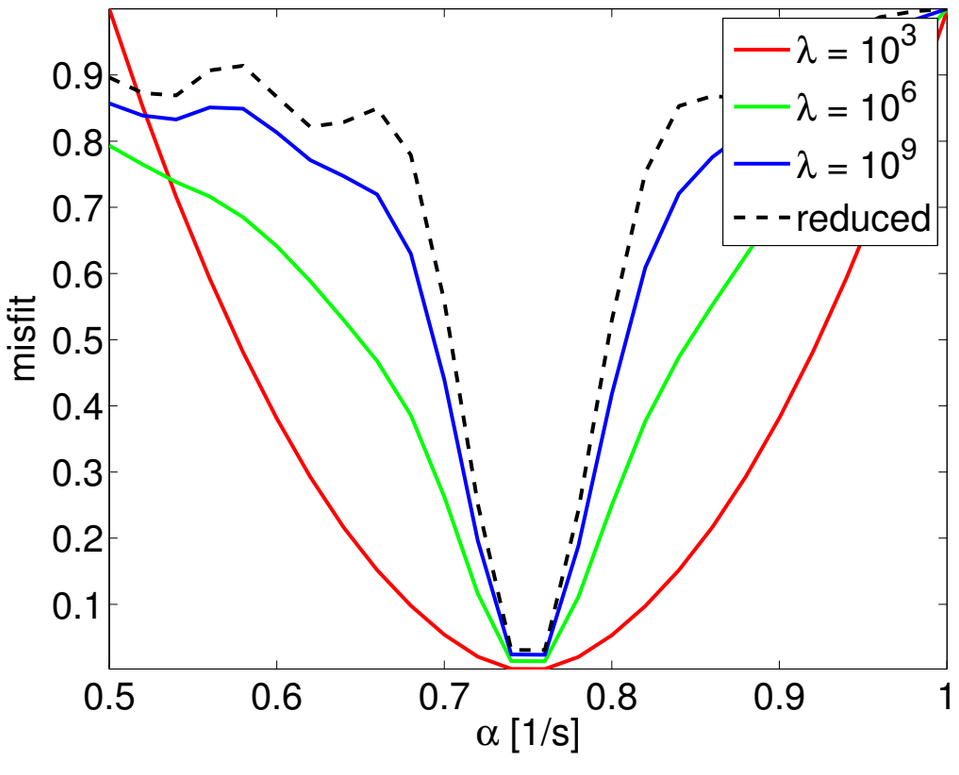
$$H_{\text{pen}} = (\lambda^2 - 1) \sum_{i=1}^{M} \omega_i^4 \text{diag}(\mathbf{u}_i)^* \text{diag}(\mathbf{u}_i).$$

Local minima

single shot, single frequency data for *linear* velocity profile $v(z) = v_0 + \alpha z$,

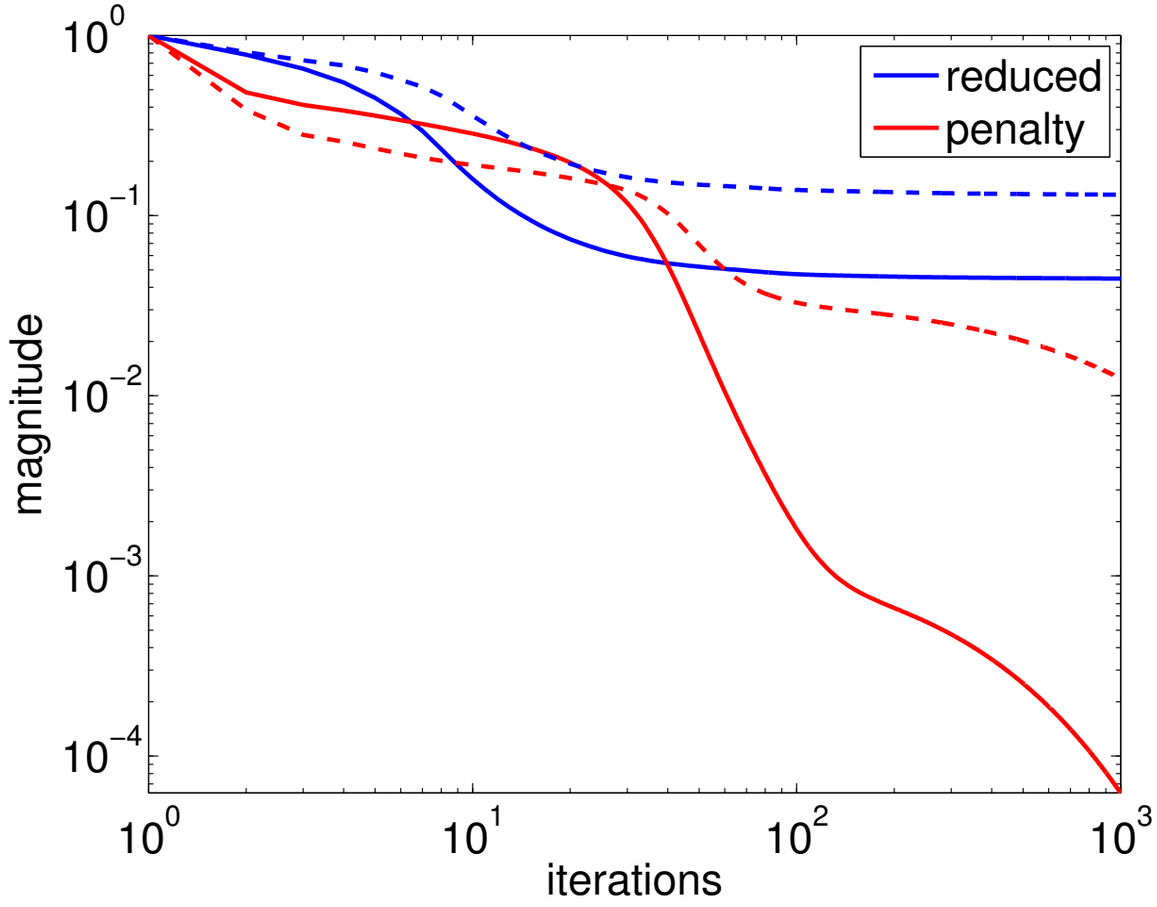
misfit as function of (v_0, α):





Local minima

Invert *single* shot, *single* frequency data for 1D velocity profile using *steepest* descent.

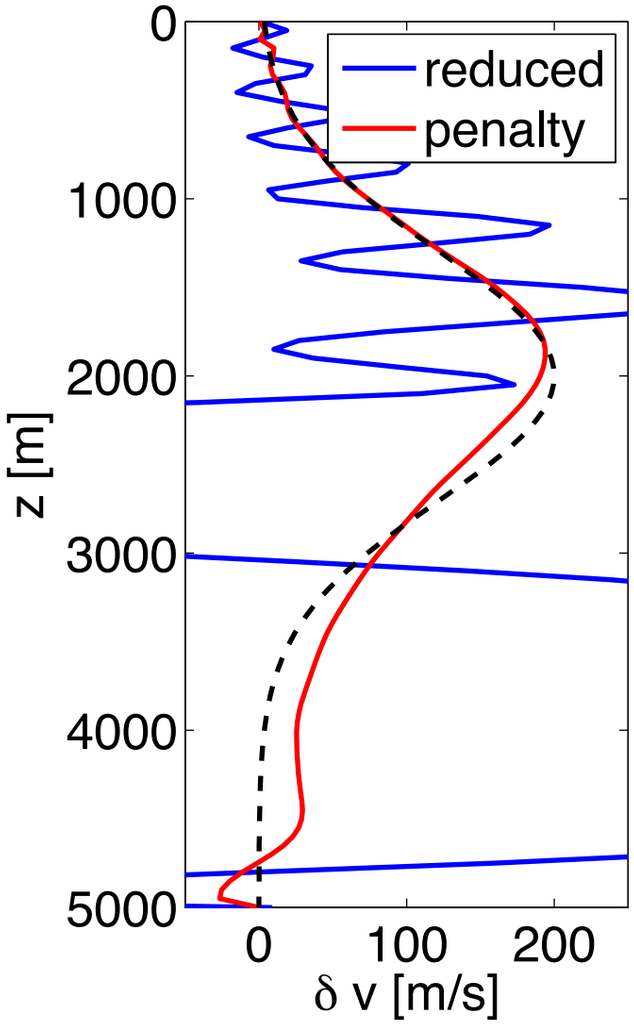


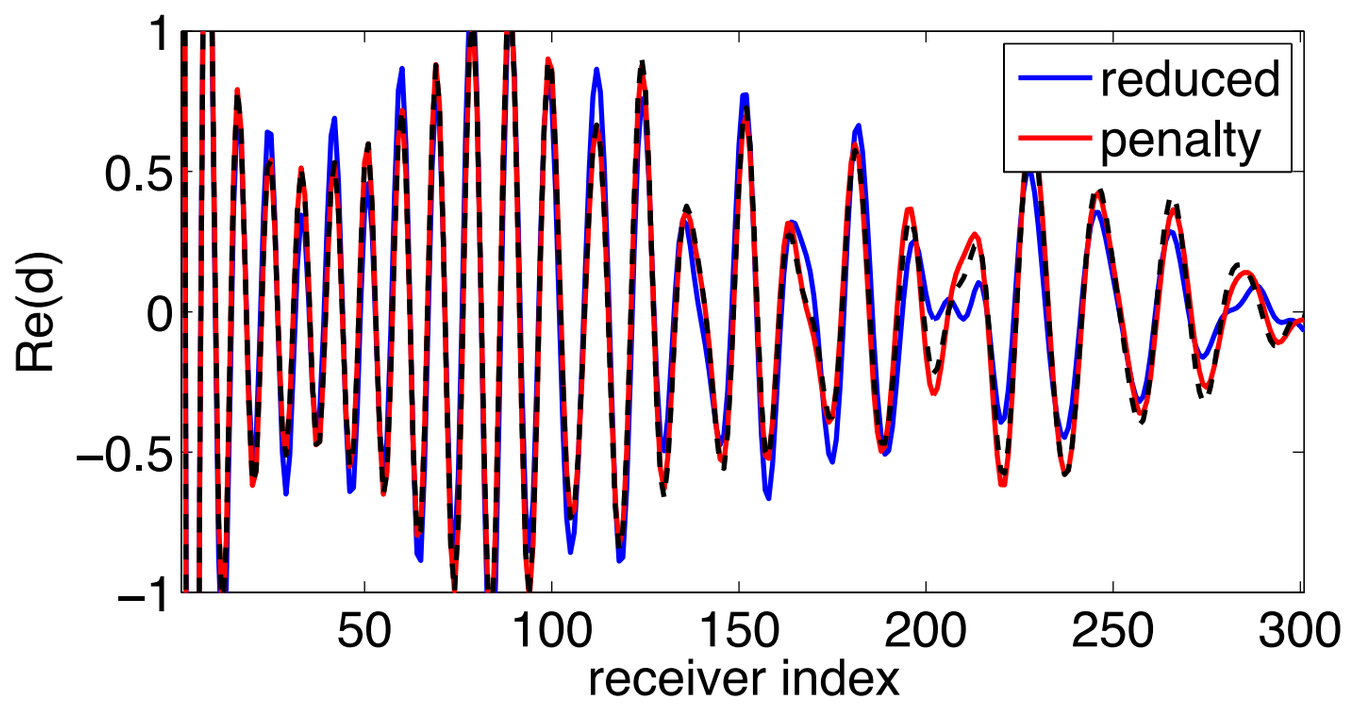
- misfit

-- gradient

Local Minima

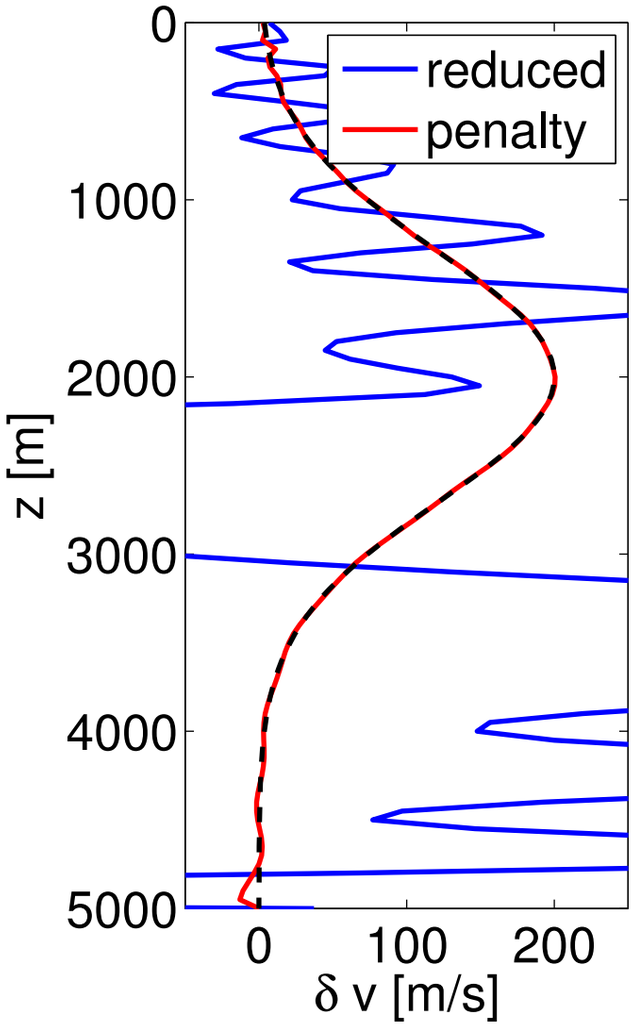
After *100* iterations

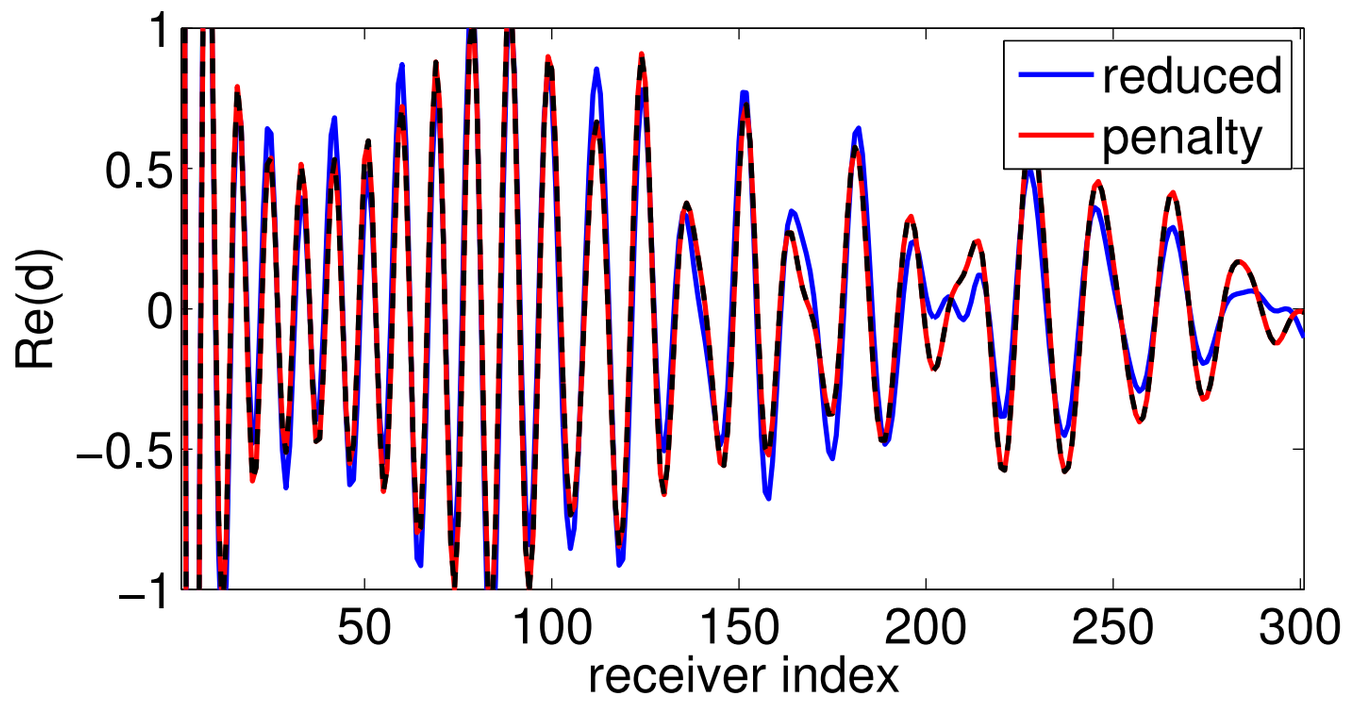




Local minima

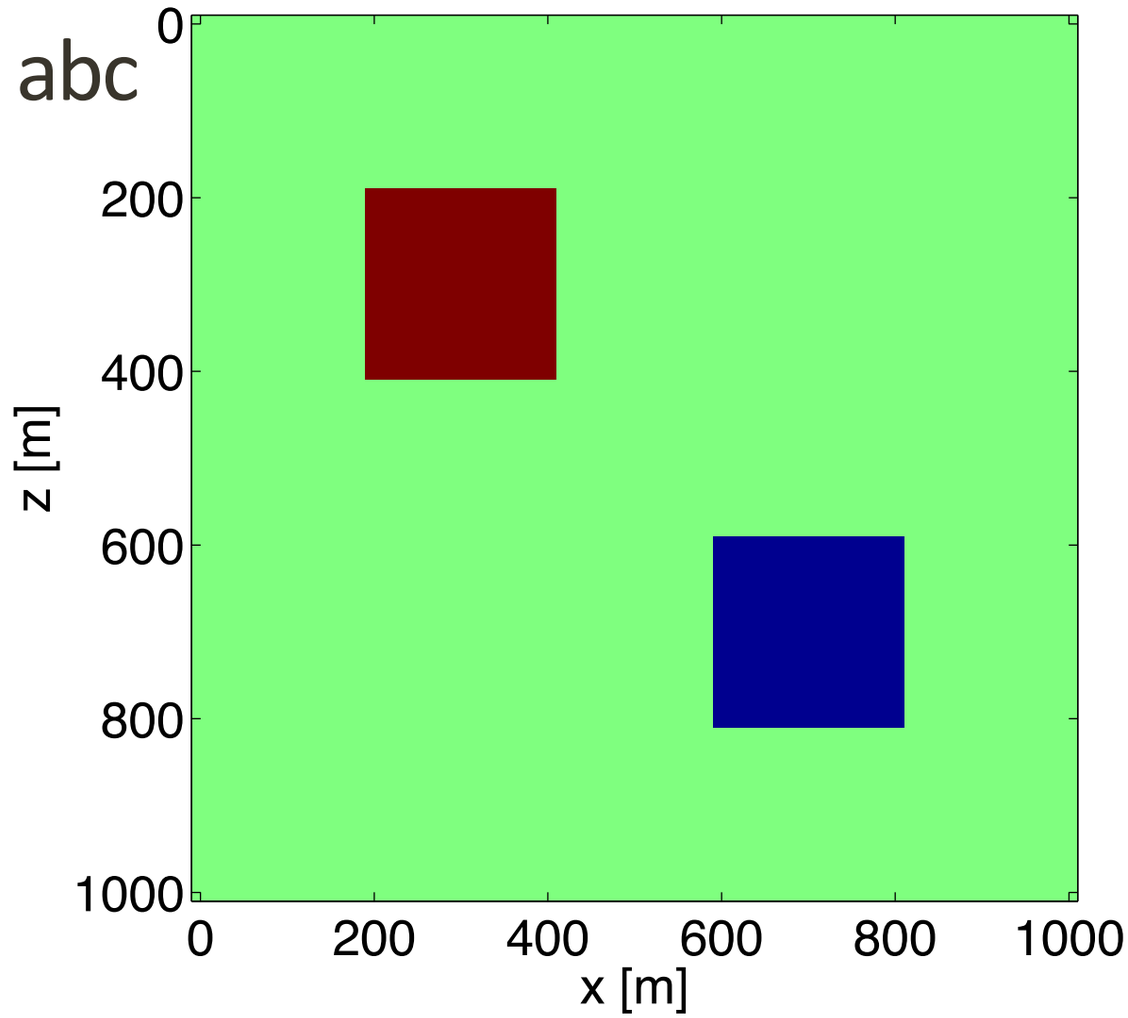
After *1000* iterations:



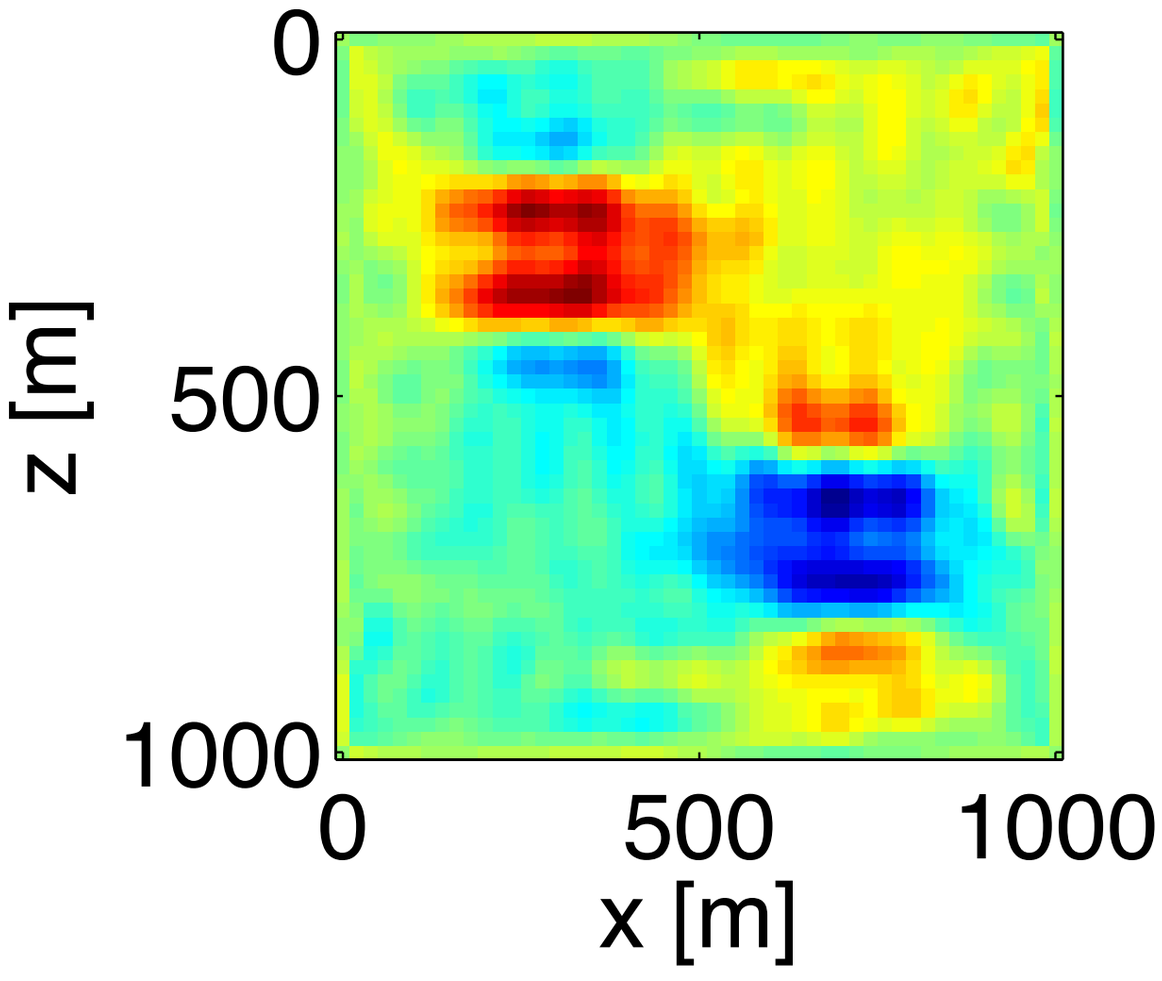


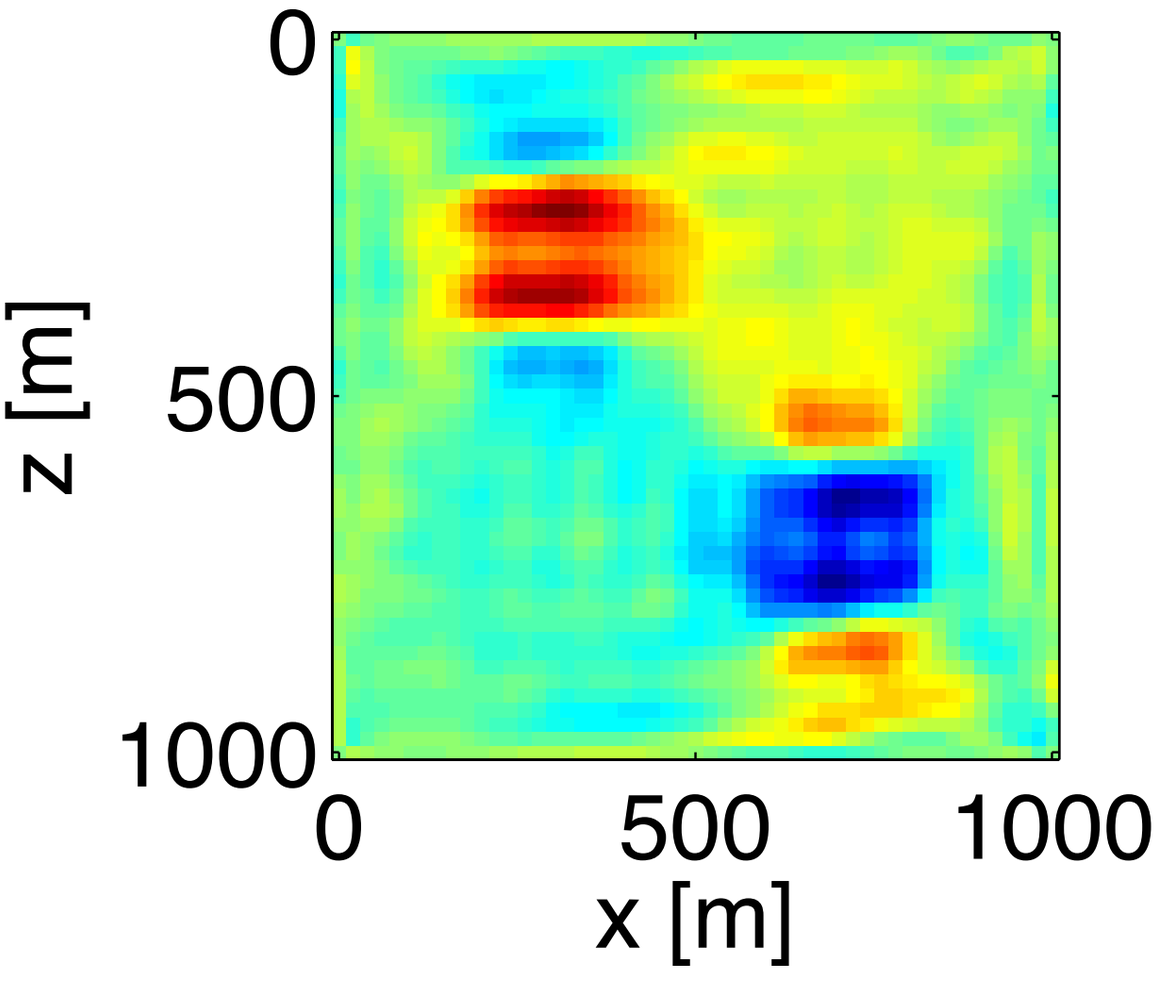
Cross-well FWI

- 2D cross-well configuration
- 5 point Helmholtz FD operator with abc
- Direct solver for PDEs (`A\b`)
- invert a single frequency (10 Hz)
- L-BFGS

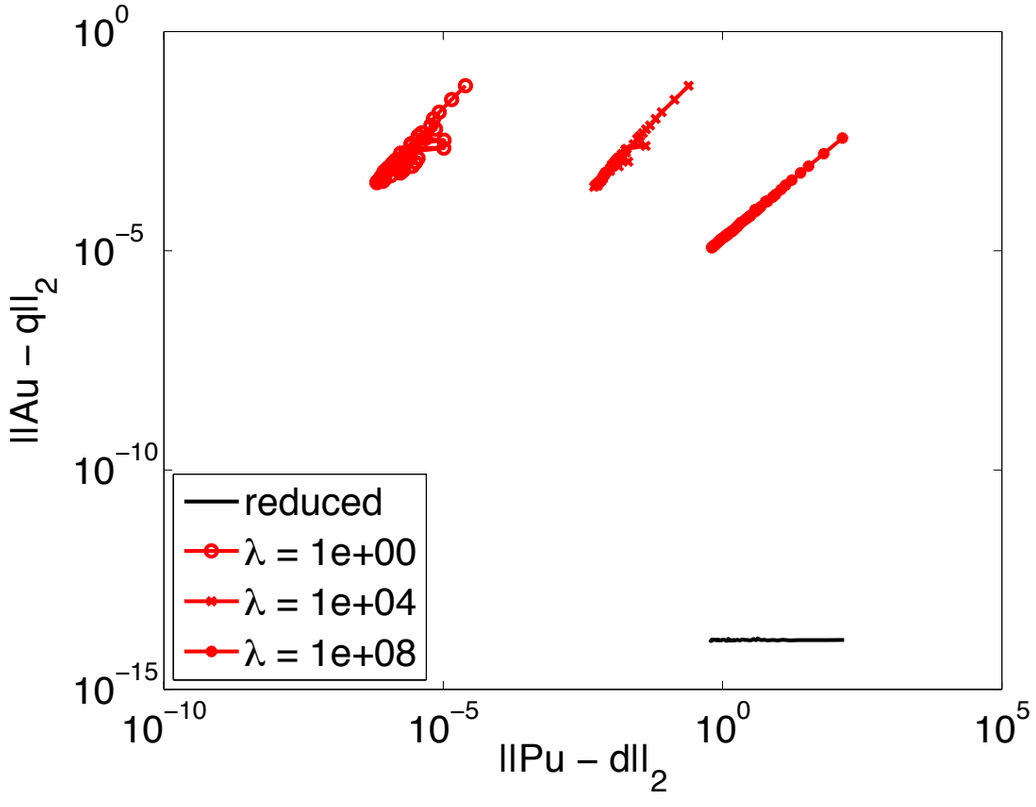


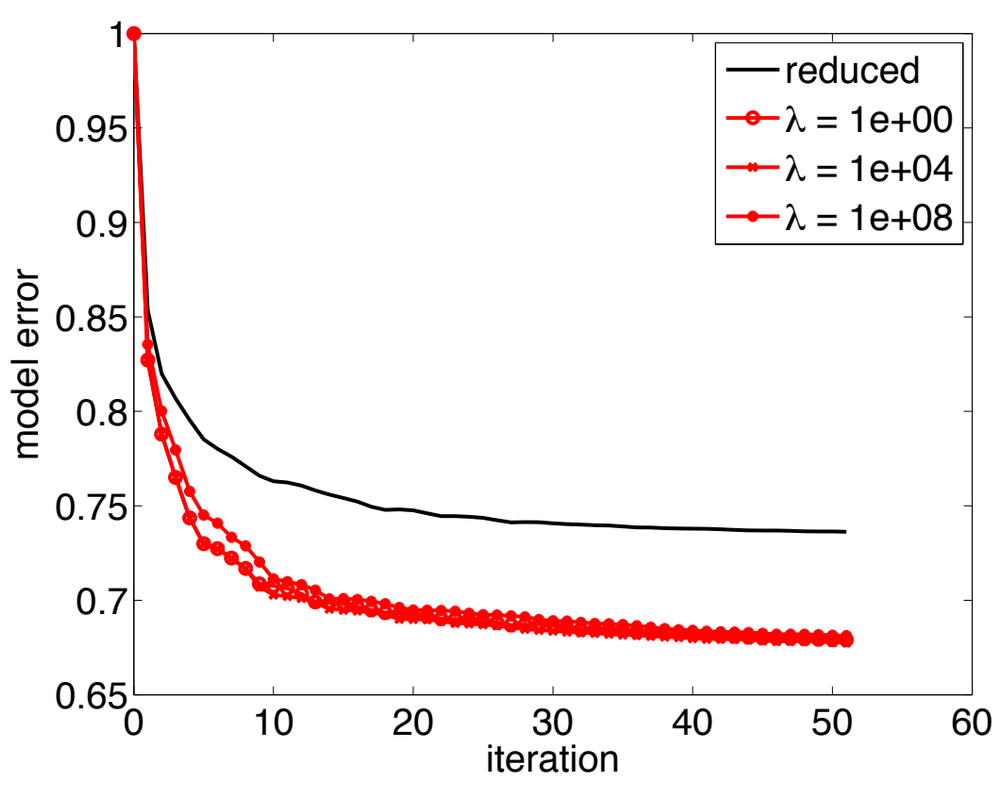
Cross-well FWI





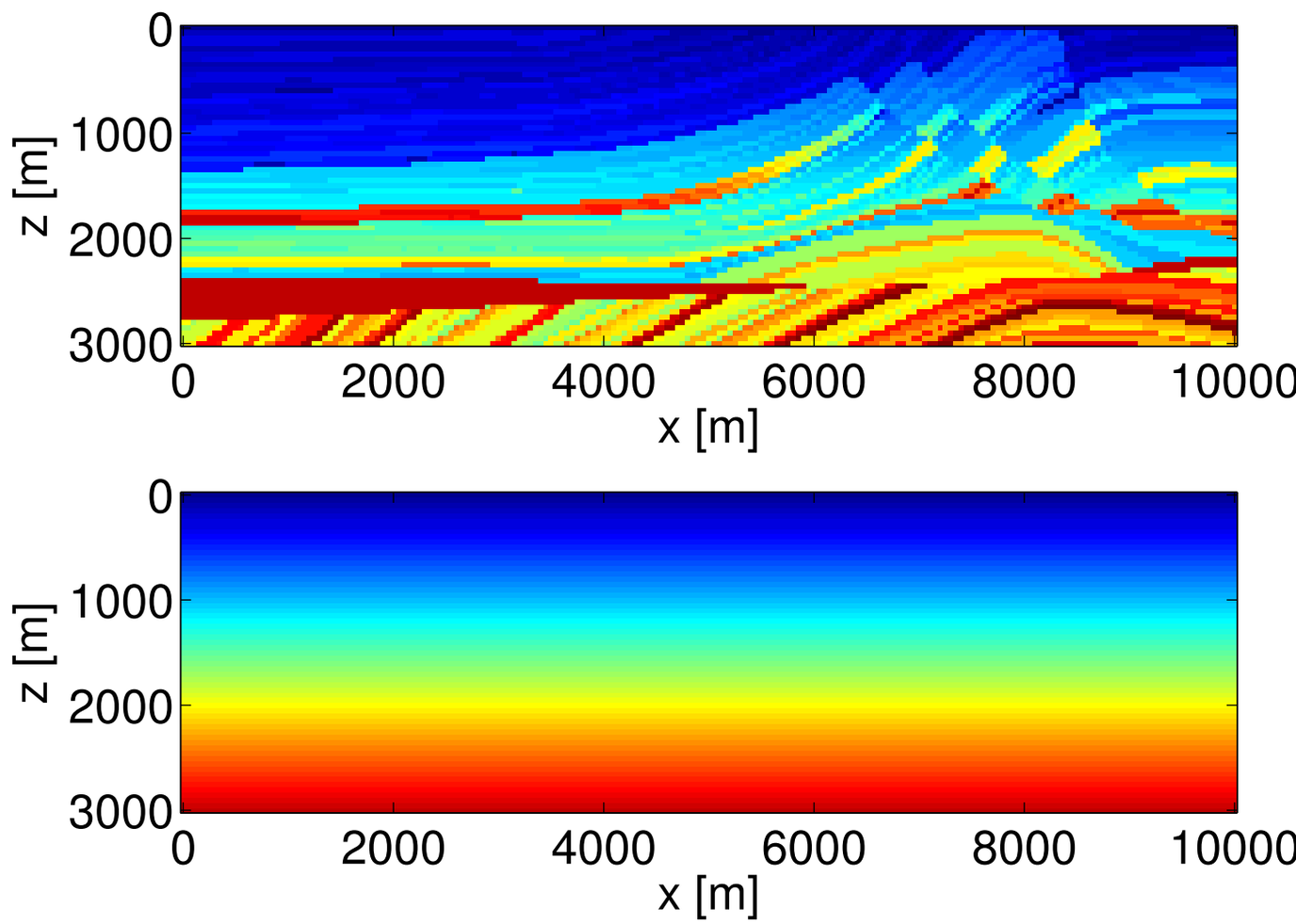
Cross-well FWI





Waveform inversion

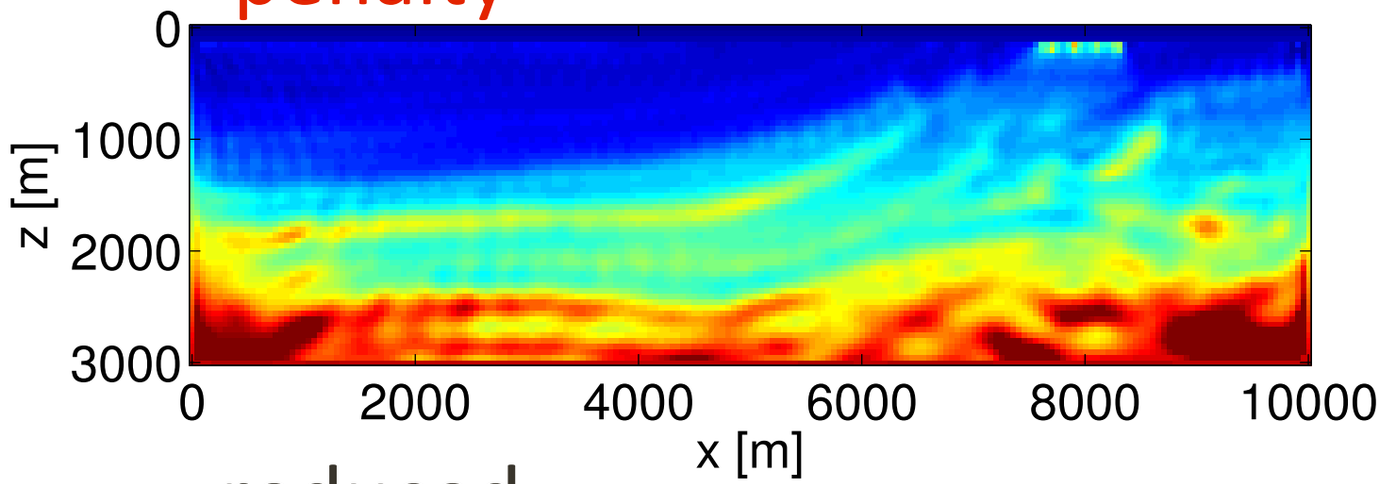
- 2D reflection configuration
- 5 point Helmholtz FD operator with abc
- Direct solver for PDEs (`A\b`)
- frequency continuation
- L-BFGS



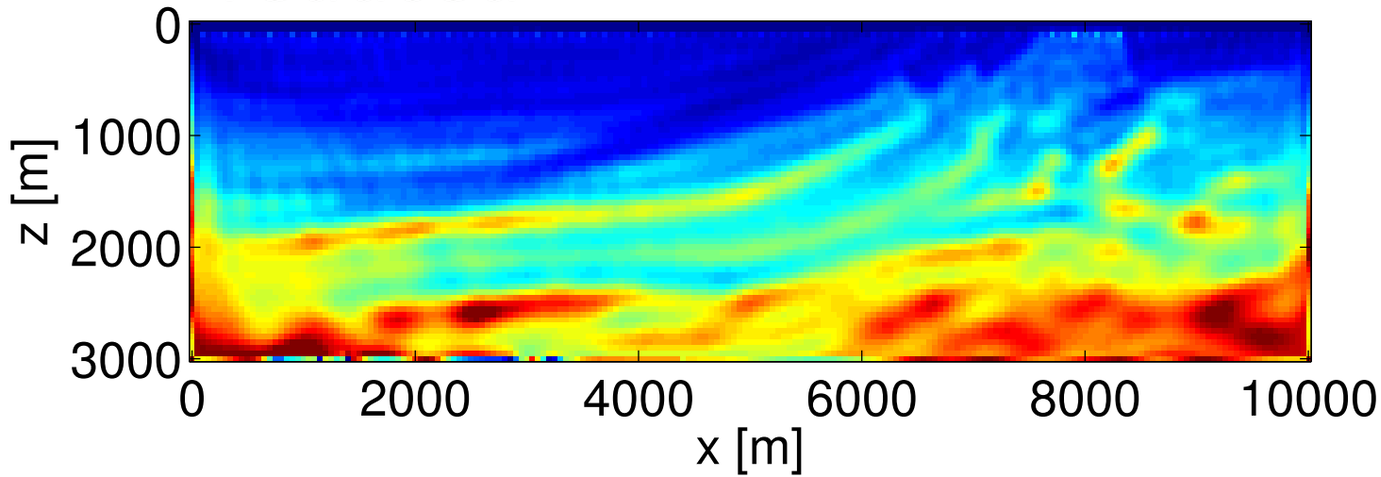
Waveform inversion

frequencies 1-5 Hz, no noise

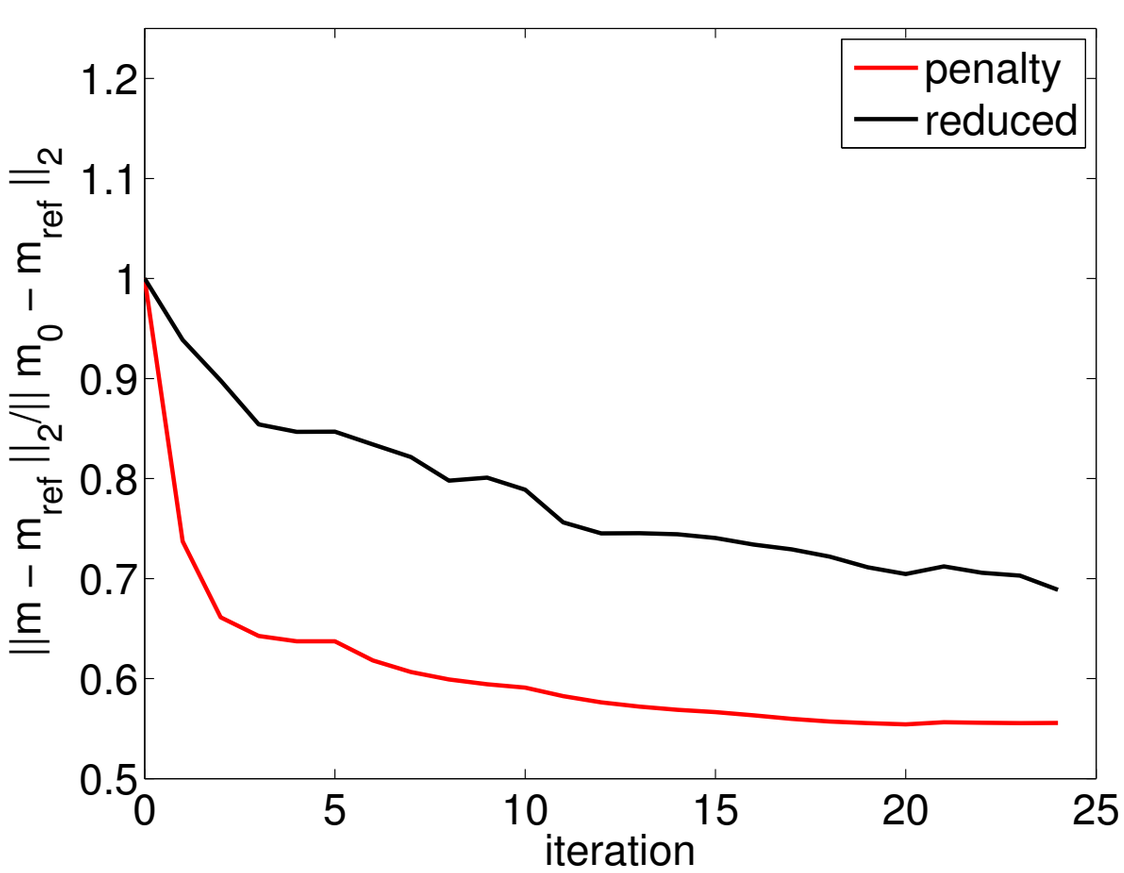
penalty



reduced



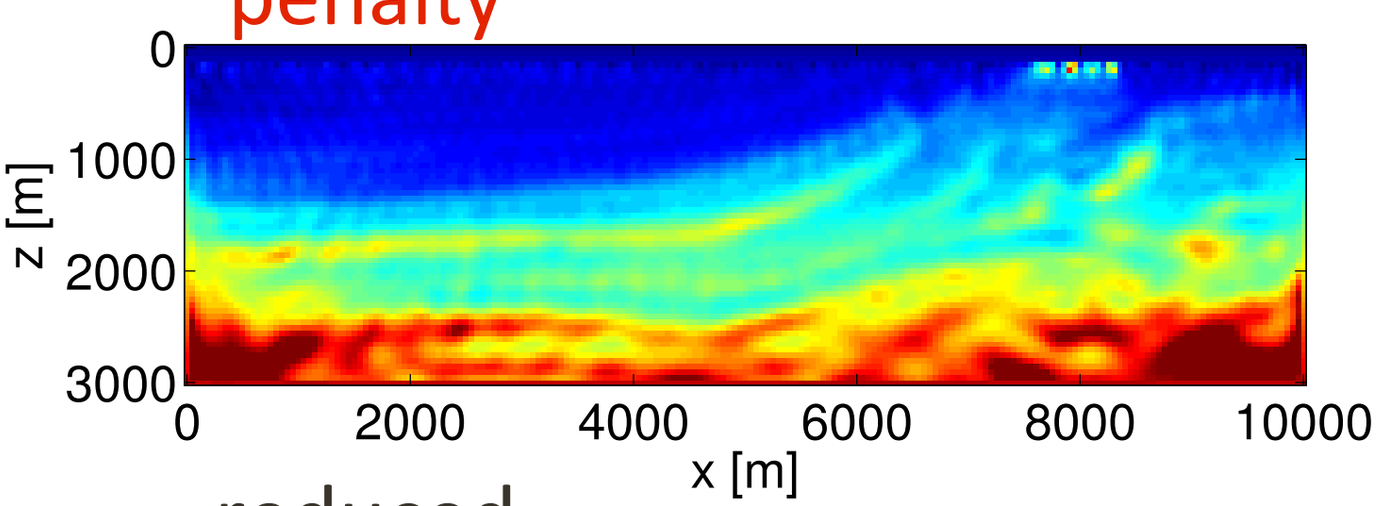
model error



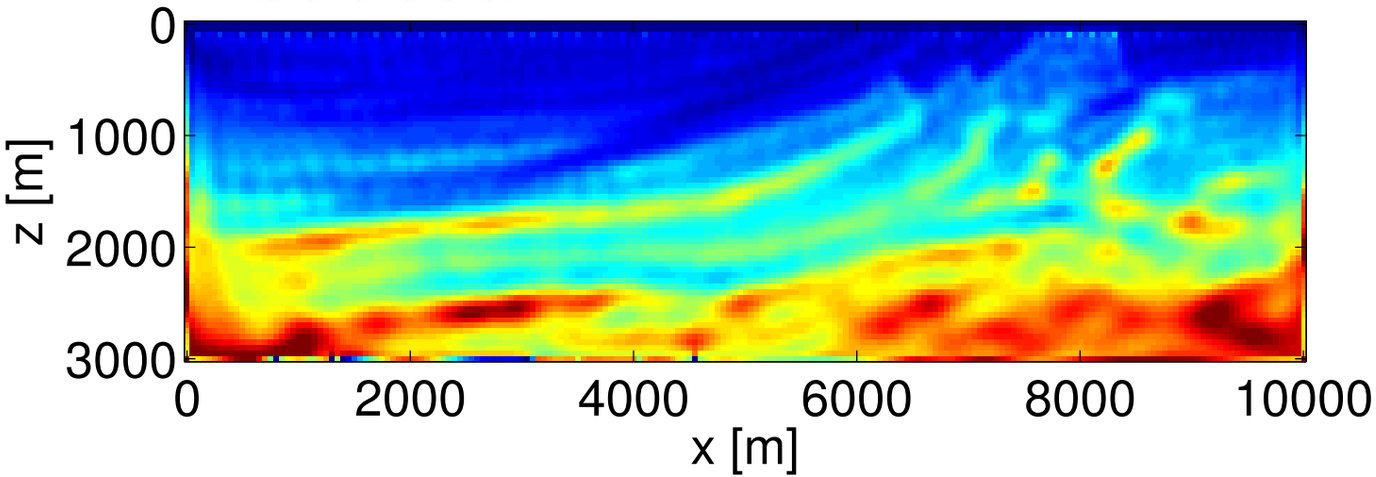
Waveform inversion

frequencies 1-5 Hz, 10 % Gaussian noise

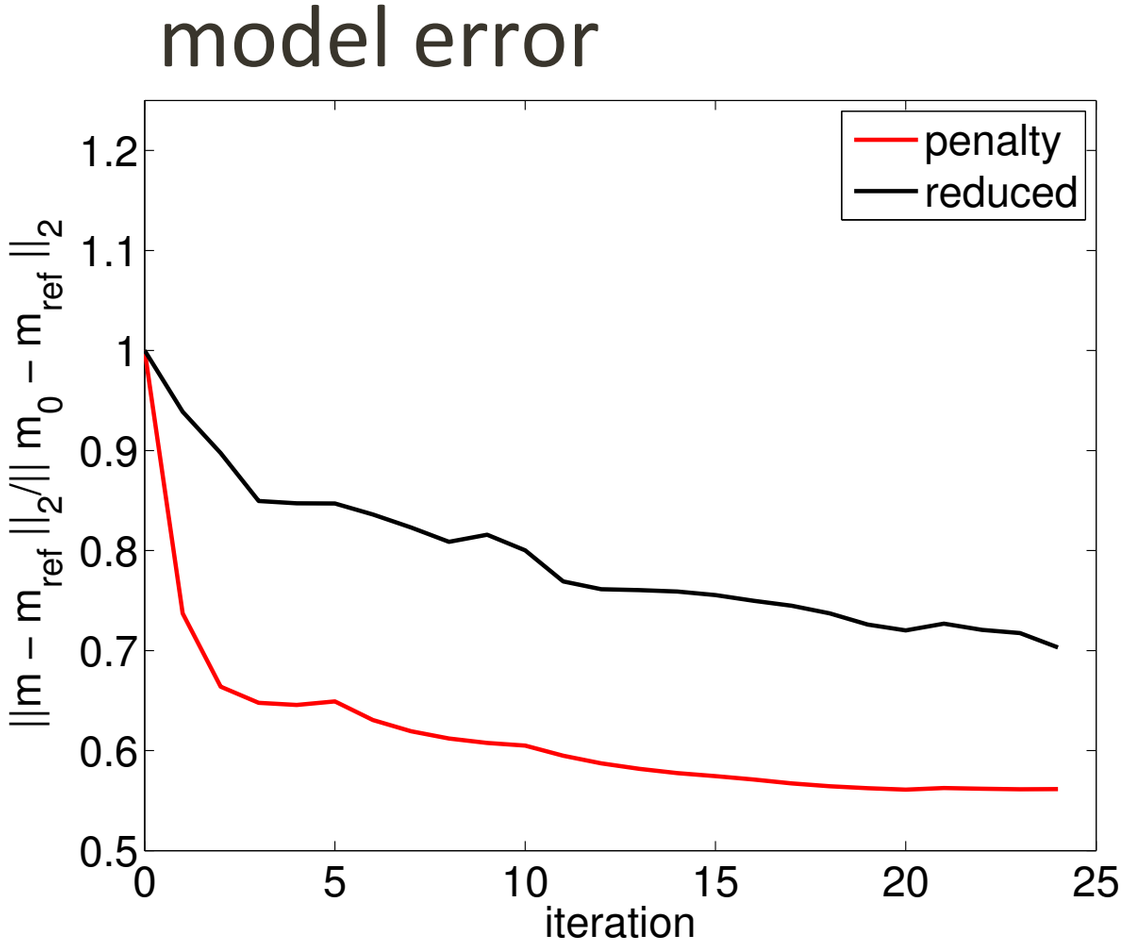
penalty



reduced



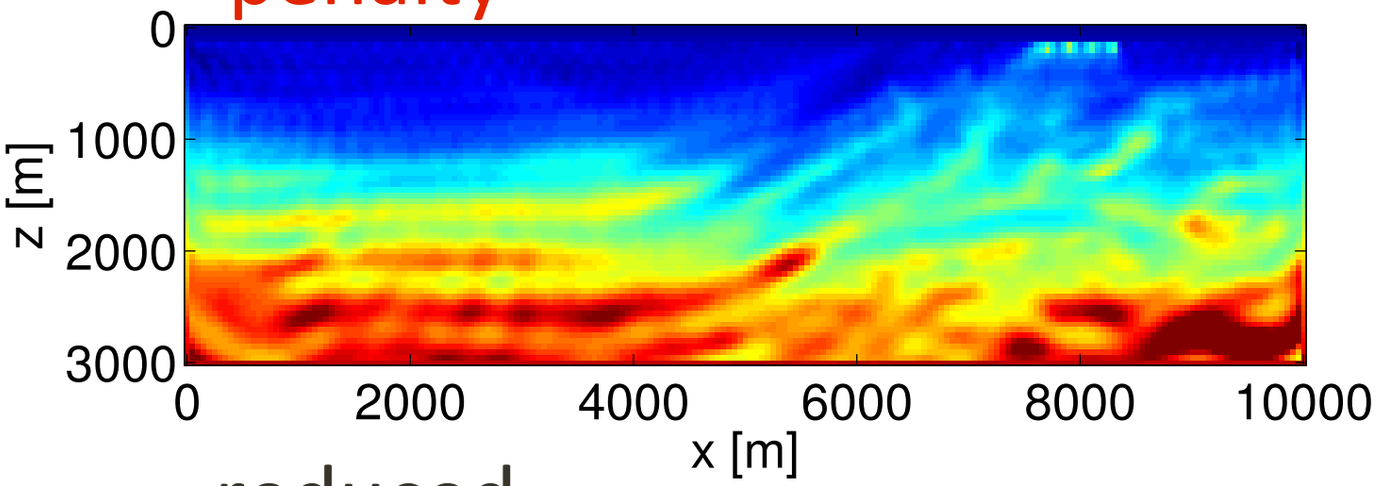
model error



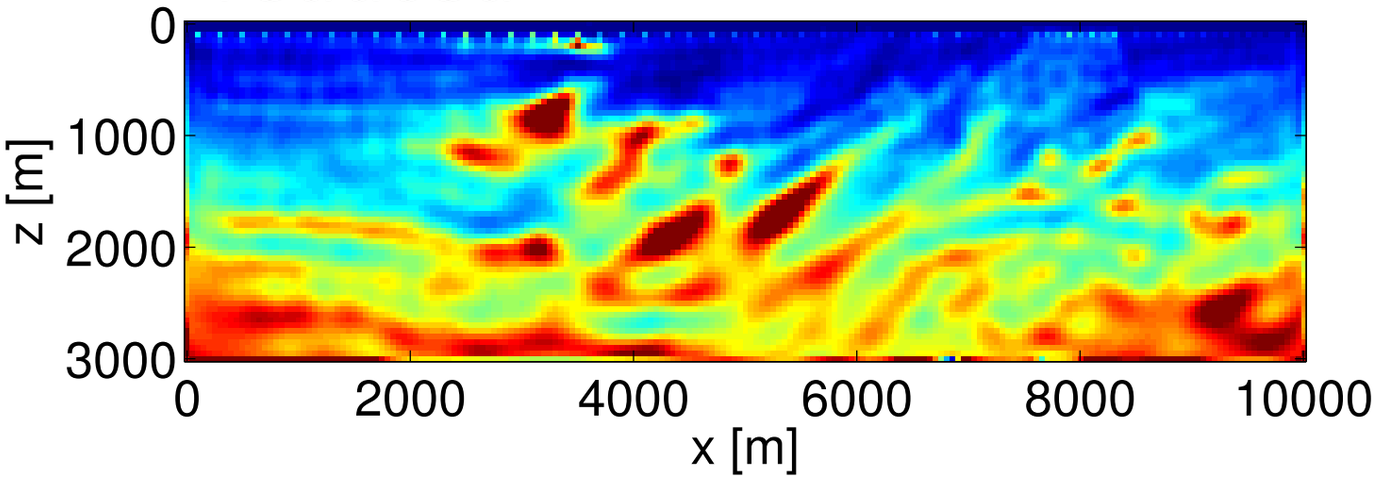
Waveform inversion

- frequencies 2-5 Hz, 10 % Gaussian noise

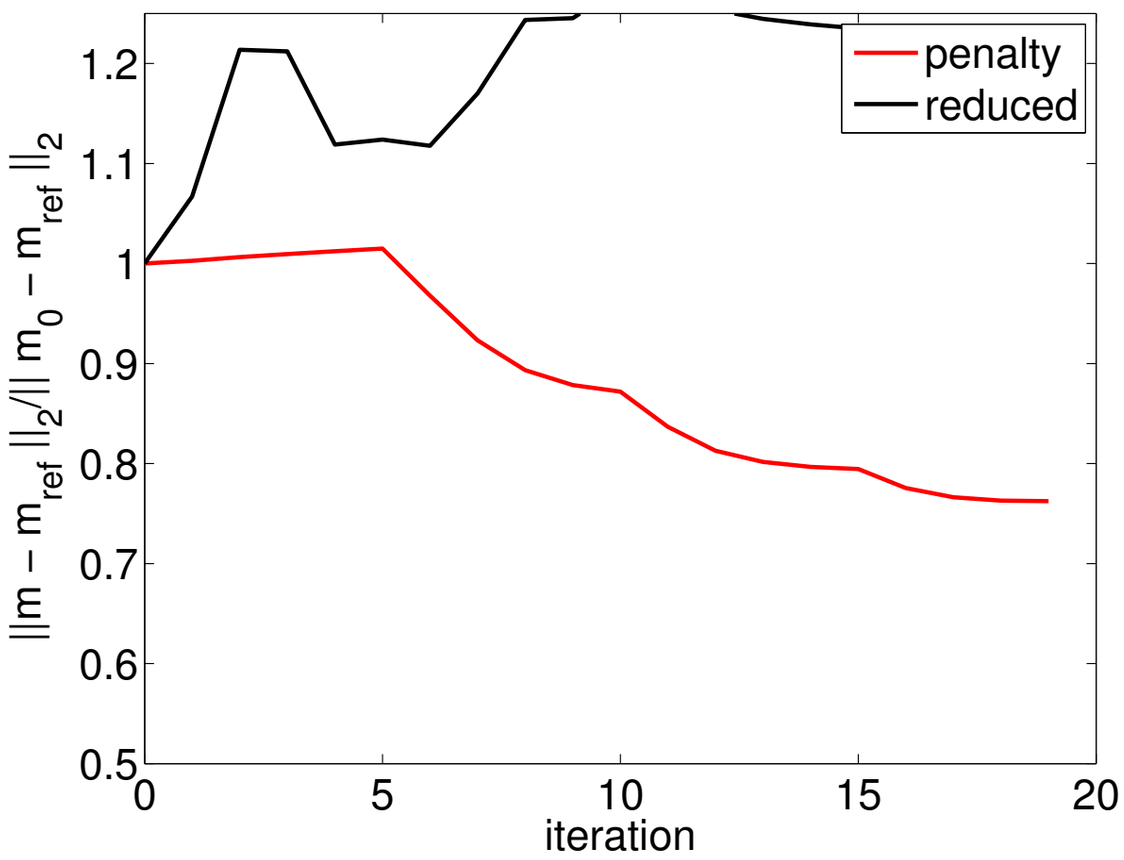
penalty



reduced

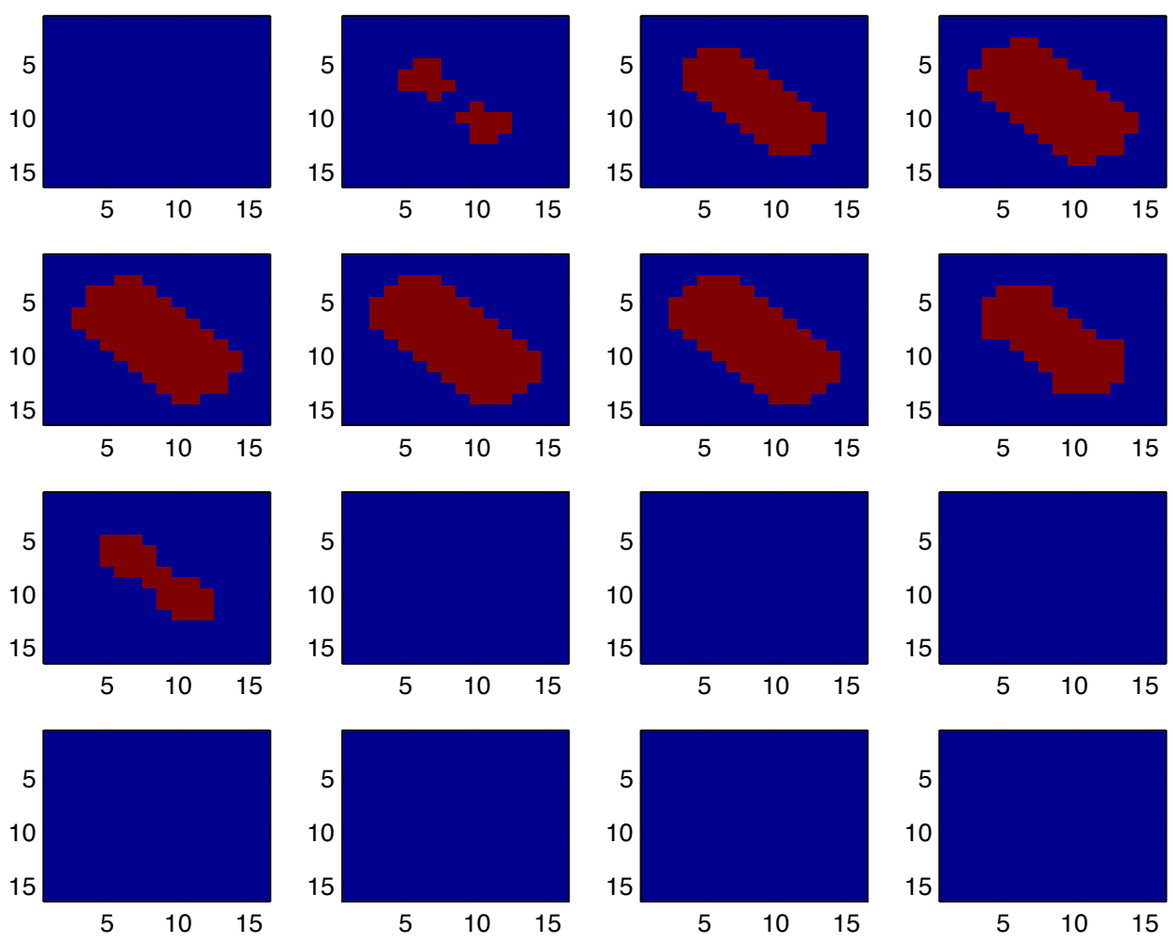


model error



3D DC resistivity

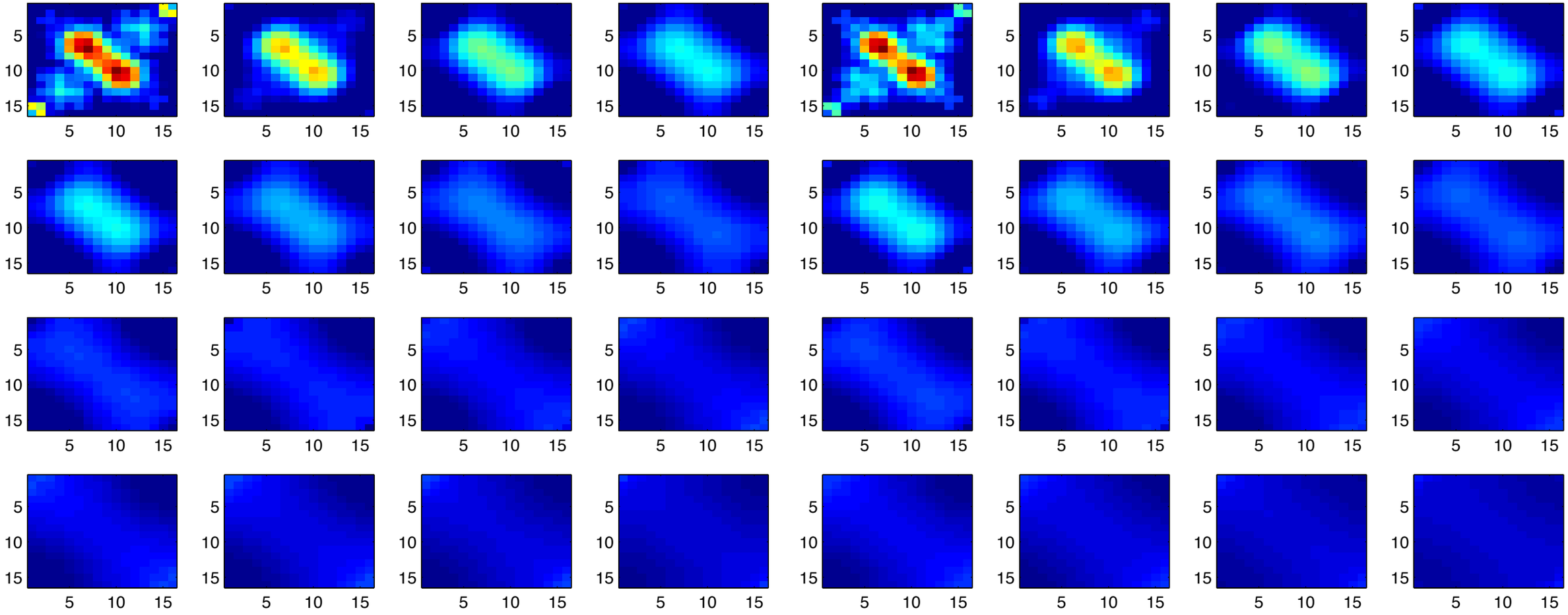
- PDE: $\nabla \cdot \sigma \nabla u = q$
- 16x16x16 grid
- 83 sources (crosswell)
- rec. on the surface
- FD discretization [Haber ’07]
- iterative solvers for PDEs
- L-BFGS



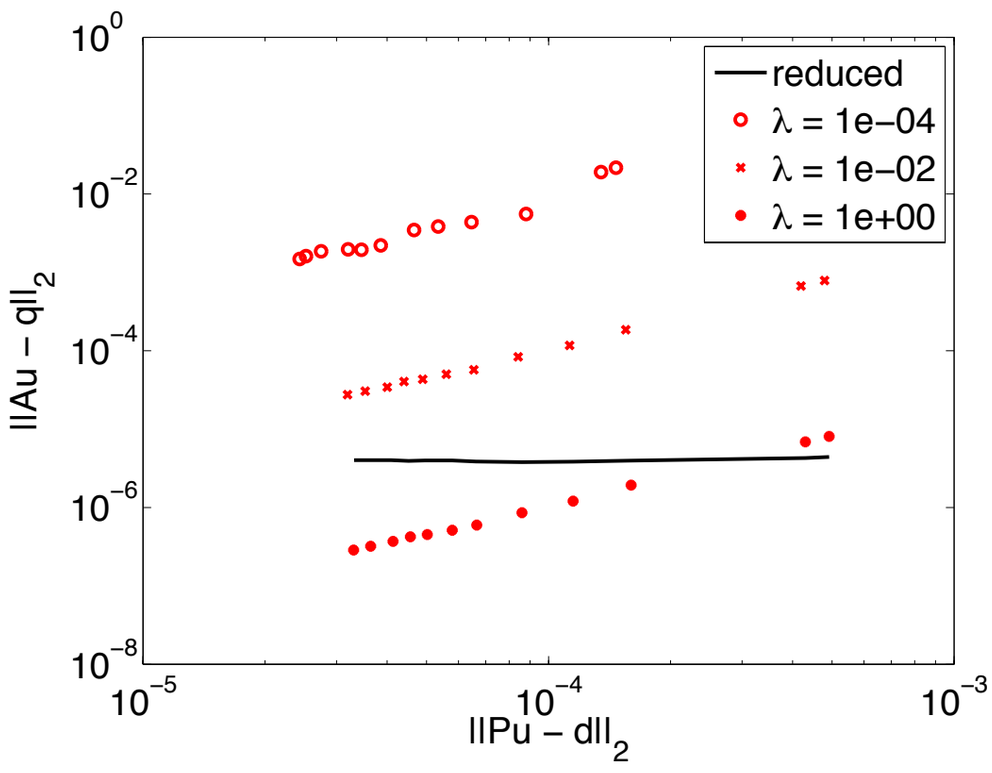
3D DC resistivity

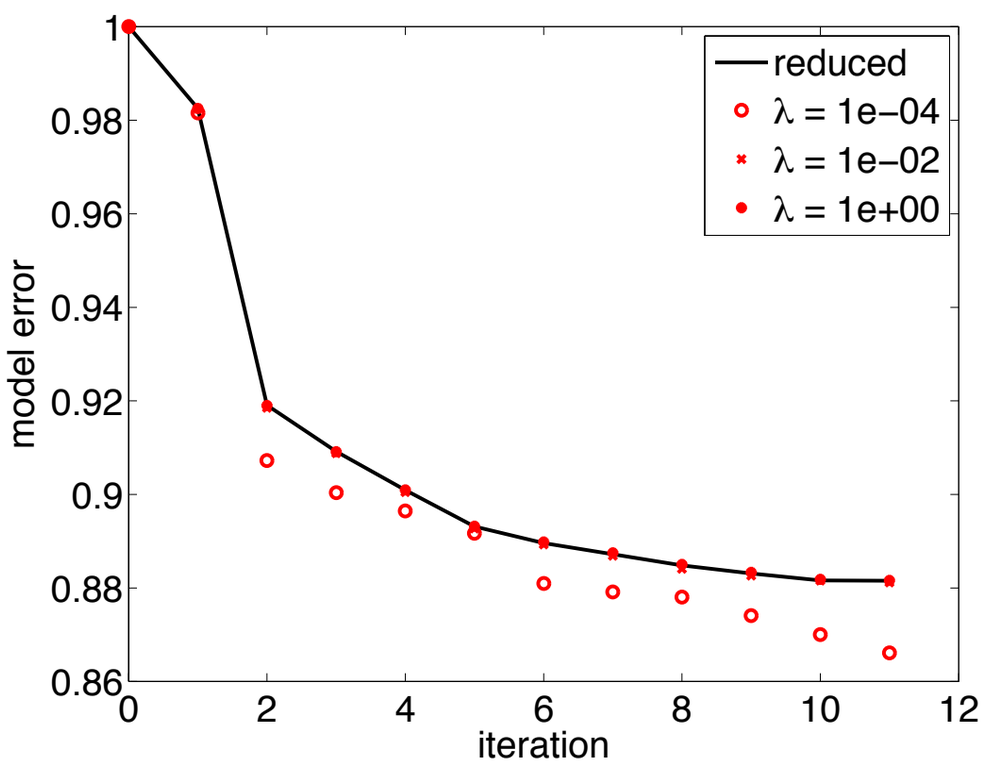
reduced

penalty



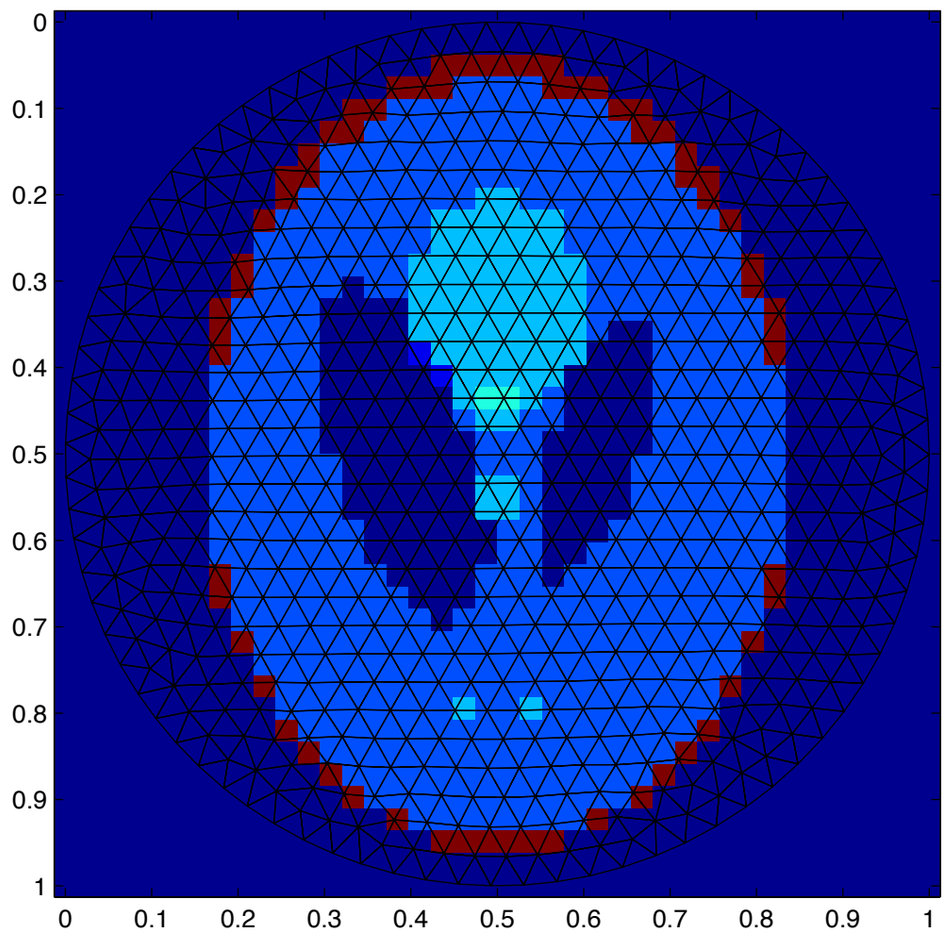
3D DC resistivity





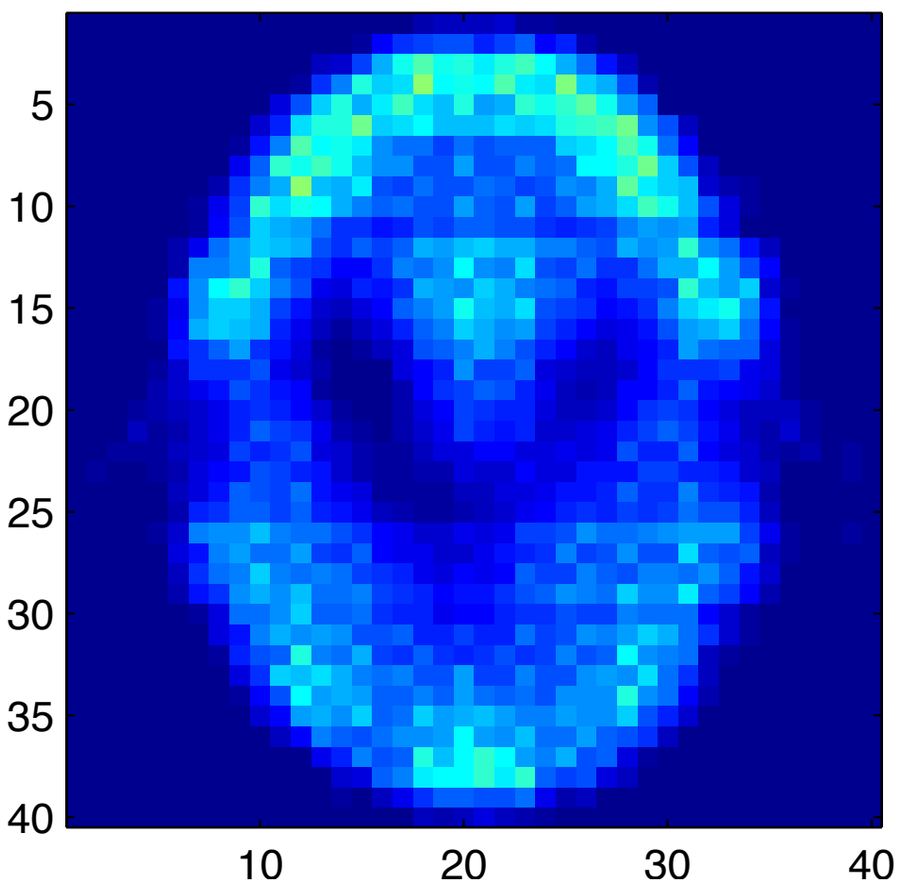
Optical Tomography

- PDE: $(-\imath\omega/c + \Omega \cdot \nabla + \mu_t)u = 0$
- FV discretization [Abdoulaev '05]
- 1 frequency (400 MHz), 21 angles
- Direct solvers for PDEs
- L-BFGS

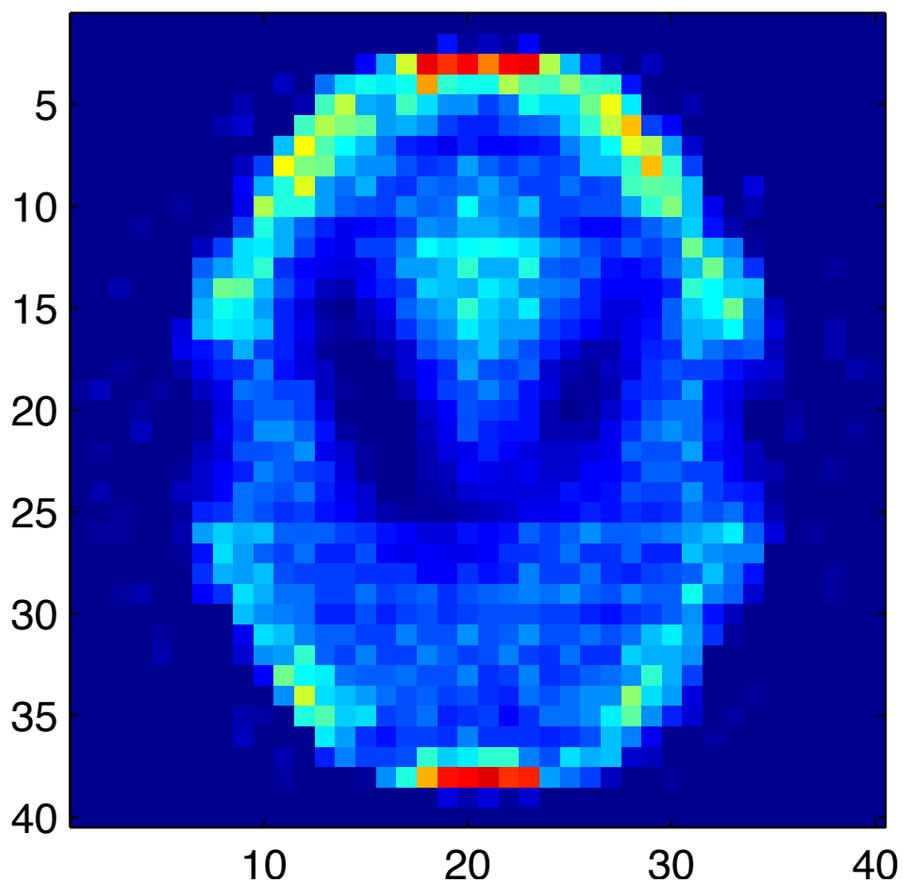


Optical Tomography

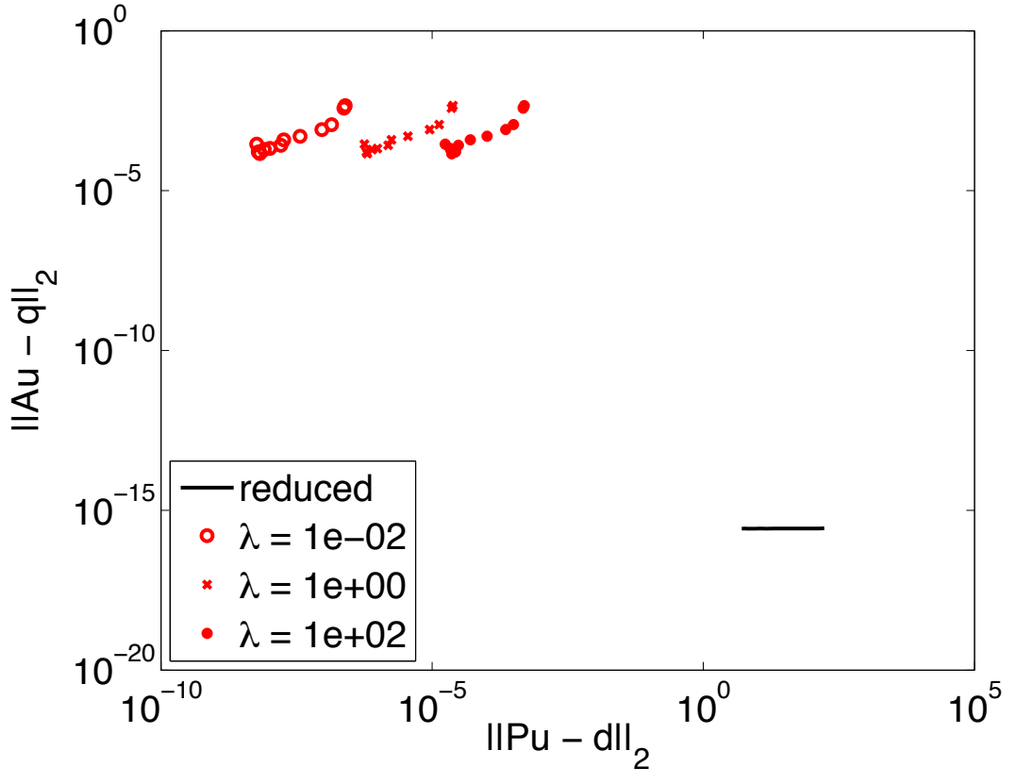
reduced

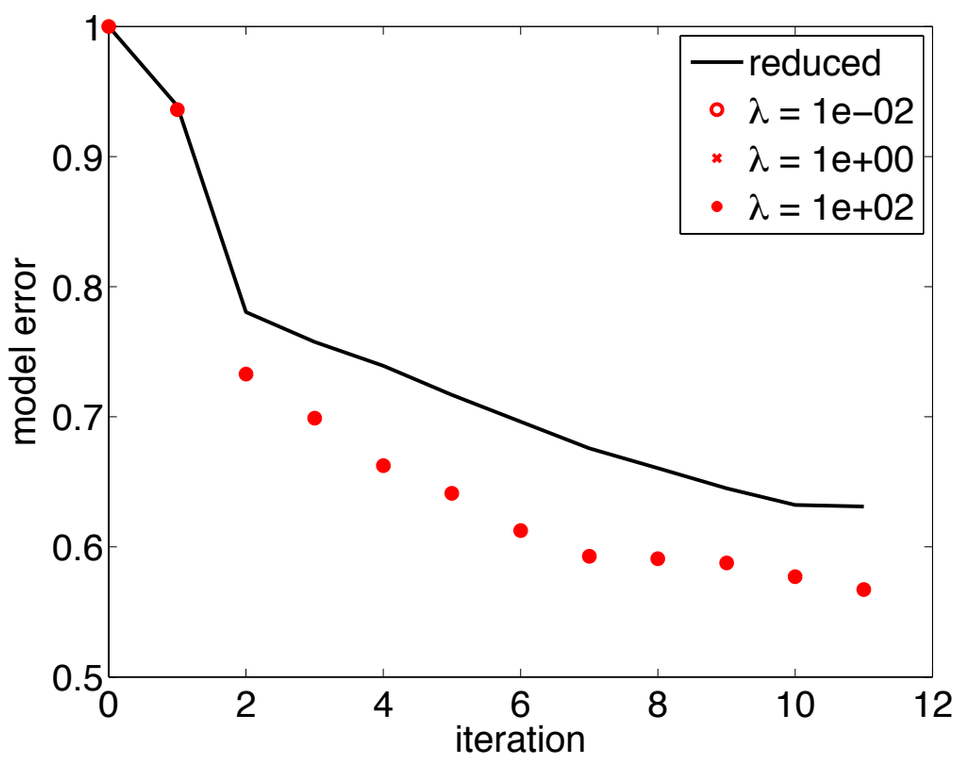


penalty



Optical Tomography





Conclusions

New method for wave-equation based inversion:

- *extended* search space as in *all-at-once* similar but with memory & CPU requirements as in *reduced* approach
- *no* adjoints & *sparse* GN-Hessian approximation
- `less non-linear' problem

Applicable to other PDE-constrained problems:

- DC resistivity
- Optical tomography

Future work

- fast solves for augmented PDE, preconditioning
- (GN) Hessian approximations
- misfit penalties & (sparse) regularization
- other PDEs
- extensions to combat non-uniqueness
- …