Constraints versus penalties for edge-preserving full-waveform inversion

Felix J. Herrmann

SLIM

University of British Columbia

Constraints versus penalties for edge-preserving full-waveform inversion

Felix J. Herrmann & Bas Peters

Bas Peters and Felix J. Herrmann, "Constraints versus penalties for edge-preserving full-waveform inversion". 2016. Submitted to TLE.

SLIM
University of British Columbia

Anagaw, A. Y., and Sacchi, M. D., 2011, Full waveform inversion with total variation regularization
Epanomeritakis, I., Akçelik, V., Ghattas, O., and Bielak, J., 2008, A newton-CG method for large-scale three-dimensional elastic full-waveform seismic inversion: Inverse Problems, 24, 034015.

Motivation

Full-waveform inversion (FWI):

- hampered by poor data & parasitic local minima
- ill-posed <=> missing frequencies & finite aperture
- should benefit from bounds & structure-promoting priors (TV- or ℓ_1-norms)

Efforts met w/ limited success:

- unpredictable dependence on (unnecessary) hyper parameters
- poor conditioning of structure promoting regularization
- difficulties handling multiple pieces of prior information

FWI

Unconstrained optimization problem:

$$\underset{\mathbf{m}\in\mathbb{R}^m}{\text{minimize}}\ \overbrace{f(\mathbf{m})}^{\text{objective}} = \frac{1}{2}\sum_{i=1}^{N_s}\|\overbrace{F(\mathbf{m})}^{\text{modelling}}\mathbf{q}_i - \overbrace{\mathbf{d}_i}^{\text{data}}\|_2^2$$

Local derivative information is used to update the model:

$$\underbrace{\mathbf{m}_{k+1}}_{\text{iterate}} = \mathbf{m}_k - \gamma\underbrace{\nabla_{\mathbf{m}} f(\mathbf{m}_k)}_{\text{gradient}}$$

- no insurance model iterates remain (physically/geologically) feasible
- no mitigation of inversion artifacts by controlling model's complexity

Stylized example

Forward model:

$$\mathbf{d} = F(\mathbf{c})\mathbf{q} \equiv \mathbf{c} * \mathbf{q}$$

Unconstrained inversion:

$$\underset{\mathbf{c} \in \mathbb{R}^m}{\text{minimize}} \ \frac{1}{2} \|F(\mathbf{c})\mathbf{q} - \mathbf{d}\|_2^2$$

- when source **q** misses low frequencies
- w/o regularization

Stylized example w/ constraints

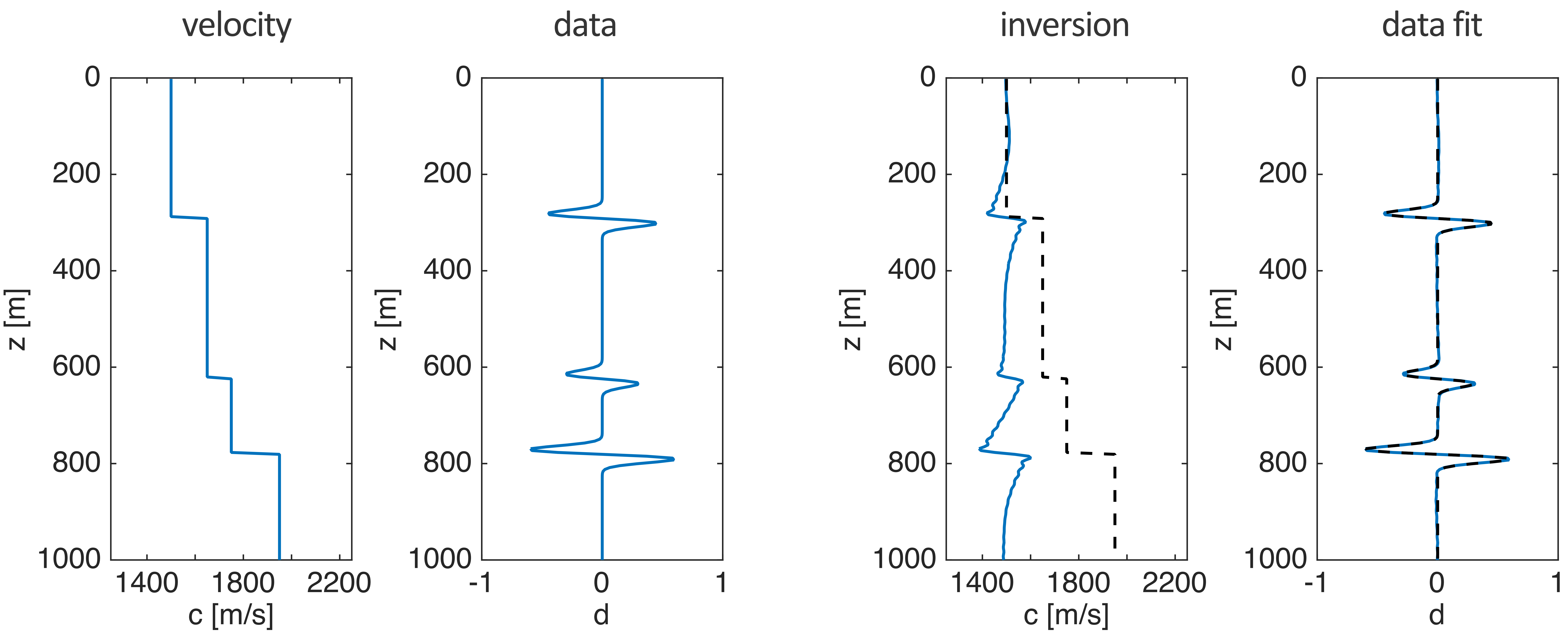
Regularization via constraints on model:

$$\underset{\mathbf{c}\geq\mathbf{c}_0}{\text{minimize}}\ \frac{1}{2}\|F(\mathbf{c})\mathbf{q}-\mathbf{d}\|_2^2 \quad \text{subject to} \quad \mathbf{D}\mathbf{c}\geq\mathbf{0}$$

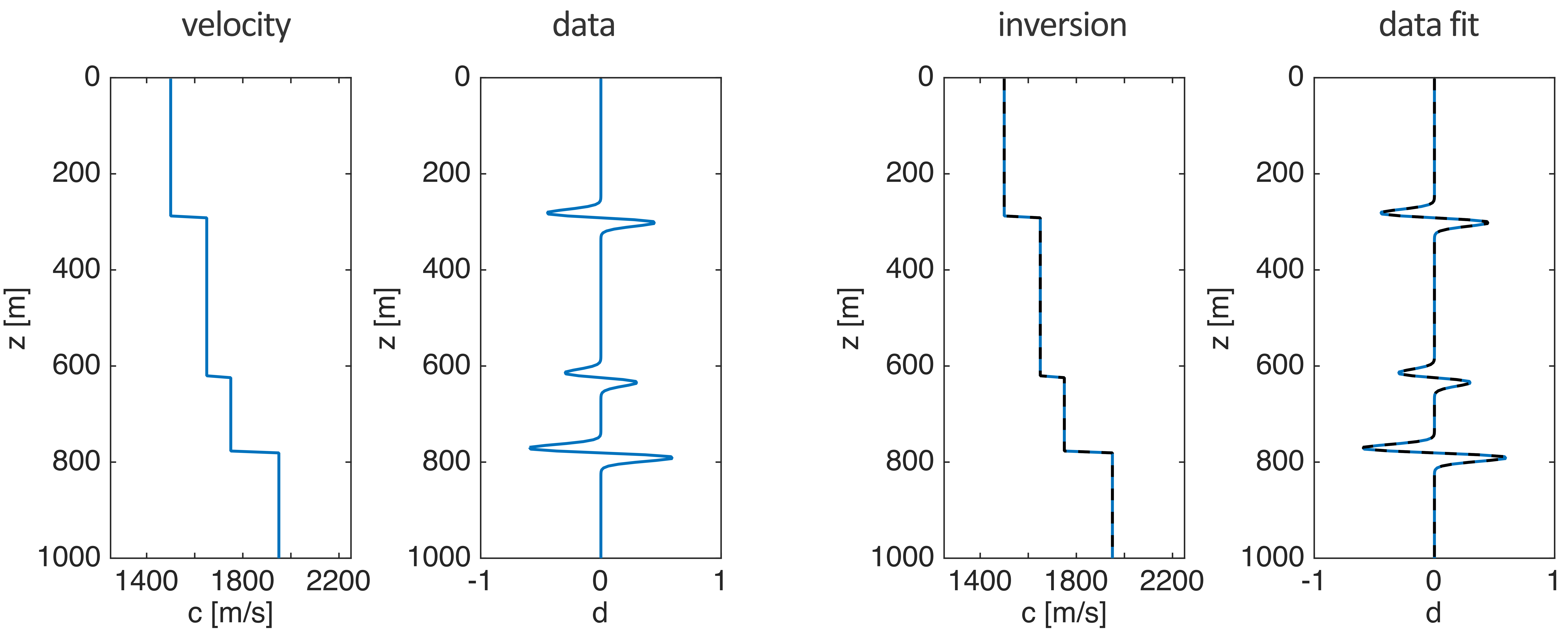
- minimal velocity
- monotonic increasing gradient of the velocity

Leads to successful recovery...

Inversion w/o constraints



Inversion w/ constraints

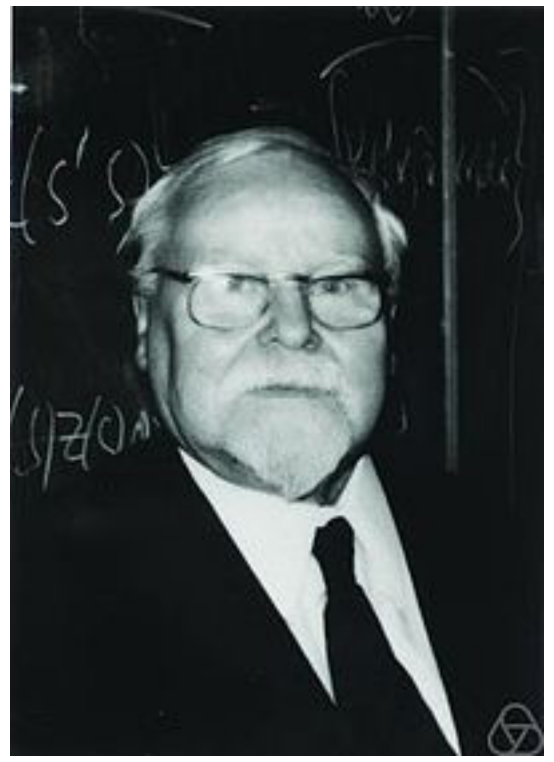


Tikhonov regularization

Add quadratic penalty terms:

$$\underset{\mathbf{m}}{\text{minimize}}\, f(\mathbf{m}) + \frac{\alpha}{2}\|R_1\mathbf{m}\|^2 + \frac{\beta}{2}\|R_2\mathbf{m}\|^2$$

- well-known & successful technique
- is differentiable
- not an exact penalty
- regularization may adversely affect gradient & Hessian
- requires non-trivial choices for hyper parameters
- not easily extended to edge-preserving ℓ_1 - norms
- **no guarantees that all model iterates are regularized**



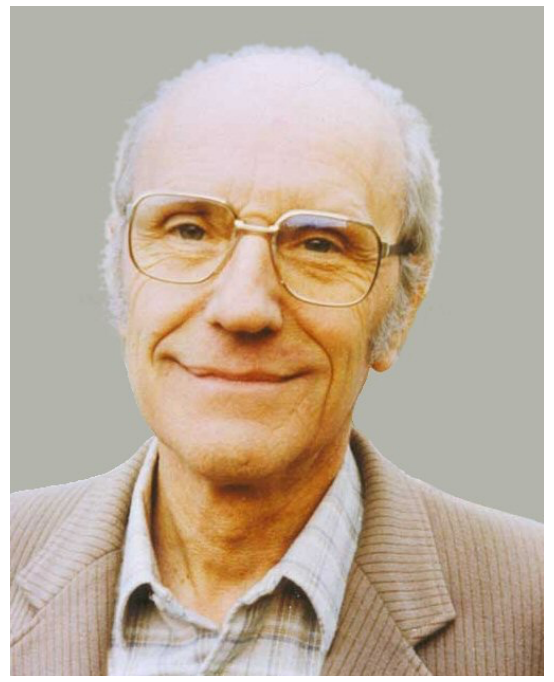
Andrey Tikhonov
1906–1993

Regularization w/ constraints

Add multiple constraints:

$$\underset{\mathbf{m}}{\text{minimize}}\, f(\mathbf{m}) \quad \text{subject to} \quad \mathbf{m} \in \mathcal{C}_1 \bigcap \mathcal{C}_2$$

- not well-known in our community
- requires understanding of latest optimization techniques
- does not affect gradient & Hessian
- easier parameterization
- able to uniquely project onto intersection of multiple constraint sets
- constraints do not need to be differentiable
- **constraints are satisfied at *every* model iterate**



Jean Jacques Moreau
1923–2014

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J., 2011, doi:10.1561/2200000016

SLIM

POCS vs. best approximation

"Projection"-onto-convex-sets solves convex *feasibility* problems:

$$\text{find} \quad \mathbf{m} \in \mathcal{C}_1 \bigcap \mathcal{C}_2$$

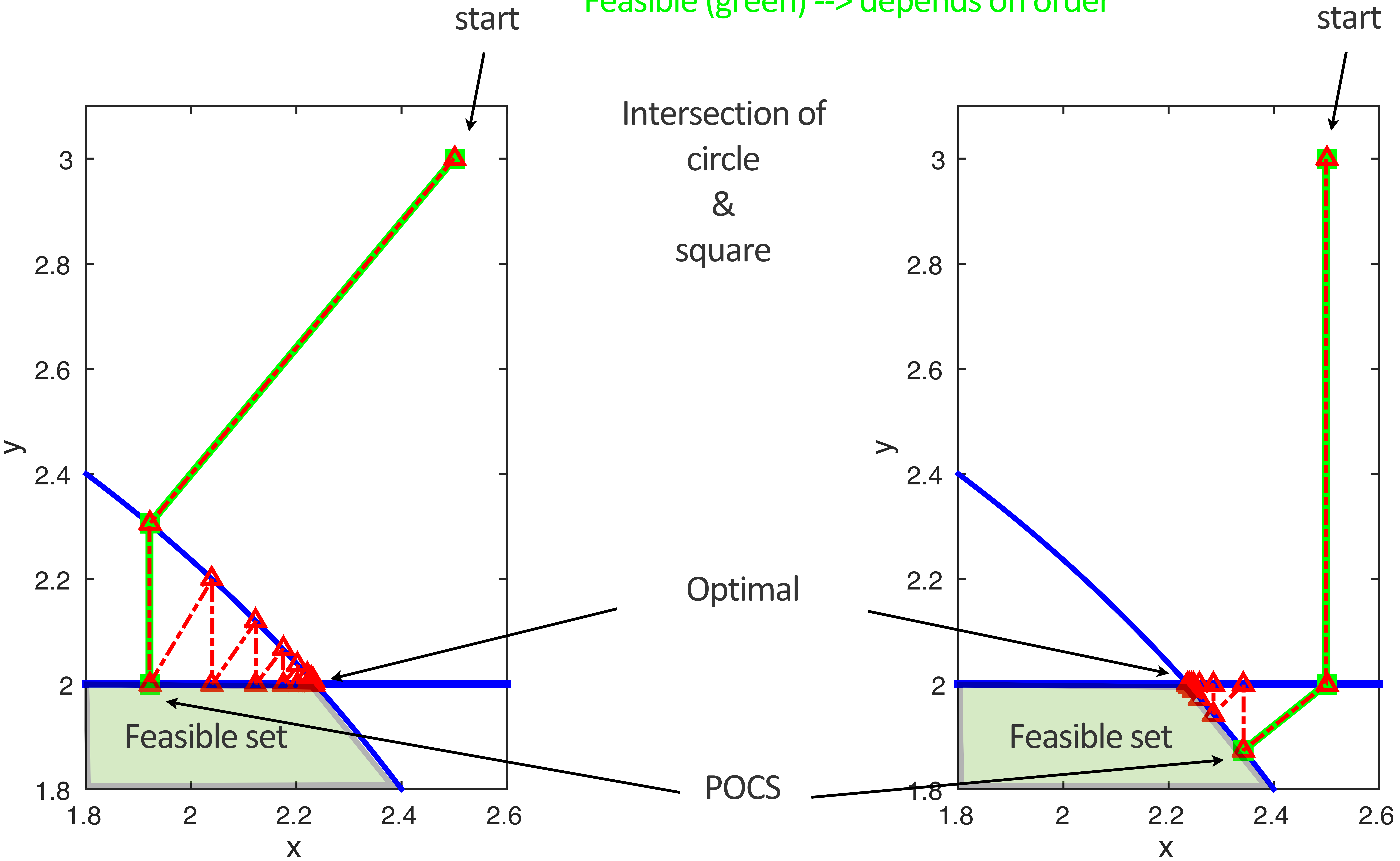
Instead, we solve convex *projection* problems:

$$\underset{\mathbf{m}}{\text{minimize}} \, \|\mathbf{m} - \mathbf{x}\|_2^2 \quad \text{subject to} \quad \mathbf{m} \in \mathcal{C}_1 \bigcap \mathcal{C}_2$$

- obtain optimal (also feasible) approximations
- project *uniquely* w/ DYKSTRA onto intersections of convex sets

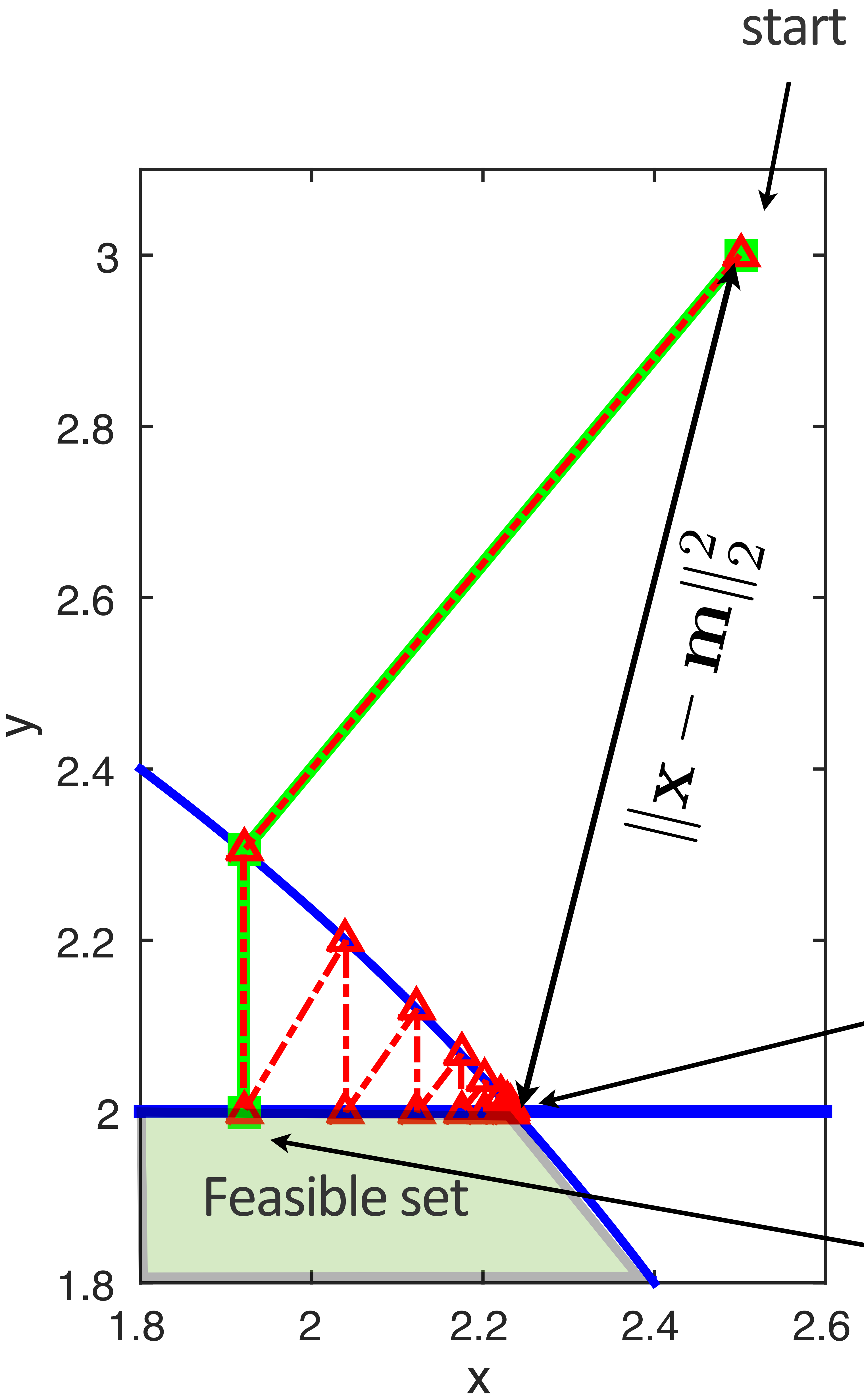
Feasibility vs optimal

Optimal (red) --> unique solution
Feasible (green) --> depends on order



Optimal (red) --> unique solution

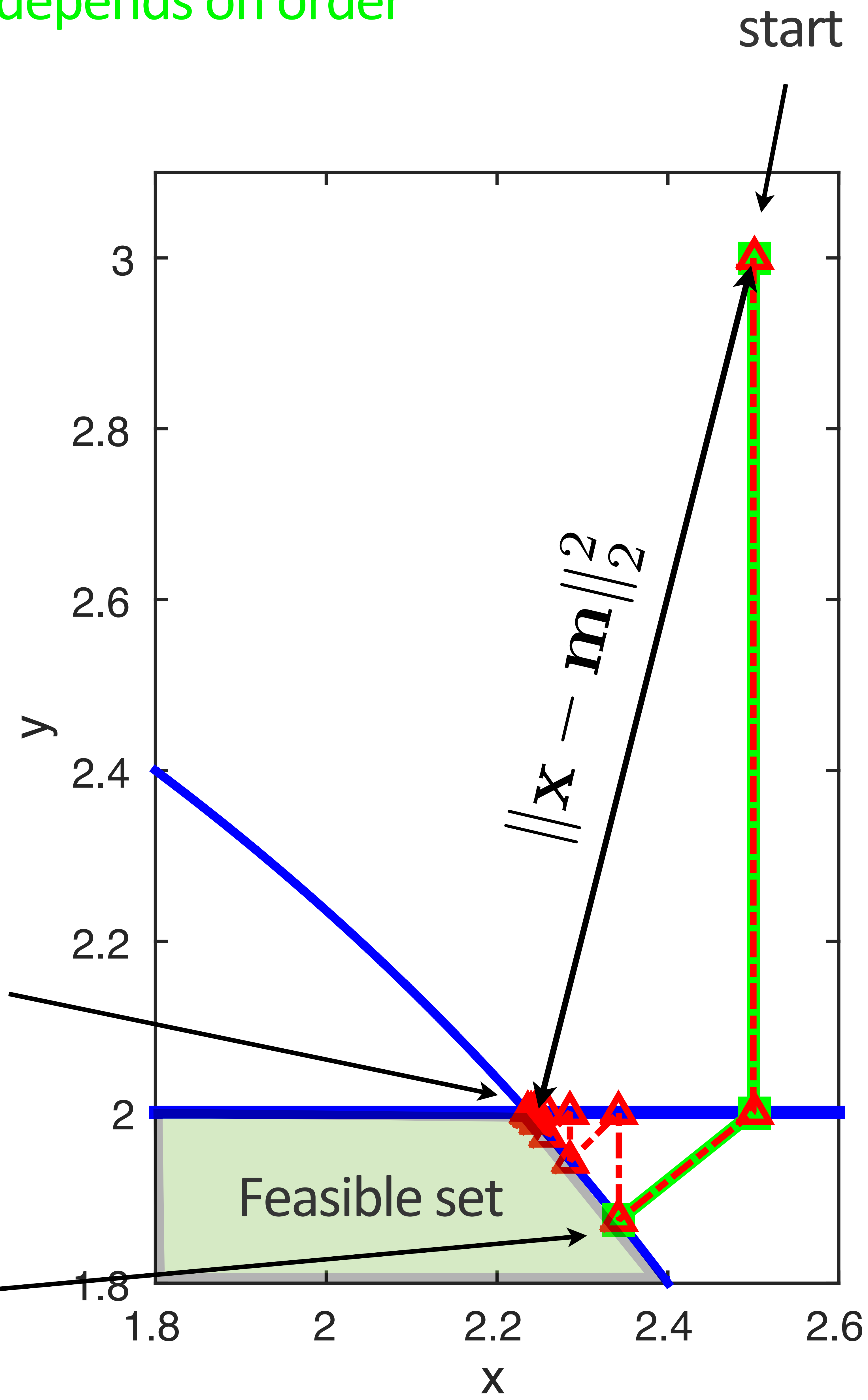
Feasible (green) --> depends on order



Intersection of circle & square

Optimal

POCS



Our constraints

bounds: $\mathcal{C}_1 = \{\mathbf{m} : \mathbf{m} \in \mathrm{Box}\}$ where $\mathbf{m} \in \mathrm{Box}$ means

$$l_{i,j} \leq m_{i,j} \leq u_{i,j} \quad \forall i, j$$

total-variation norm ball: $\mathcal{C}_2 = \{\mathbf{m} : \mathrm{TV}(\mathbf{m}) \leq \tau\}$

$$\mathrm{TV}(\mathbf{m}) = \frac{1}{h} \sum_{ij} \sqrt{(m_{i+1,j} - m_{i,j})^2 + (m_{i,j+1} - m_{i,j})^2}$$

Proximal projection

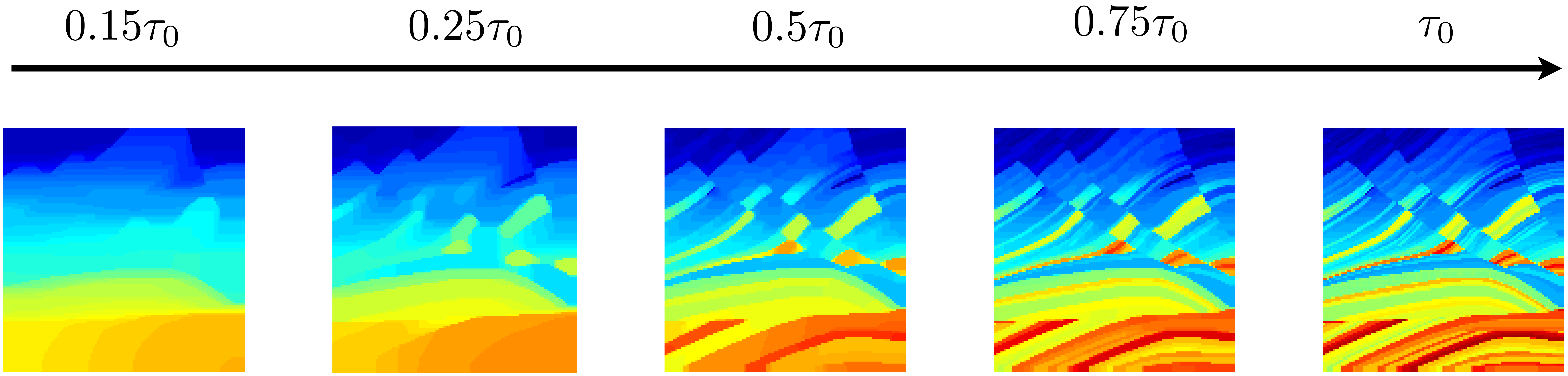
Find a model $\mathbf{m}$, closest to $\mathbf{x}$, such that it satisfies the constraints:

$$\mathcal{P}_{\mathcal{C}}(\mathbf{x}) = \arg\min_{\mathbf{m}} \|\mathbf{m} - \mathbf{x}\|_2^2 \quad \text{subject to} \quad \mathbf{m} \in \mathcal{C}_1 \bigcap \mathcal{C}_2$$

yields

- nonlinear “minimal complexity” best approximation of $\mathbf{m}$
- $\mathbf{m} \rightarrow \mathbf{x}$ when $\tau \rightarrow \tau_0 = \mathrm{TV}(\mathbf{x})$ and $\mathbf{m} \in \mathrm{Box}$
- edge preserving

Best approximations



FWI w/ non-differentiable penalties

Nonlinear least-squares objective for FWI w/ TV:

$$\underset{\mathbf{m}}{\text{minimize}} \, \|\mathbf{d}^{\text{obs}} - \mathbf{d}^{\text{sim}}(\mathbf{m})\|^2 + \alpha \text{TV}(\mathbf{m})$$

- $\text{TV}(\mathbf{m})$ is not differentiable so no access to $\nabla f(\mathbf{m})$ and $\nabla^2 f(\mathbf{m})$
- gradient-descent/quasi-Newton/(Gauss-Newton) solvers need local derivate information

FWI w/ non-differentiable penalties

Nonlinear least-squares objective for FWI w/ TV:

$$\cancel{\underset{\mathbf{m}}{\text{minimize}} \|\mathbf{d}^{\text{obs}} - \mathbf{d}^{\text{sim}}(\mathbf{m})\|^2 + \alpha \text{TV}(\mathbf{m})}$$

- $\text{TV}(\mathbf{m})$ is not differentiable so no access to $\nabla f(\mathbf{m})$ and $\nabla^2 f(\mathbf{m})$
- gradient-descent/quasi-Newton/(Gauss-Newton) solvers need local derivate information

FWI w/ non-differentiable constraints

Possible solution ϵ-smoothing of TV:

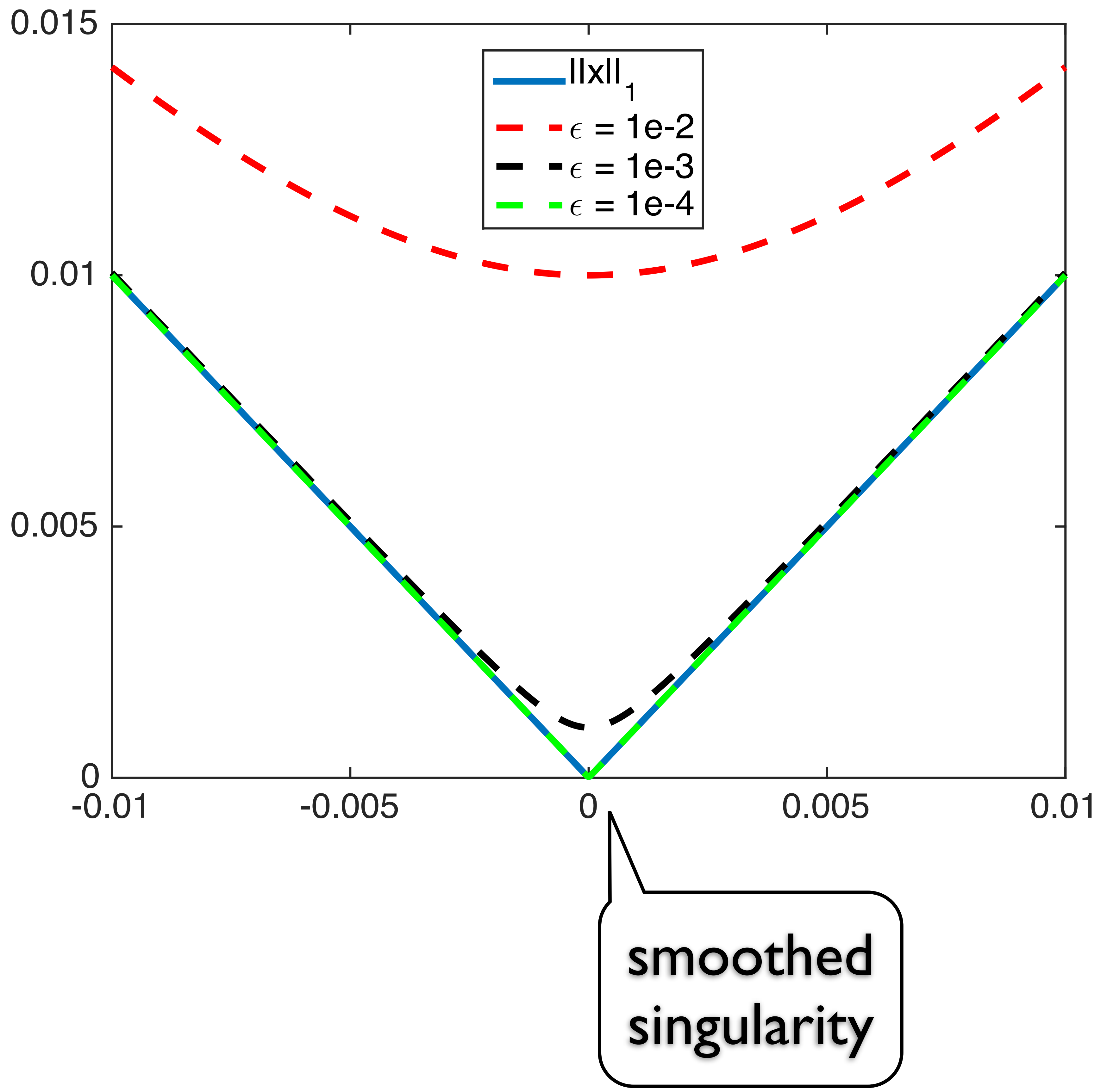
$$\mathrm{TV}_\epsilon(\mathbf{m}) = \frac{1}{h}\sum_{ij}\sqrt{(m_{i+1,j} - m_{i,j})^2 + (m_{i,j+1} - m_{i,j})^2 + \epsilon^2}$$

Differentiable objective with penalty-parameter α:

$$\underset{\mathbf{m}}{\text{minimize}}\; f(\mathbf{m}) + \alpha \mathrm{TV}_\epsilon(\mathbf{m})$$

Problem: need to select 2 unintuitive hyper parameters

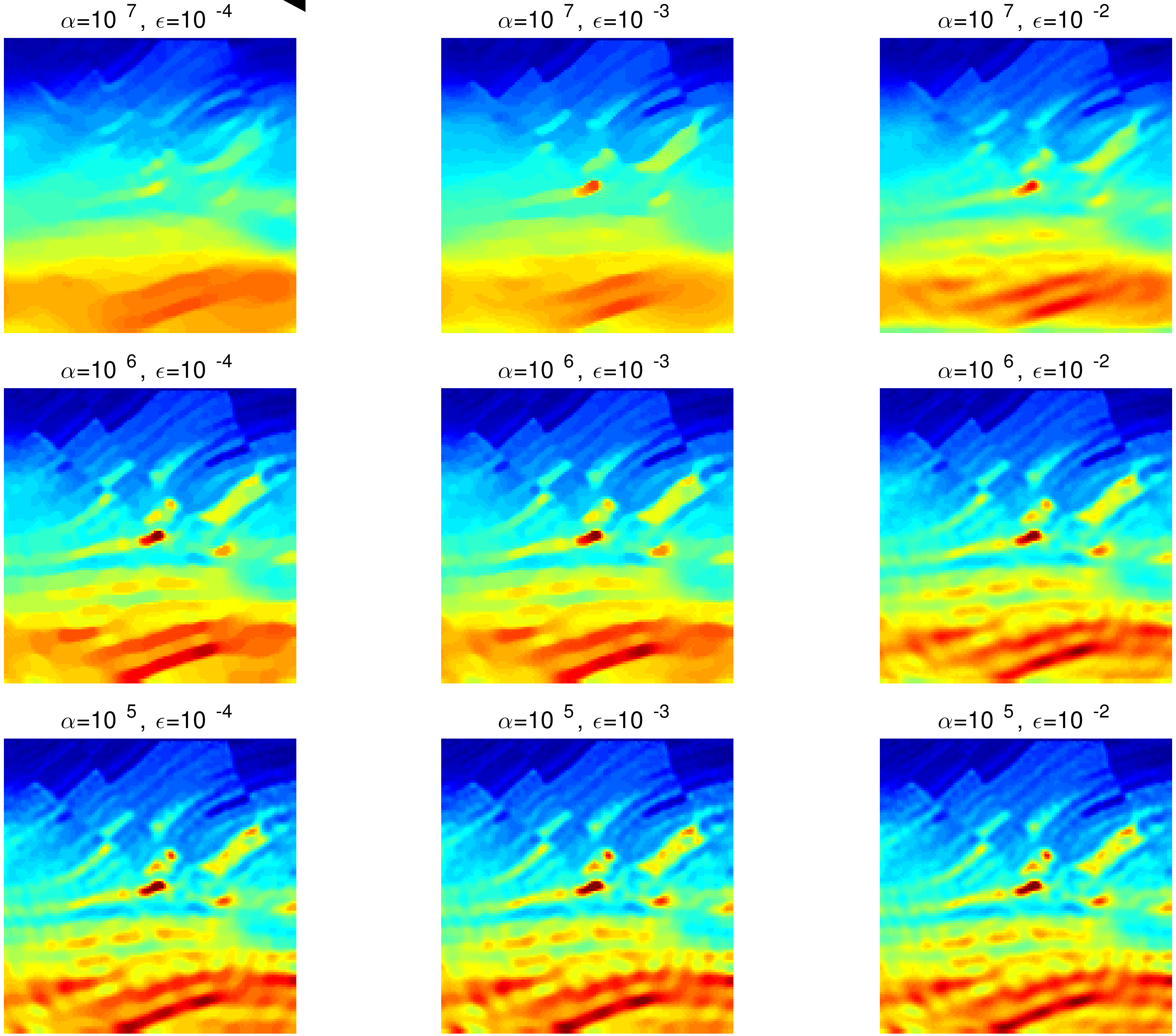
Numerical example 1



- FWI w/ smoothed TV-penalty & box constraints
- data w/ zero-mean random noise, $\|\text{noise}\|_2 / \|\text{signal}\|_2 = 0.25$
- starting model = smoothed true model
- frequency batches from 3 Hz to 10 Hz

increased "blockiness"

FWI results using smoothed TV for various α, ϵ combinations



Constrained formulation

Problem statement:

$$\underset{\mathbf{m}}{\text{minimize}}\, f(\mathbf{m}) \quad \text{subject to} \quad \mathbf{m} \in \text{Box and TV}(\mathbf{m}) \leq \tau$$

Our approach: solve this problem directly.

There are many ways to solve it.

Algorithm design – wish list

- application of constraints should not require additional expensive gradient & objective calculations
- updated models need to satisfy all constraints after each iteration
- arbitrary number of constraints should be handled as long as their intersection is non-empty
- manual tuning of parameters should be limited to bare minimum
- constraints should work w/ black-box gradients & objectives

Nested optimization strategy

Constrained optimization:

$$\underset{\mathbf{m}}{\text{minimize}}\, f(\mathbf{m}) \quad \text{subject to} \quad \mathbf{m} \in \mathcal{C} = \bigcap_{i=1}^{p} \mathcal{C}_i$$

via 3 levels of nested optimization:

1. Projected gradients = expensive step
2. Dykstra's algorithm*
3. Projection onto each set separately (closed form or w/ ADMM)

* parameter free

Projected gradients

Algorithm:

$$\mathbf{m}_{k+1} = \mathcal{P}_{\mathcal{C}}(\mathbf{m}_k - \nabla_{\mathbf{m}} f(\mathbf{m}_k))$$

Constrained formulation

Algorithm:

$$\mathbf{m}_{k+1} = \underbrace{\mathcal{P}_{\mathcal{C}}}_{\text{projection onto constraint set}}\big(\underbrace{\mathbf{m}_k - \nabla_{\mathbf{m}} f(\mathbf{m}_k)}_{\text{gradient step (proposed model)}}\big)$$

$$\mathcal{P}_{\mathcal{C}}(\mathbf{m}) = \arg\min_{\mathbf{x}} \|\mathbf{x} - \mathbf{m}\|_2 \quad \text{s.t.} \quad \mathbf{x} \in \underbrace{\bigcap_{i=1}^{p} \mathcal{C}_i}_{\text{intersection of constraint sets}}.$$

Constrained formulation

Projection onto intersection:

$$\mathcal{P}_{\mathcal{C}}(\mathbf{m}) = \underset{\mathbf{x}}{\arg\min} \, \|\mathbf{x} - \mathbf{m}\|_2 \quad \text{s.t.} \quad \mathbf{x} \in \bigcap_{i=1}^{p} \mathcal{C}_i.$$

Typically no closed-form solution -> use Dykstra's algorithm.

Requires:

- projections onto each set separately
- vector additions

Dykstra splitting

Toy example:

find projection onto intersection of circle & square

Algorithm 1 Dykstra.

$x_0 = \mathbf{m},\ p_0 = 0,\ q_0 = 0$

For $k = 0, 1, \ldots$

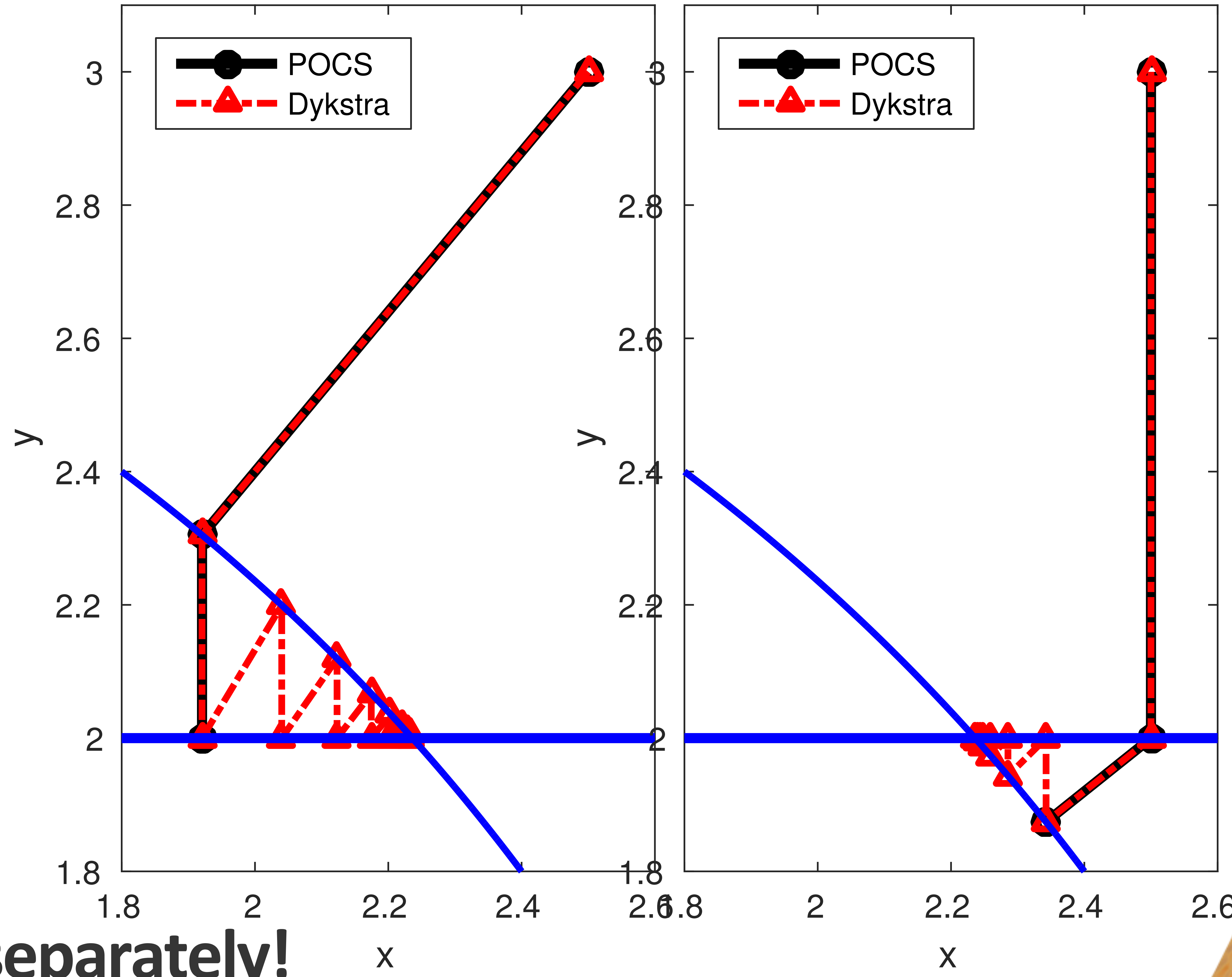
$\longrightarrow$ $\boxed{y_k = \mathscr{P}_{C_1}(x_k + p_k)}$

$p_{k+1} = x_k + p_k - y_k$

$\longrightarrow$ $\boxed{x_{k+1} = \mathscr{P}_{C_2}(y_k + q_k)}$

$q_{k+1} = y_k + q_k - x_{k+1}$

End



Only needs projections onto each set separately!

Individual projections

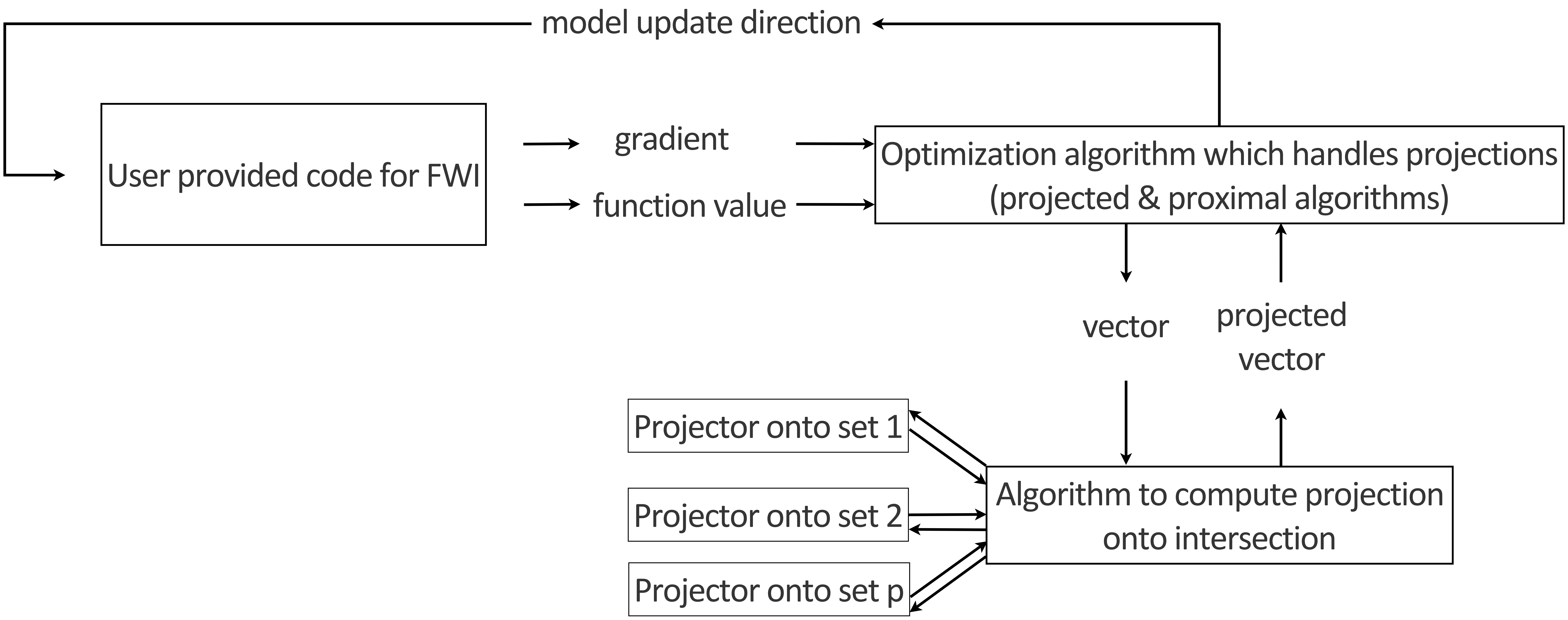
Projection onto bounds:

$$\mathcal{P}_{\mathcal{C}_1}(m_i) = \text{median}\{l_i, m_i, u_i\} \quad \forall i$$

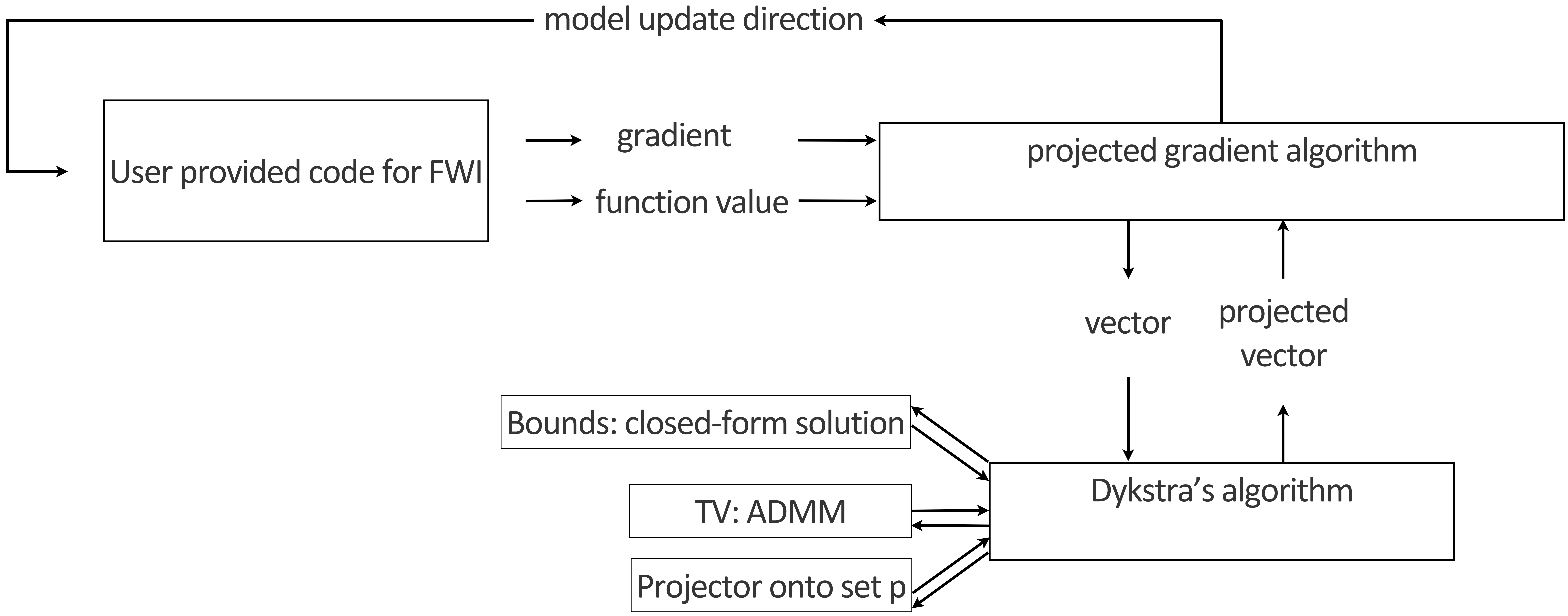
Projection onto TV-ball: split variables, then use ADMM for **x** & **z**

$$\begin{aligned}
\mathcal{P}_{\mathcal{C}}(\mathbf{m}) &= \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x}-\mathbf{m}\|_2^2 \quad \text{s.t} \quad \mathbf{x} \in \mathcal{C} \\
&= \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x}-\mathbf{m}\|_2^2 \quad \text{s.t} \quad \|\nabla \mathbf{x}\|_1 \leq \sigma \\
&= \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x}-\mathbf{m}\|_2^2 \quad \text{s.t} \quad \|\mathbf{z}\|_1 \leq \sigma \,,\ \nabla \mathbf{x} = \mathbf{z}
\end{aligned}$$

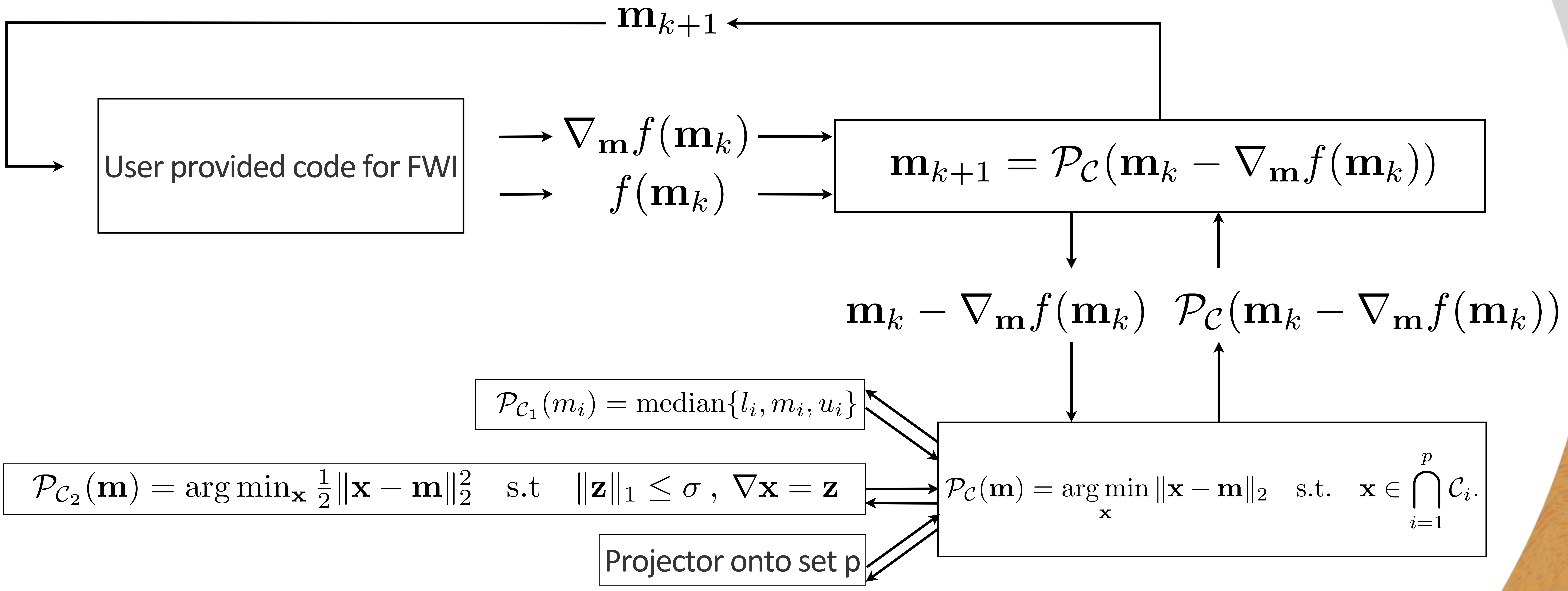
Workflow


model update direction
User provided code for FWI
gradient
function value
Optimization algorithm which handles projections
(projected & proximal algorithms)
vector
projected
vector
Projector onto set 1
Projector onto set 2
Projector onto set p
Algorithm to compute projection
onto intersection

Workflow


model update direction
User provided code for FWI
gradient
function value
projected gradient algorithm
vector
projected vector
Bounds: closed-form solution
TV: ADMM
Projector onto set p
Dykstra's algorithm

Workflow

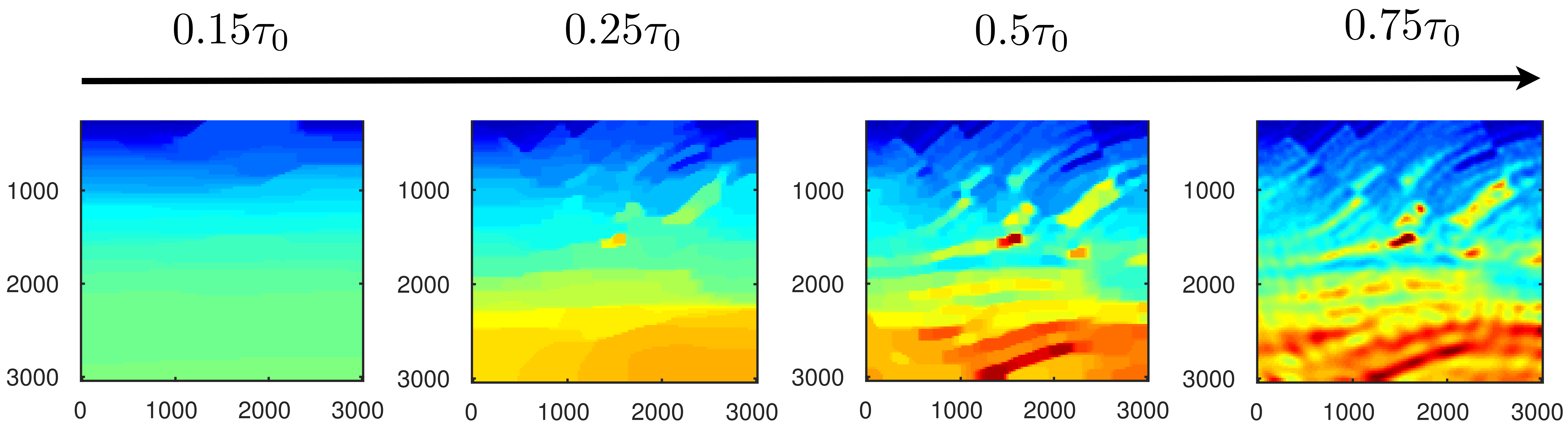


$\mathbf{m}_{k+1}$
User provided code for FWI
$\nabla_{\mathbf{m}} f(\mathbf{m}_k)$
$f(\mathbf{m}_k)$
$\mathbf{m}_{k+1} = \mathcal{P}_{\mathcal{C}}(\mathbf{m}_k - \nabla_{\mathbf{m}} f(\mathbf{m}_k))$
$\mathbf{m}_k - \nabla_{\mathbf{m}} f(\mathbf{m}_k)$
$\mathcal{P}_{\mathcal{C}}(\mathbf{m}_k - \nabla_{\mathbf{m}} f(\mathbf{m}_k))$
$\mathcal{P}_{\mathcal{C}_1}(m_i) = \text{median}\{l_i, m_i, u_i\}$
$\mathcal{P}_{\mathcal{C}_2}(\mathbf{m}) = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x} - \mathbf{m}\|_2^2 \quad \text{s.t} \quad \|\mathbf{z}\|_1 \leq \sigma \text{ , } \nabla\mathbf{x} = \mathbf{z}$
Projector onto set p
$\mathcal{P}_{\mathcal{C}}(\mathbf{m}) = \arg\min_{\mathbf{x}} \|\mathbf{x} - \mathbf{m}\|_2 \quad \text{s.t.} \quad \mathbf{x} \in \bigcap_{i=1}^{p} \mathcal{C}_i.$

Numerical example 1 – revisited

- FWI w/ TV-norm & bound constraints
- data w/ zero-mean random noise, $\|\text{noise}\|_2 / \|\text{signal}\|_2 = 0.25$
- starting model = smoothed true model
- frequency batches from 3Hz to 10 Hz

FWI w/ constraints



Penalties vs. constraints

Regularization w/ penalties:

- inversion results for various α, ϵ combinations behave unpredictably
- challenges proper parameter settings
- offers no guarantees of feasibility for each model iterate

Regularization w/ constraints:

- inversion results behave predictably for increasing τ
- edges are preserved for not too large τ
- inversion artifacts appear for too large τ

Suggests cooling technique w/ warm starts where τ is increased slowly...

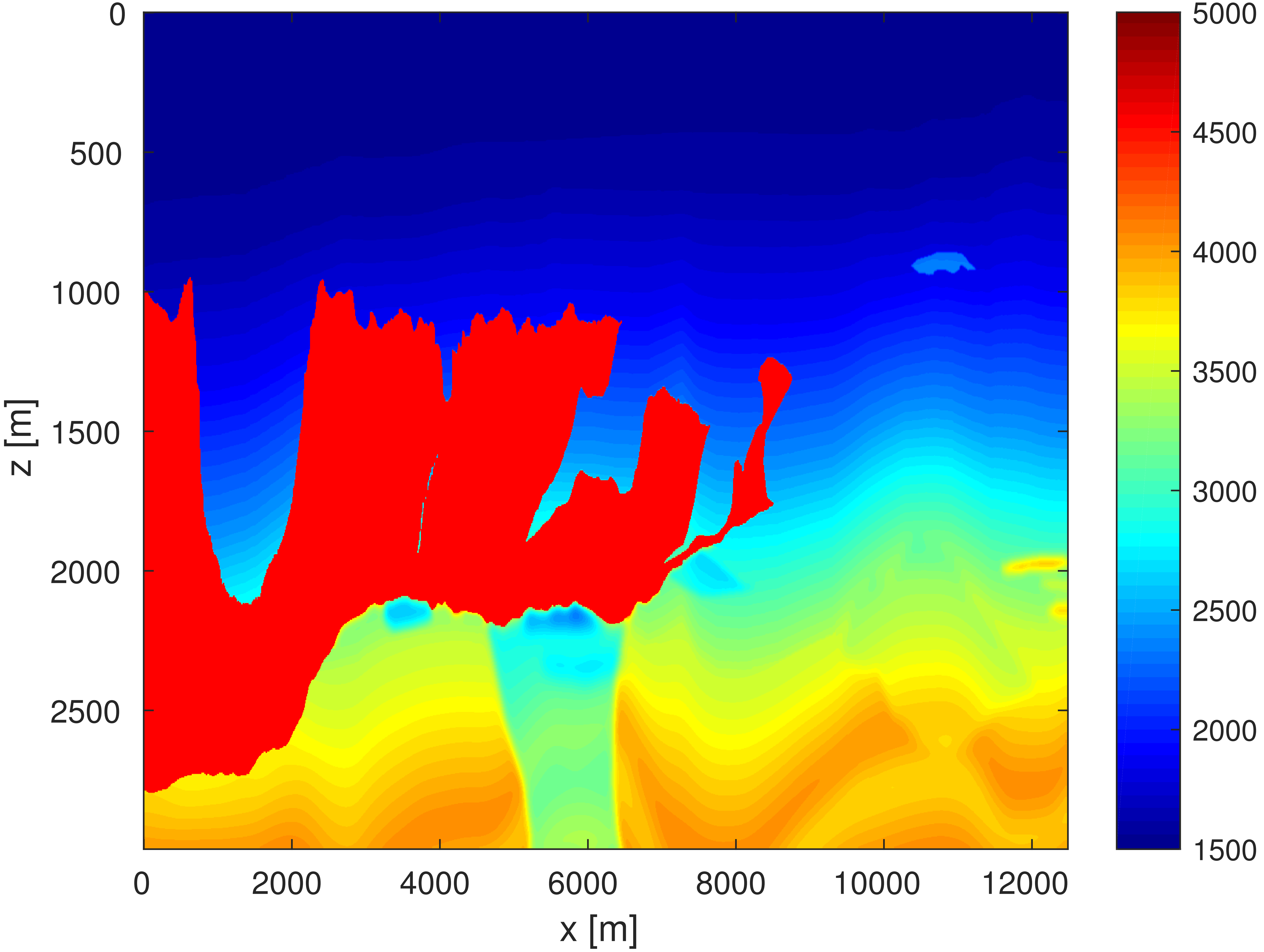
Case study. Improve delineation of salt for a good starting model but poor (6 dB) data...

Billette, F., and Brandsberg-Dahl, S., 2005

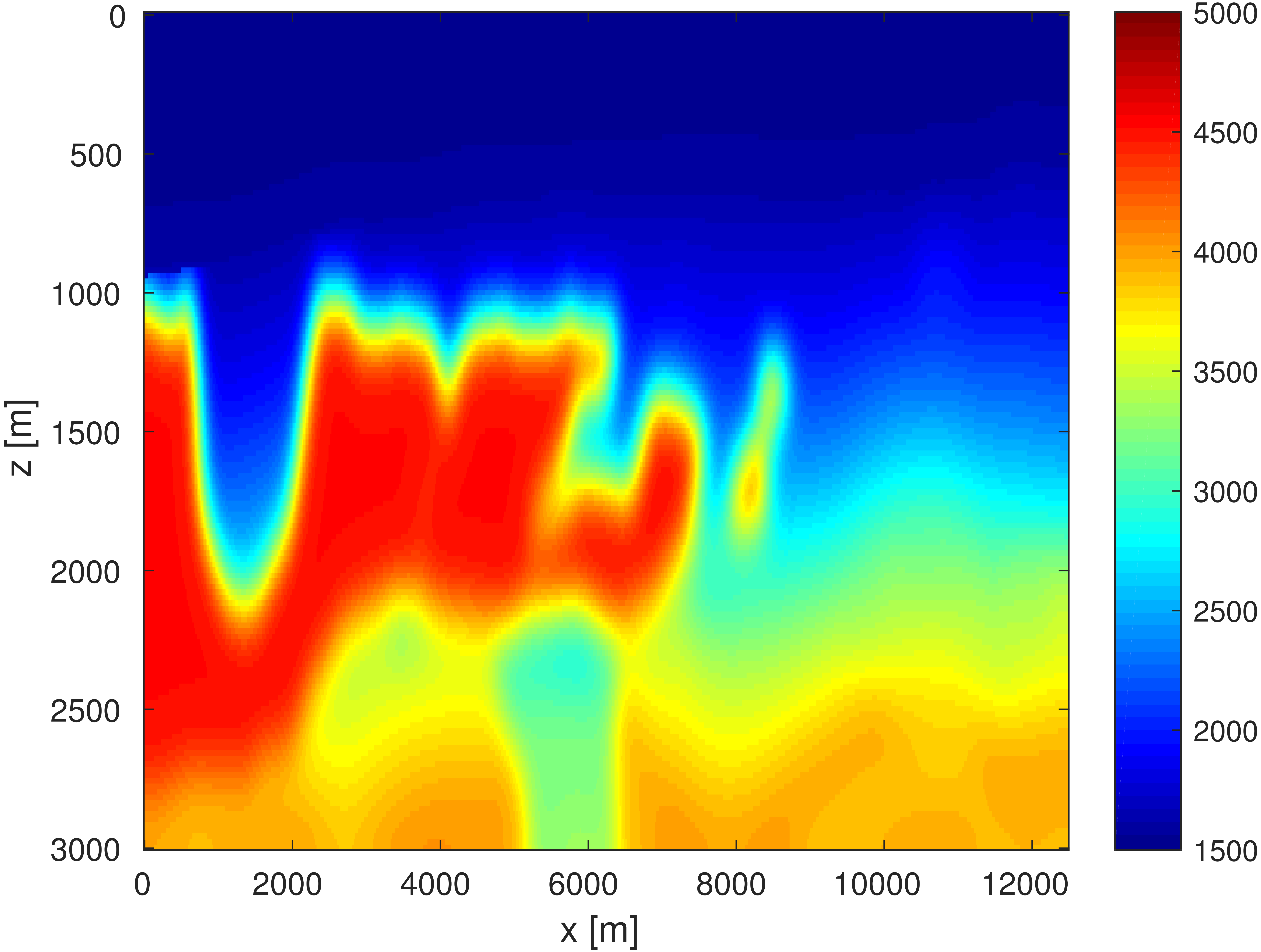
Reduced BP model – modelling parameters

- number of sources: 132; number of receivers: 311
- receiver spacing: 40m, source spacing: 80m, max offset 11.5 km
- grid size: 20 m
- known Ricker wavelet sources with 15Hz peak frequency
- data available starting at 3 Hz
- 8 simultaneous shots w/ Gaussian weights w/ redraws
- starting model = smoothed true model
- inversion crime but poor data $\|\text{noise}\|_2/\|\text{signal}\|_2 = 0.5$

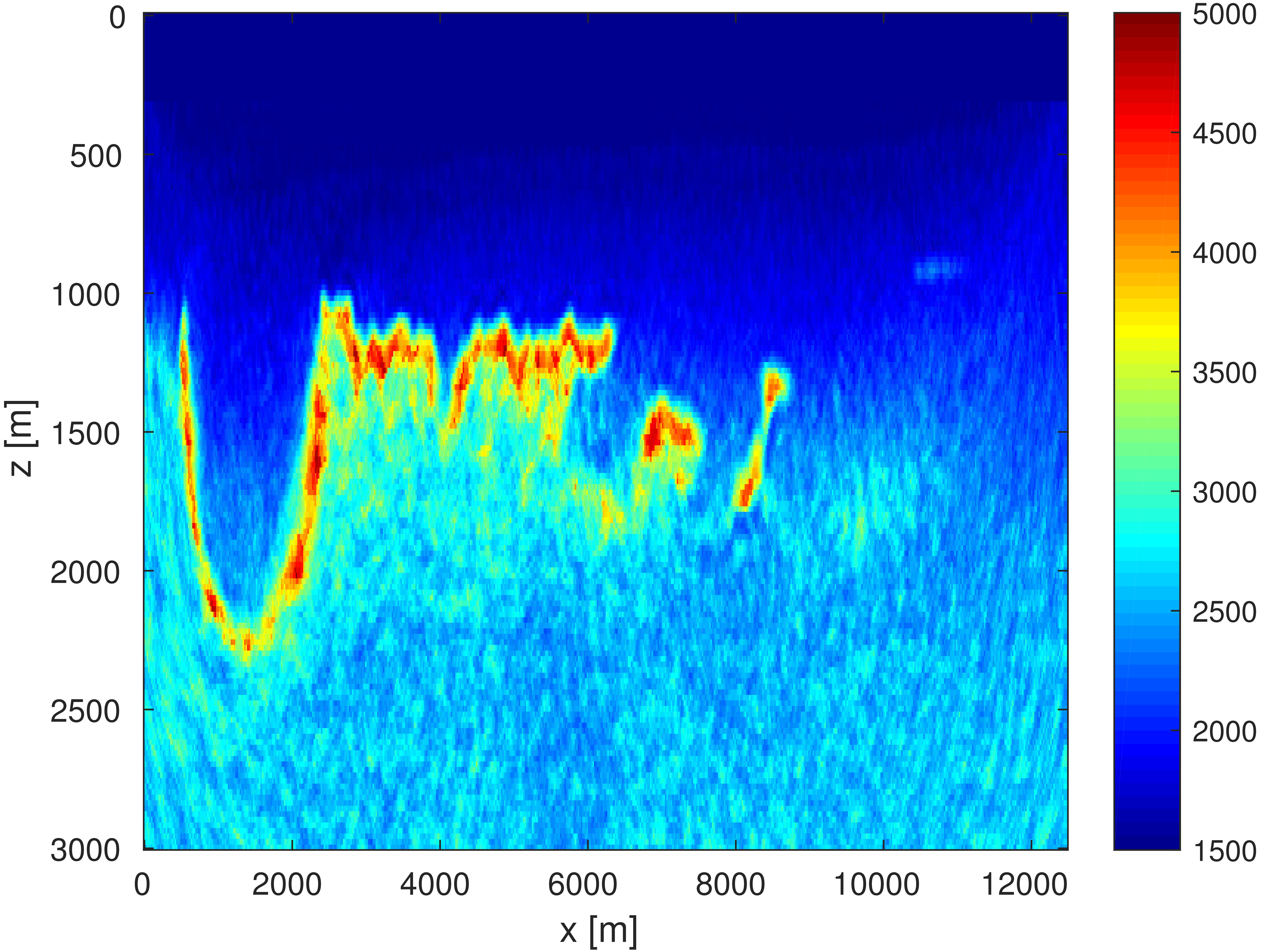
True velocity model – reduced by a factor of 2.5



Starting model

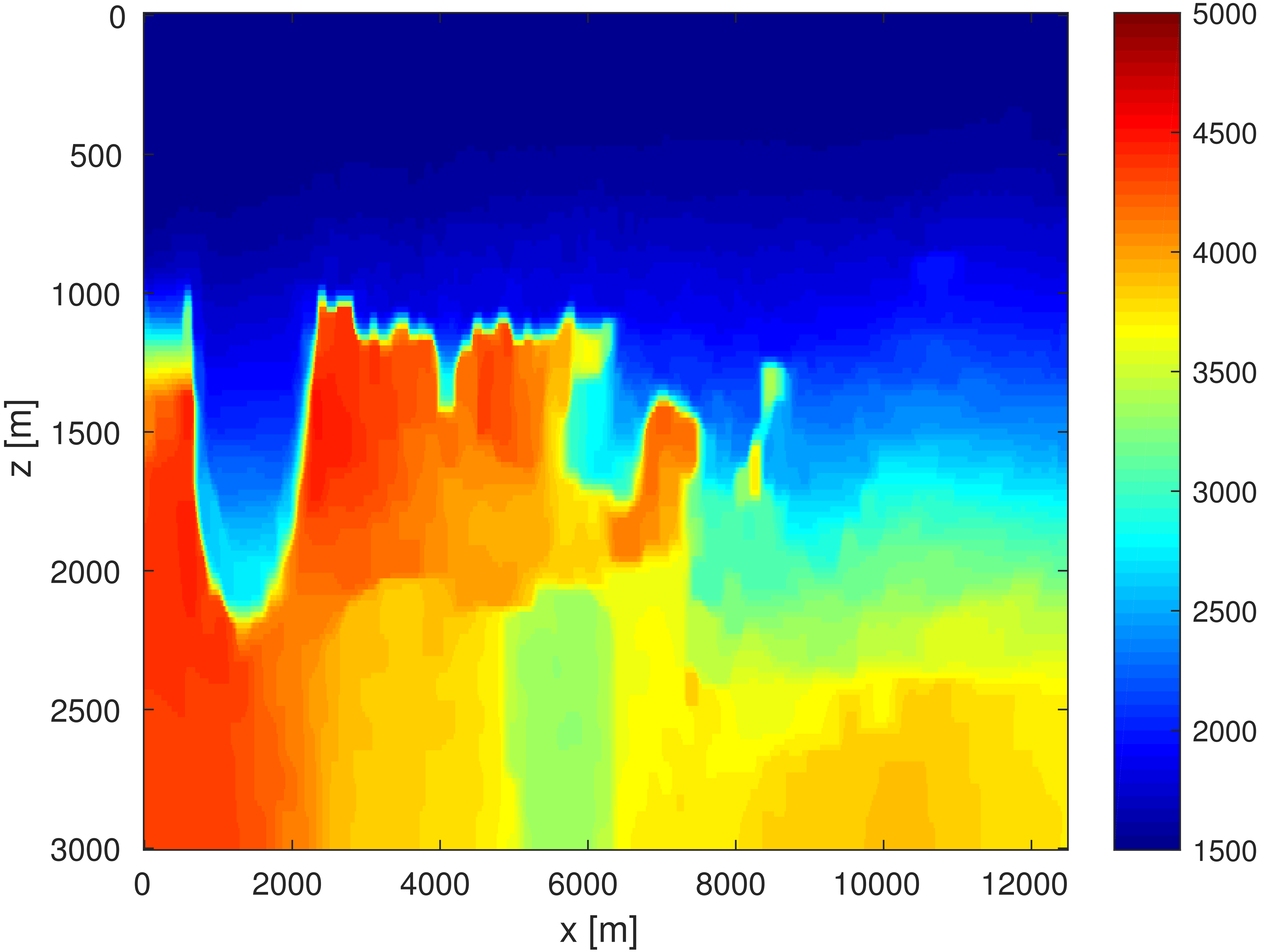


Adjoint-state w/ noisy data $\|\text{noise}\|_2/\|\text{signal}\|_2 = 0.5$



WRI /w TV-constraints

$$\|\text{noise}\|_2 / \|\text{signal}\|_2 = 0.5$$



Ernie Esser, Lluís Guasch, Felix J. Herrmann, and Mike Warner, "Constrained waveform inversion for automatic salt flooding", *The Leading Edge*, vol. 35, p. 235-239, 2016
Ernie Esser, Lluís Guasch, Tristan van Leeuwen, Aleksandr Y. Aravkin, and Felix J. Herrmann, "Total-variation regularization strategies in full-waveform inversion". 2016

SLIM

Heuristic

Multiple frequency cycles:

- warm starts
- increasingly relaxed TV constraints & fixed bound constraints
- starts w/ relaxed TV-norm of starting model

Extend search space:

- make sure data is fitted
- optimize over model & (source) wavefields
- jointly fit data & wave equation (physics)
- use noise level to automatically select trade-off data & PDE (physics) fit

Wavefield Reconstruction Inversion – gradient

Tristan van Leeuwen and Felix J. Herrmann, "A penalty method for PDE-constrained optimization in inverse problems", *Inverse Problems*, vol. 32, p. 015007, 2015.
Tristan van Leeuwen and Felix J. Herrmann, "Mitigating local minima in full-waveform inversion by expanding the search space", *Geophysical Journal International*, vol. 195, p. 661-667, 2013

WRI method:

for each source i

$$\text{solve} \quad \begin{pmatrix} P_i \\ \lambda A_i(\mathbf{m}) \end{pmatrix} \mathbf{u}_{\lambda,i} \approx \begin{pmatrix} \mathbf{d}_i \\ \lambda \mathbf{q}_i \end{pmatrix}$$

$$\mathbf{g} = \mathbf{g} + \lambda^2\omega^2 \mathrm{diag}(\bar{\mathbf{u}}_{i,\lambda})^*(A(\mathbf{m})\bar{\mathbf{u}}_{i,\lambda} - \mathbf{q}_i)$$

end

correlation proxy wavefield & PDE residual

Adjoint-state method:

for each source i

$$\text{solve} \quad A(\mathbf{m})\mathbf{u}_i = \mathbf{q}_i$$

$$\text{solve} \quad A(\mathbf{m})^*\mathbf{v}_i = P_i^*(P_i\mathbf{u}_i - \mathbf{d}_i)$$

$$\mathbf{g} = \mathbf{g} + \omega^2 \mathrm{diag}(\mathbf{u}_i)^*\mathbf{v}_i$$

end

correlation wavefield & data residual

Patent application WO2014172787 – A PENALTY METHOD FOR PDE-CONSTRAINED OPTIMIZATION pending

BP model – inversion parameters

Optimization specs:

- spectral-projected gradients
- non-monotone linesearch w/ window size of 5
- max 8 DYKSTRA iterations

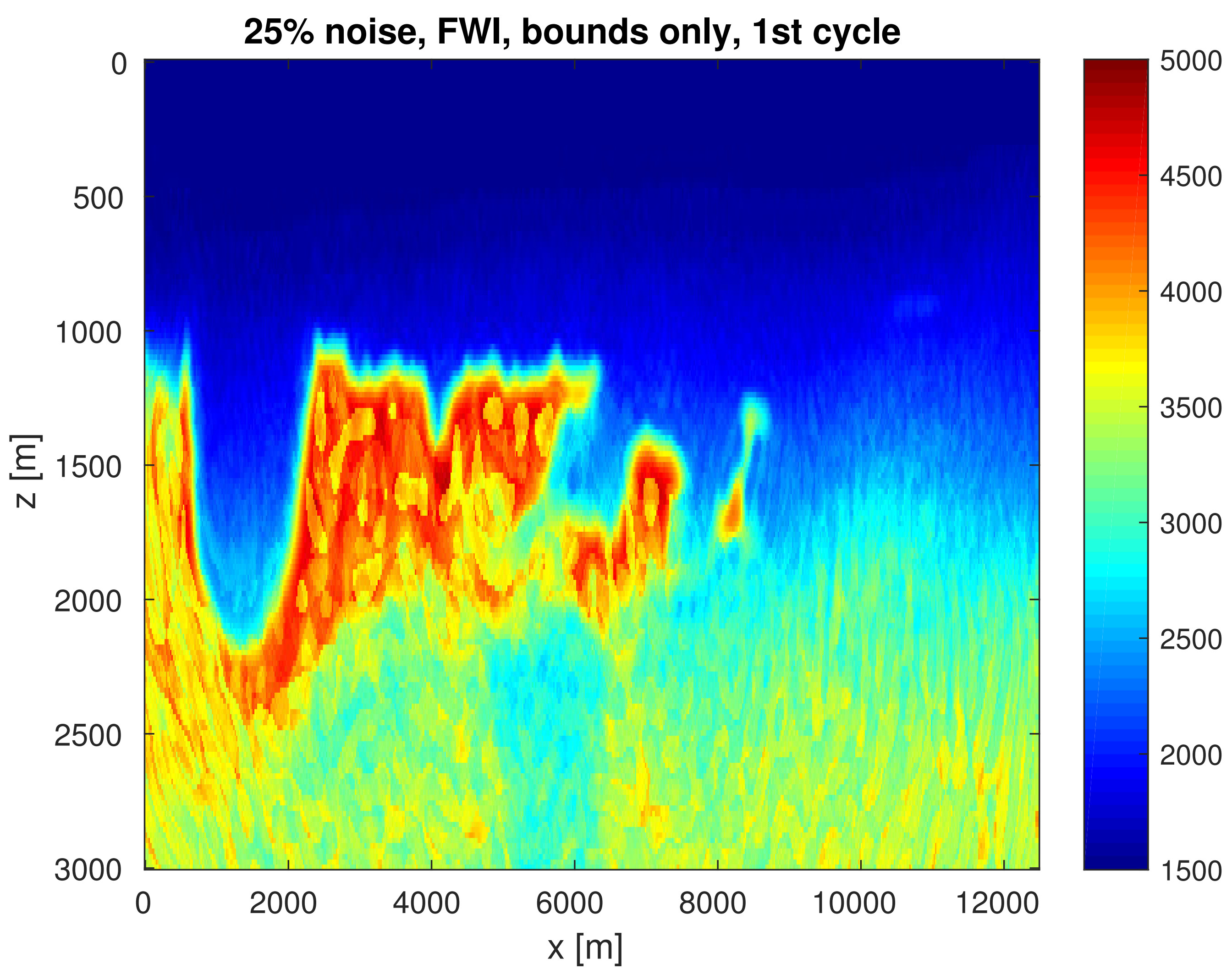
Constraint specs:

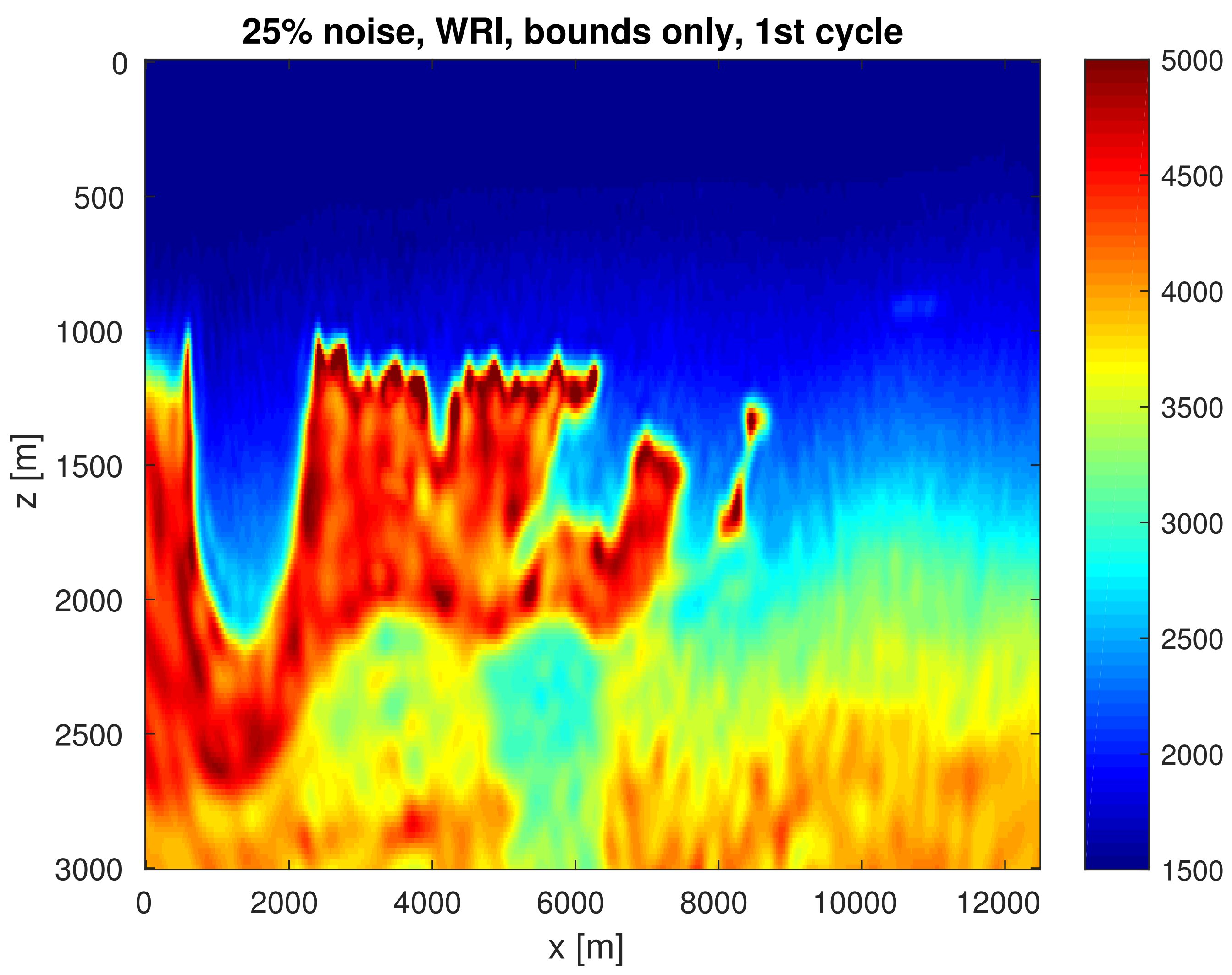
- frequency continuation 3–9 Hz in consecutive batches of 2
- 3 warm started frequency sweeps w/ $\tau^{l+1} = 1.25\tau^{l}$ & $\tau^{0} = 1.00 \times \mathrm{TV}(\mathbf{m}_0)$
- anisotropic TV

1st cycle cycle

$$\|\text{noise}\|_2 / \|\text{signal}\|_2 = 0.25$$

bounds only

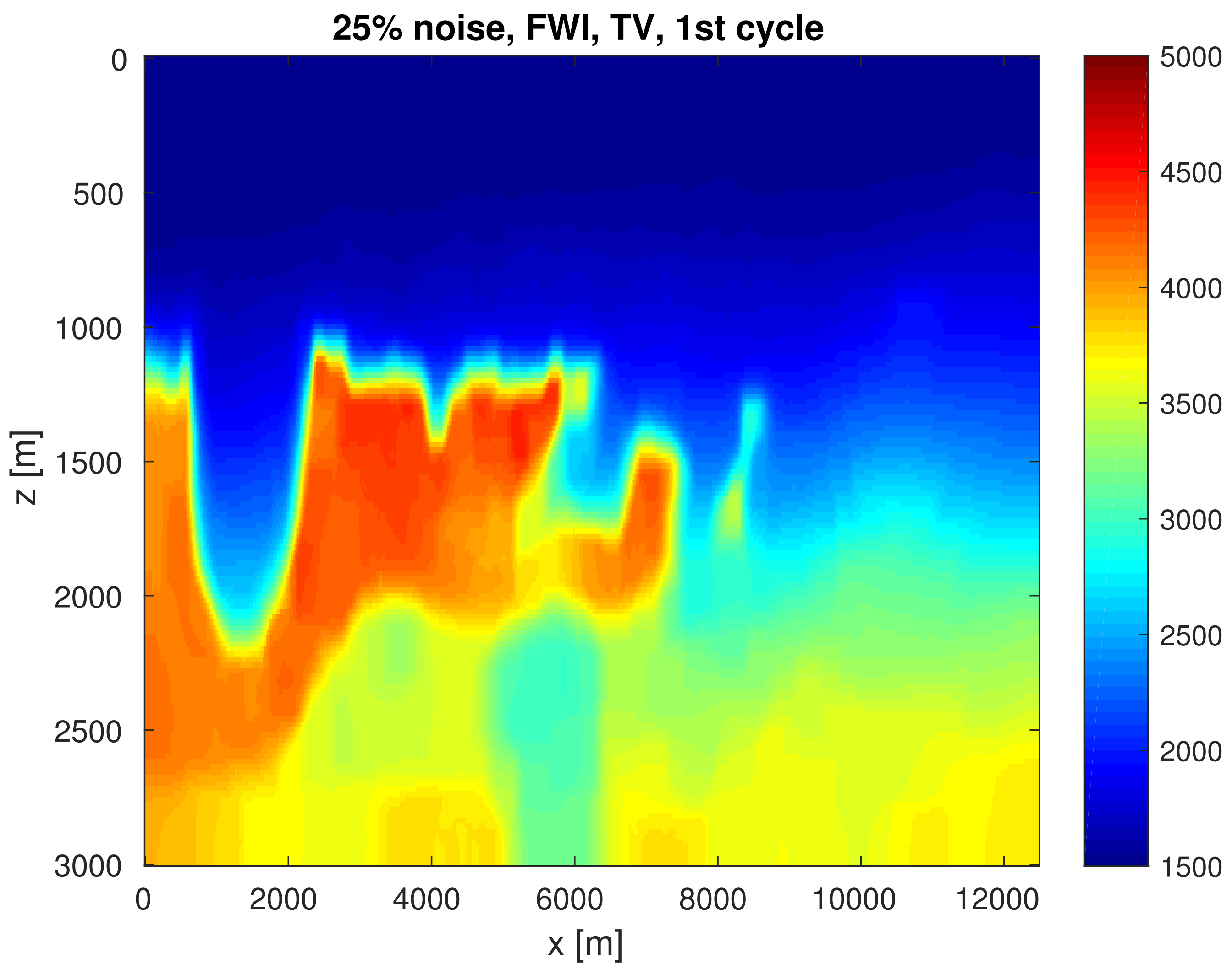


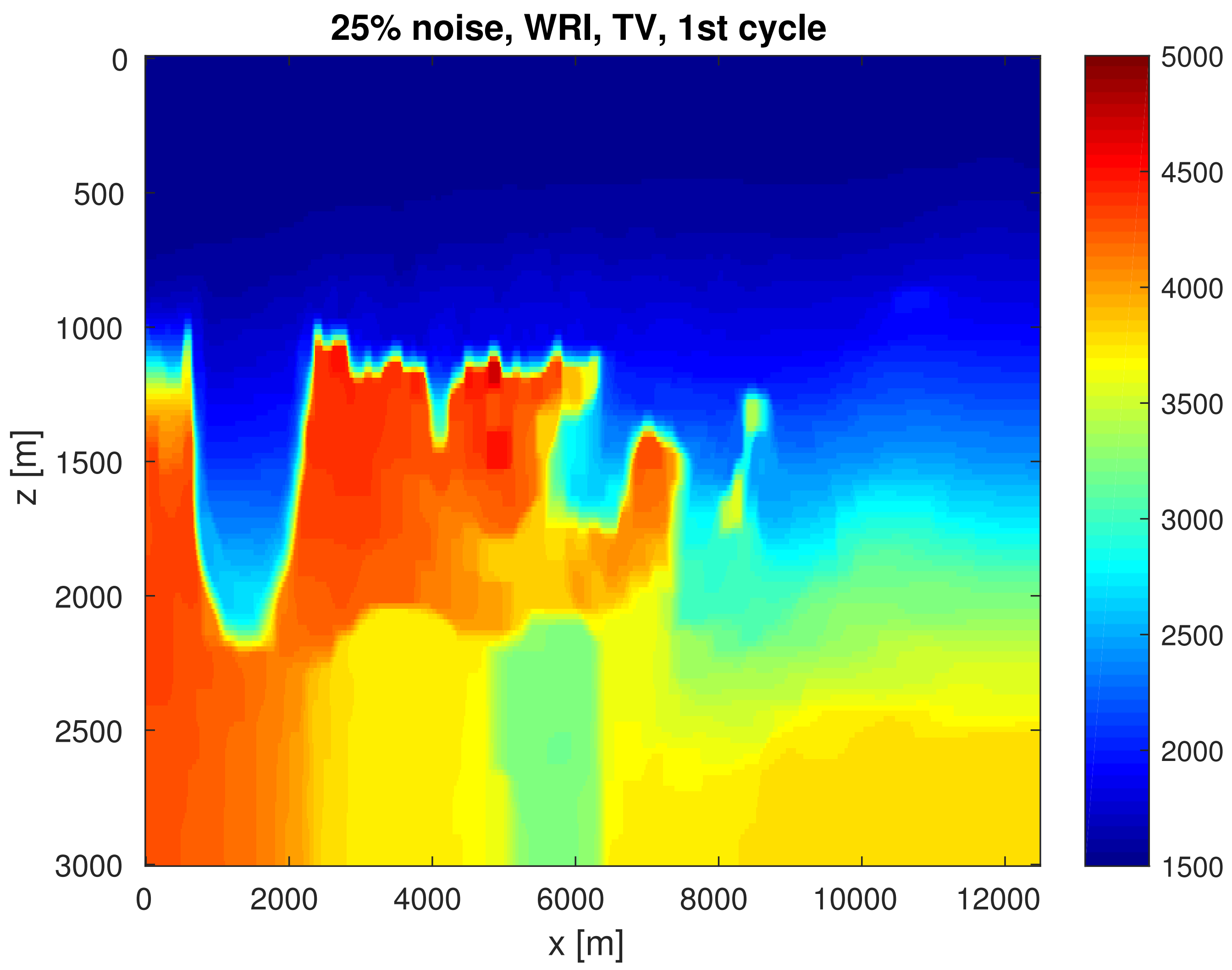


FWI

WRI

bounds & TV

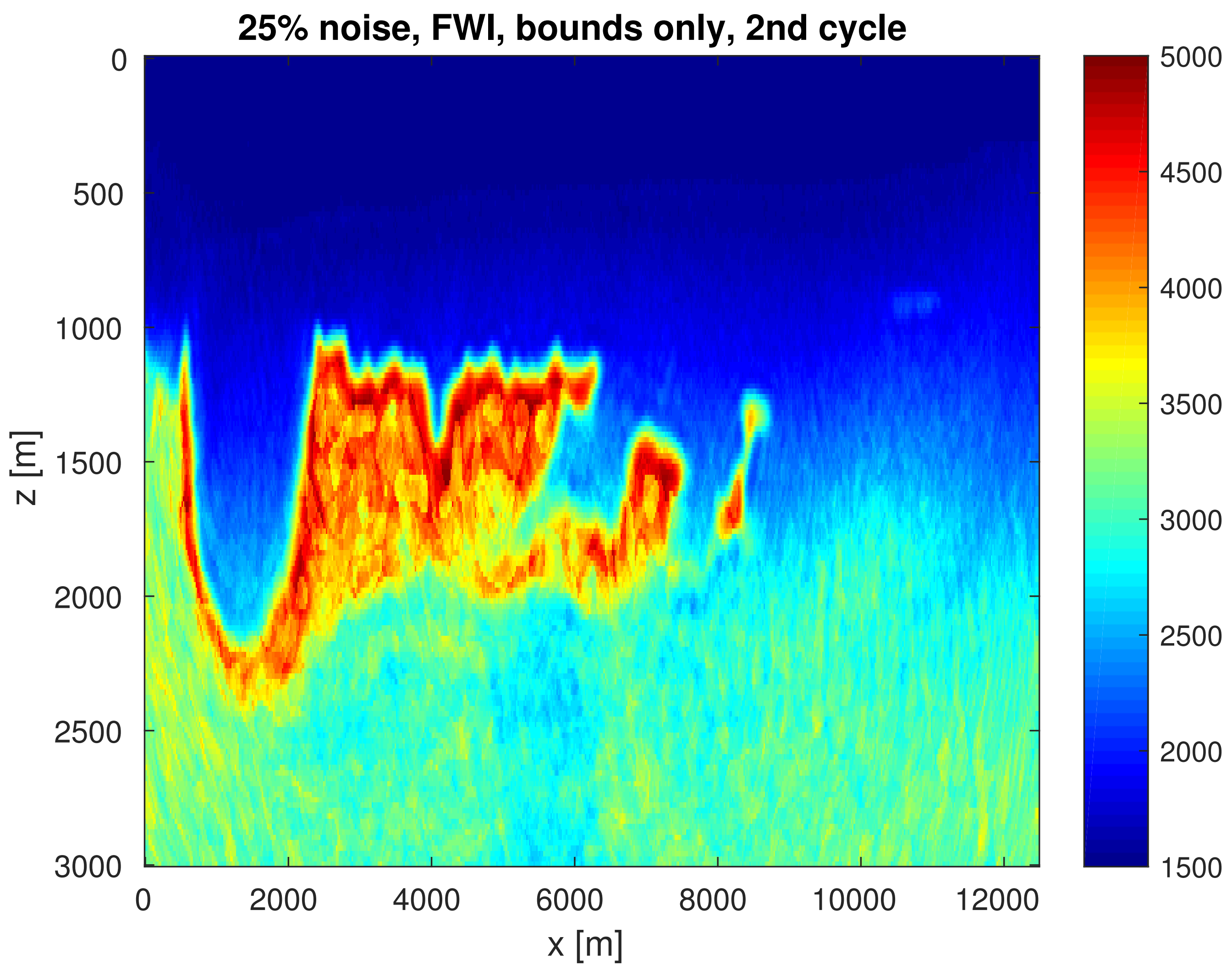


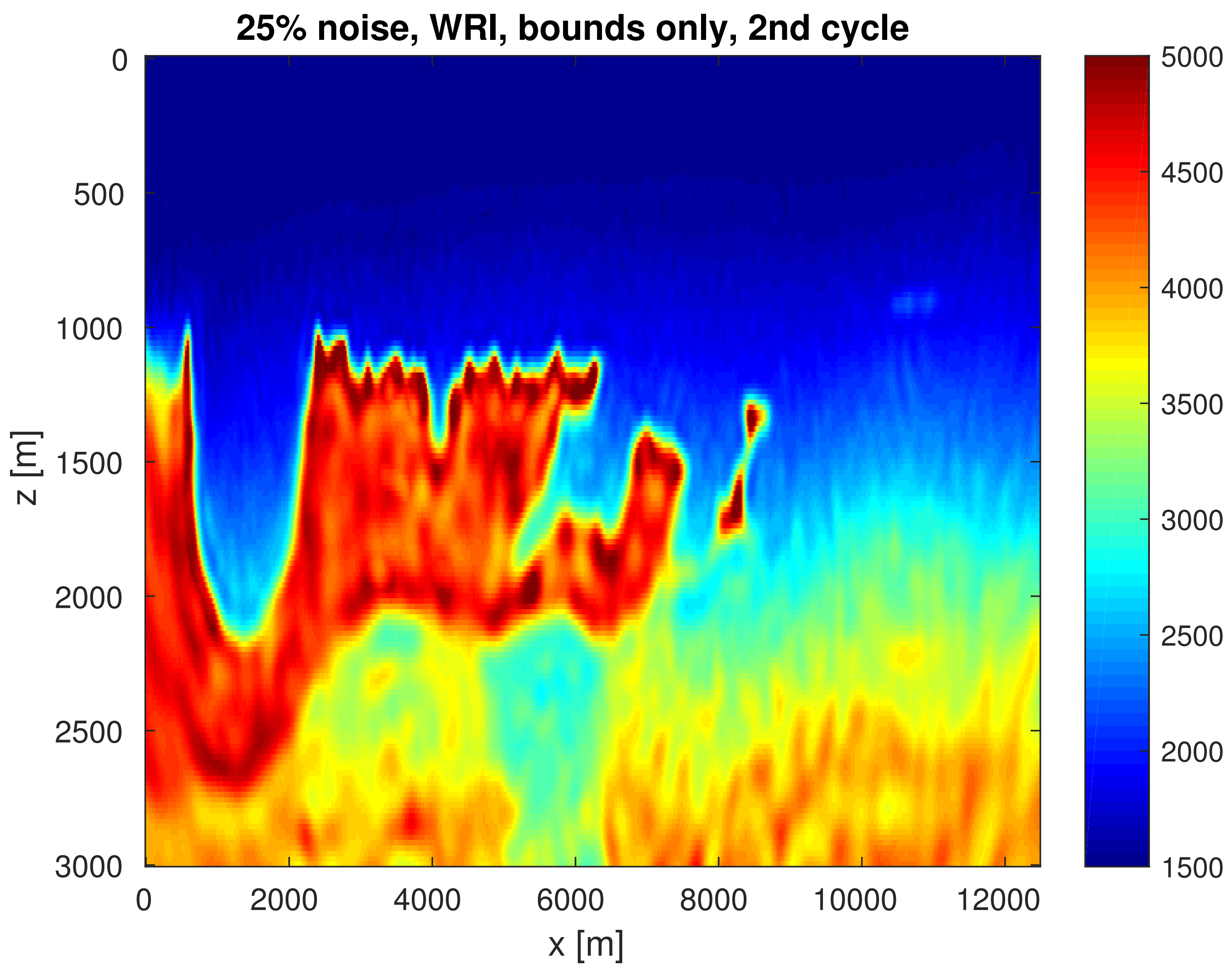


2nd cycle

$$\|\text{noise}\|_2 / \|\text{signal}\|_2 = 0.25$$

bounds only

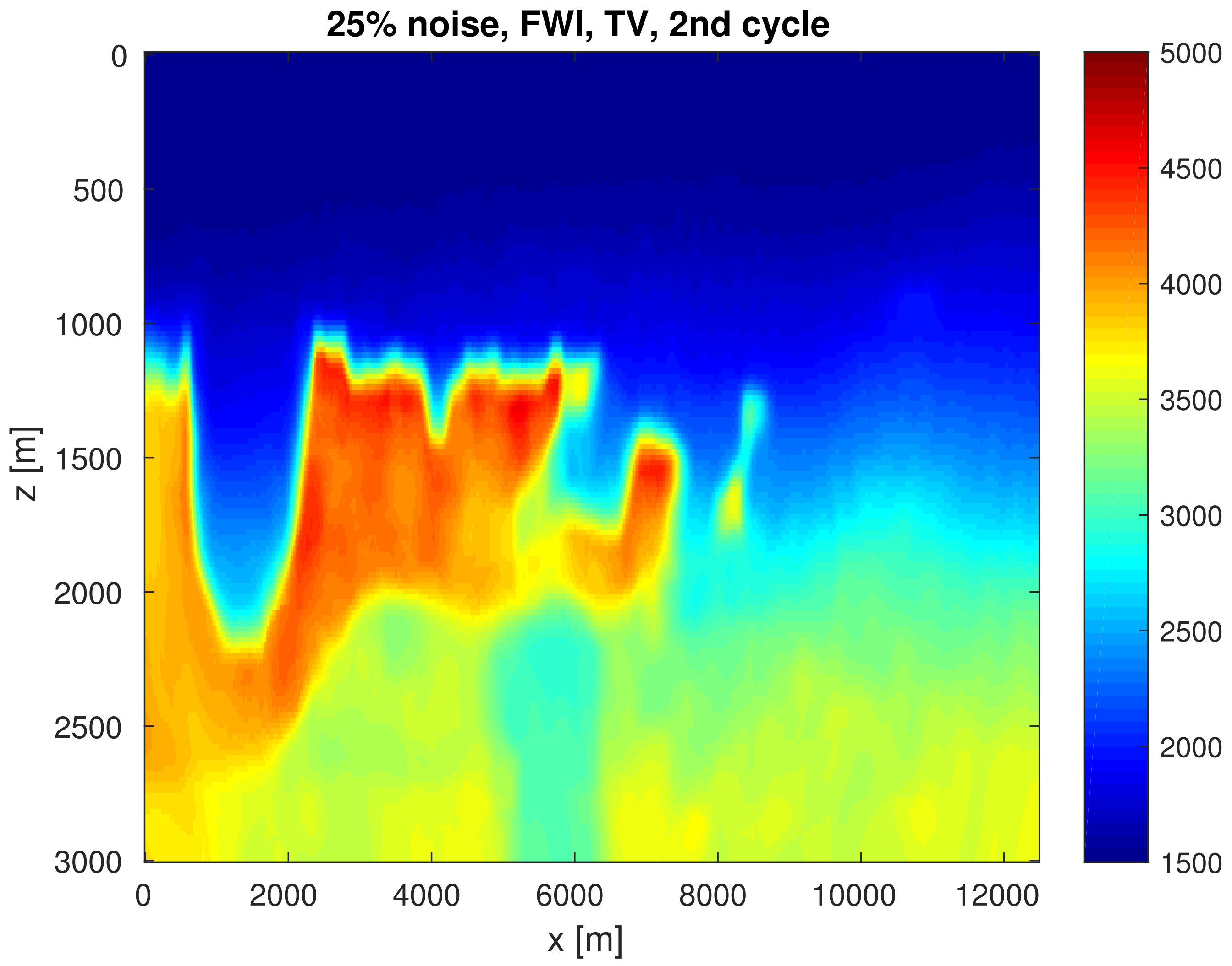


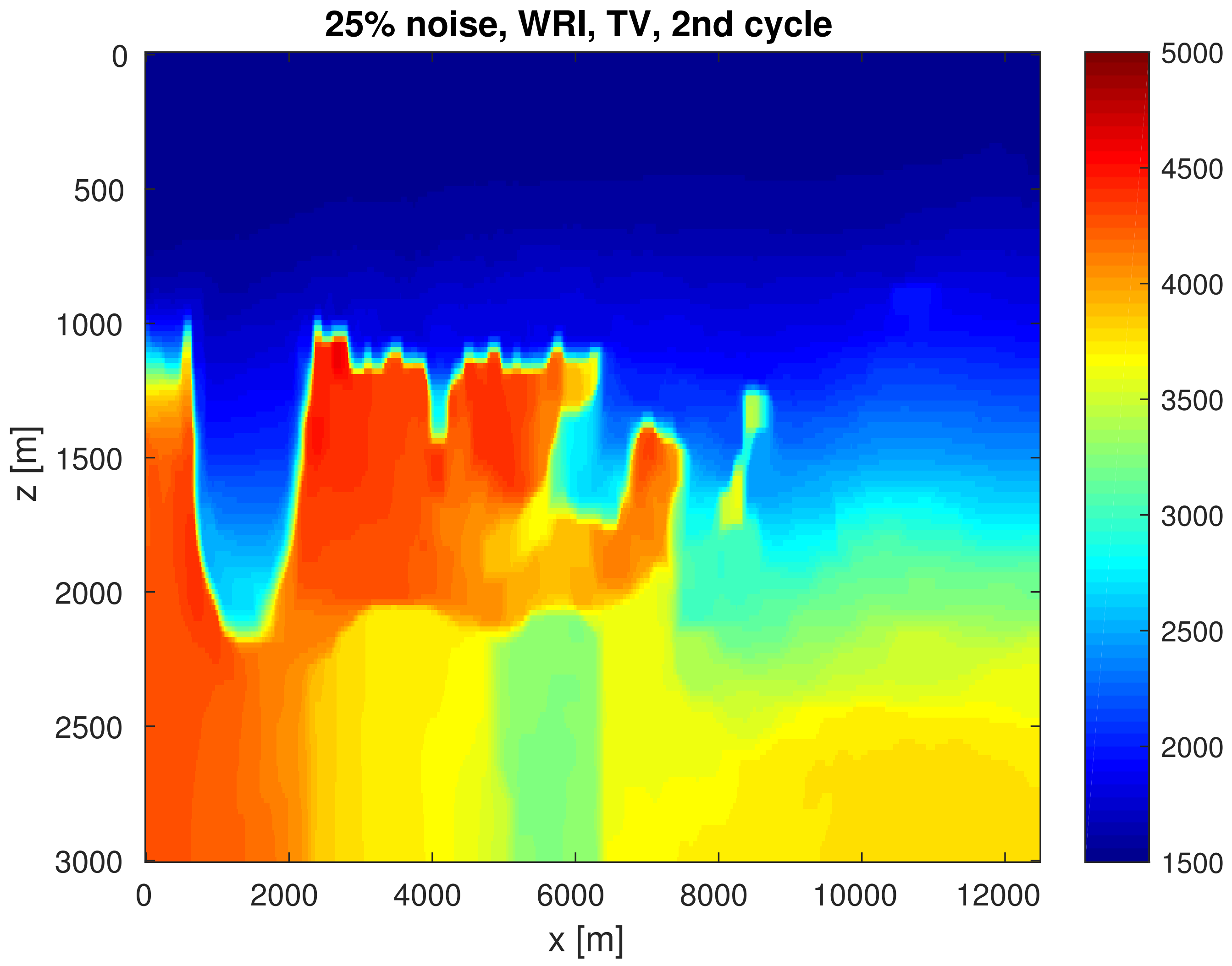


FWI

WRI

bounds & TV

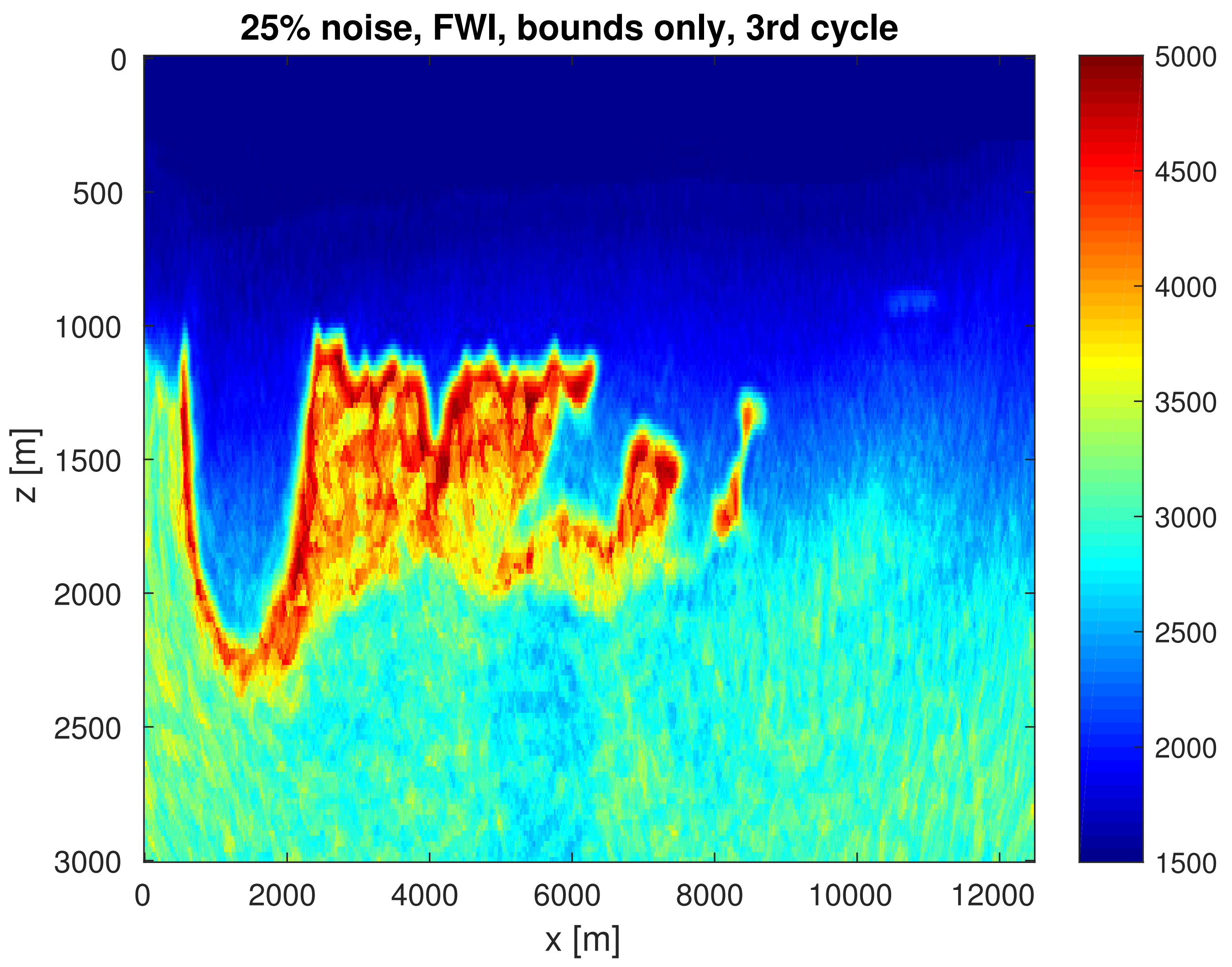


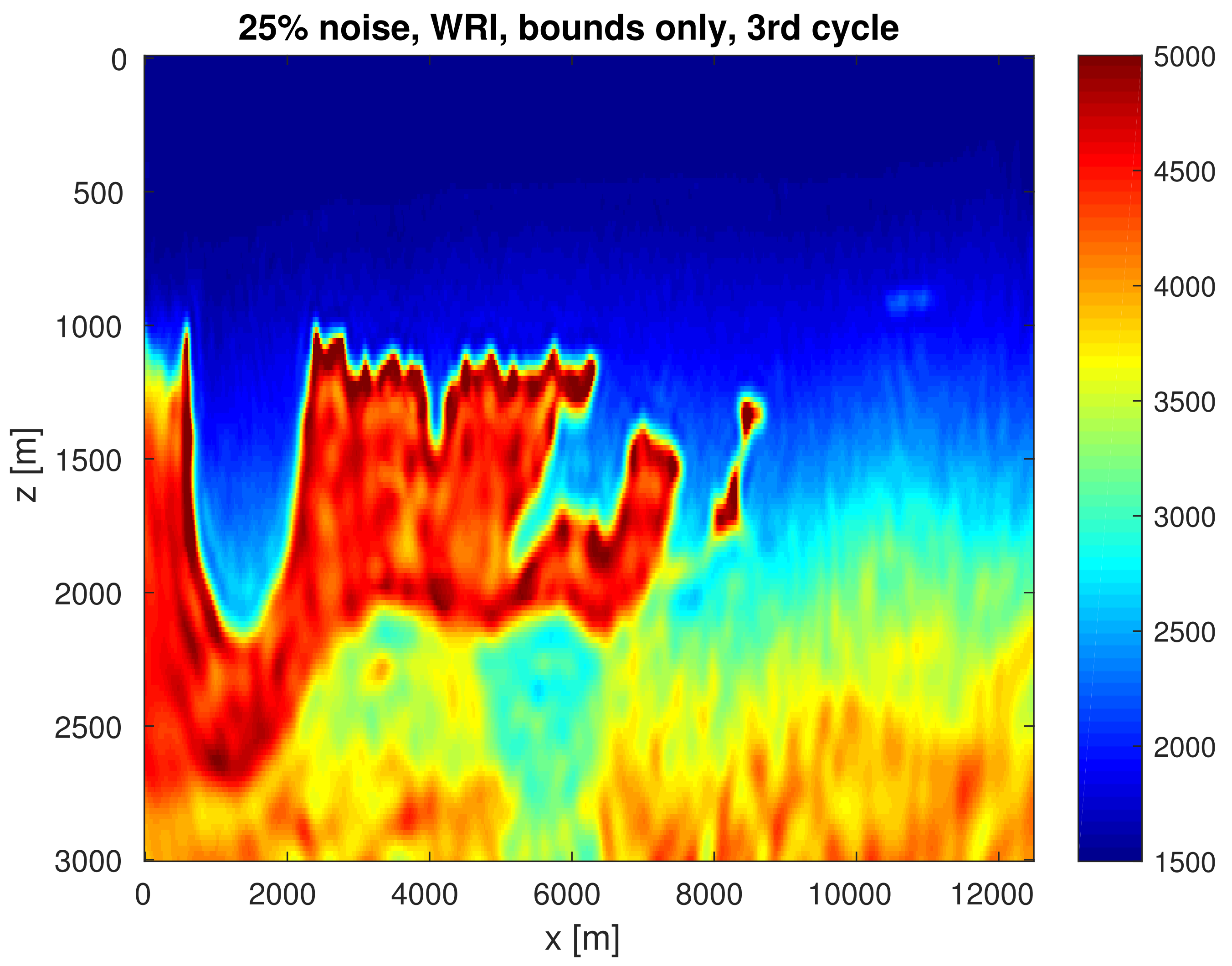


3rd cycle

$$\|\text{noise}\|_2 / \|\text{signal}\|_2 = 0.25$$

bounds only

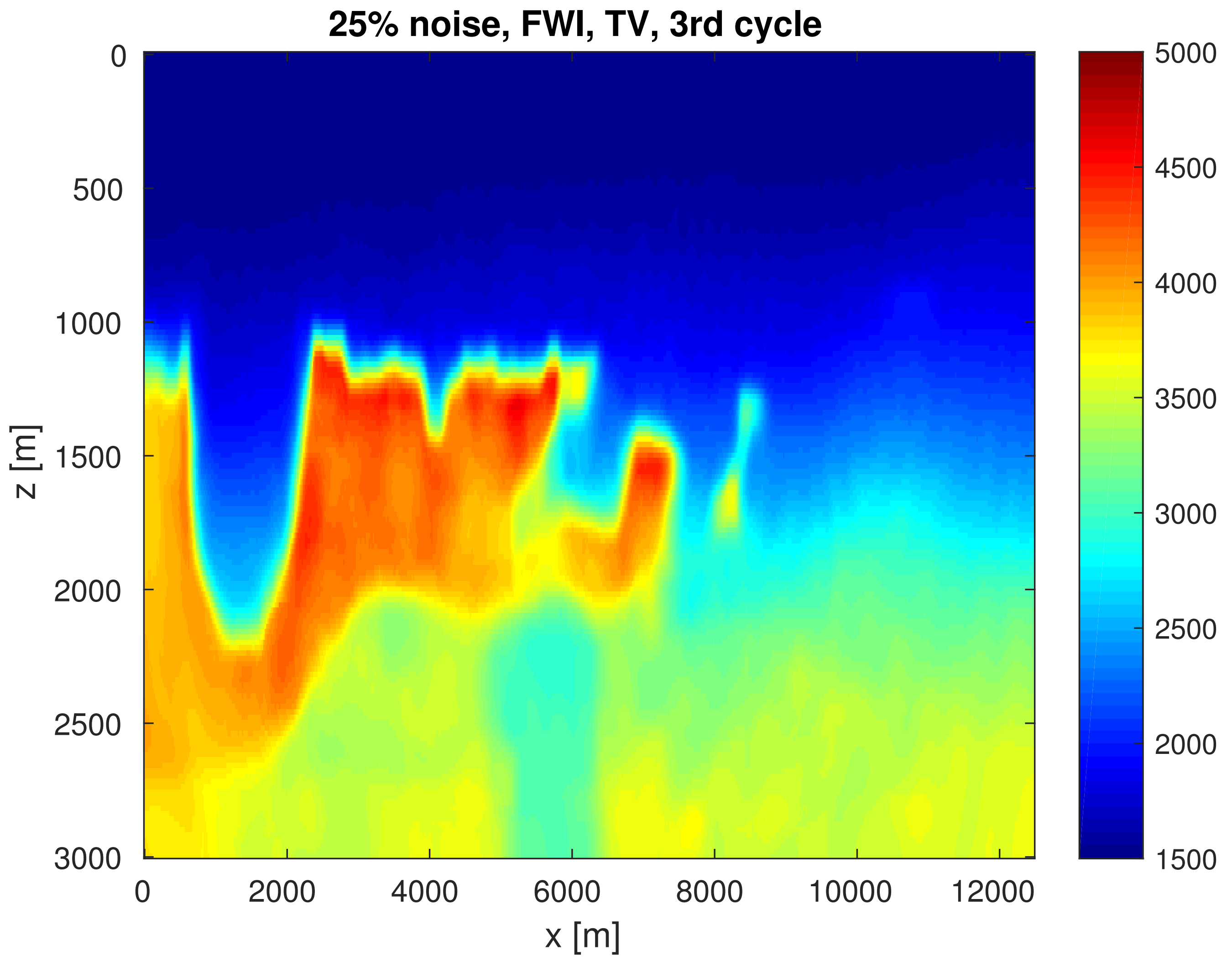


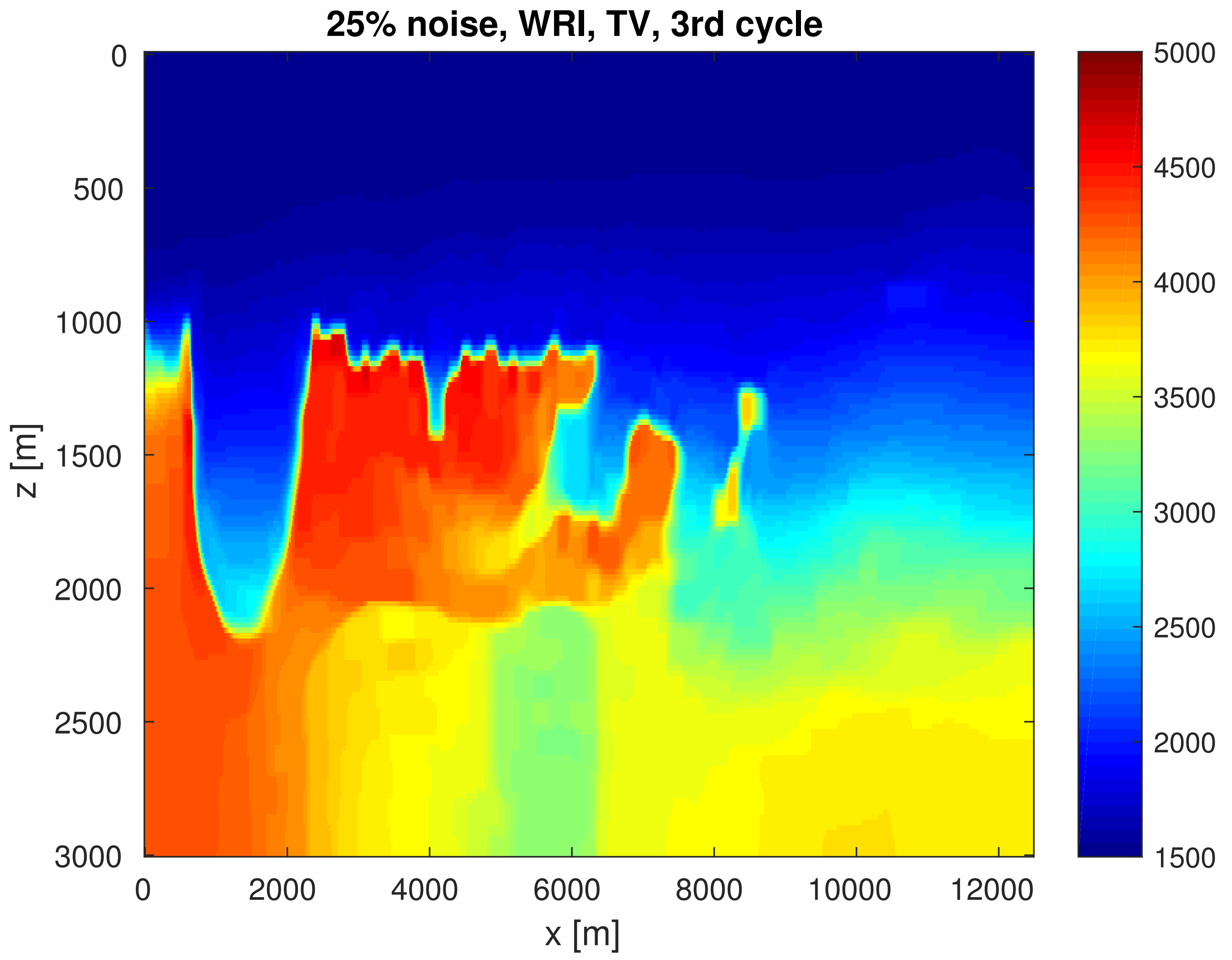


FWI

WRI

bounds & TV

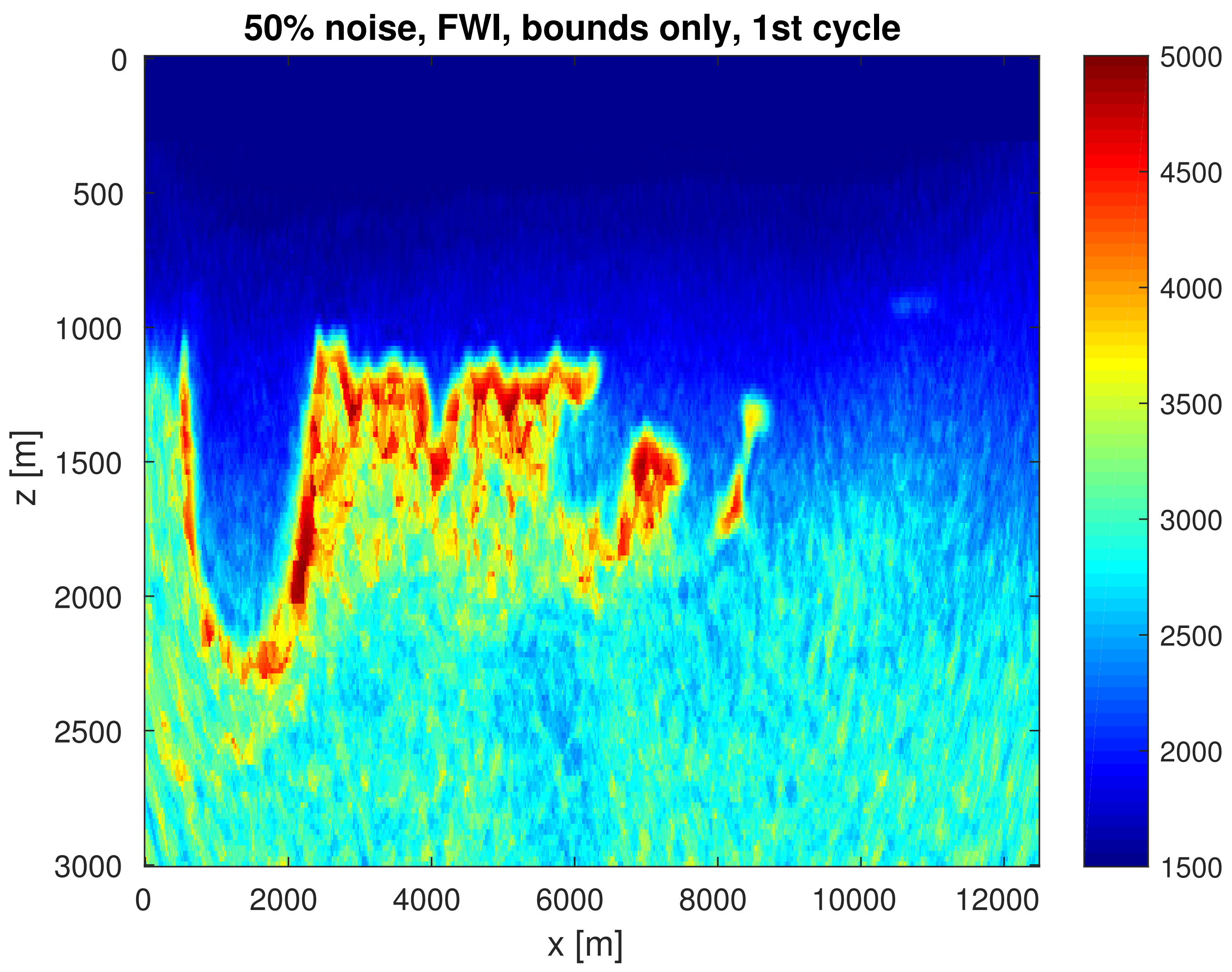


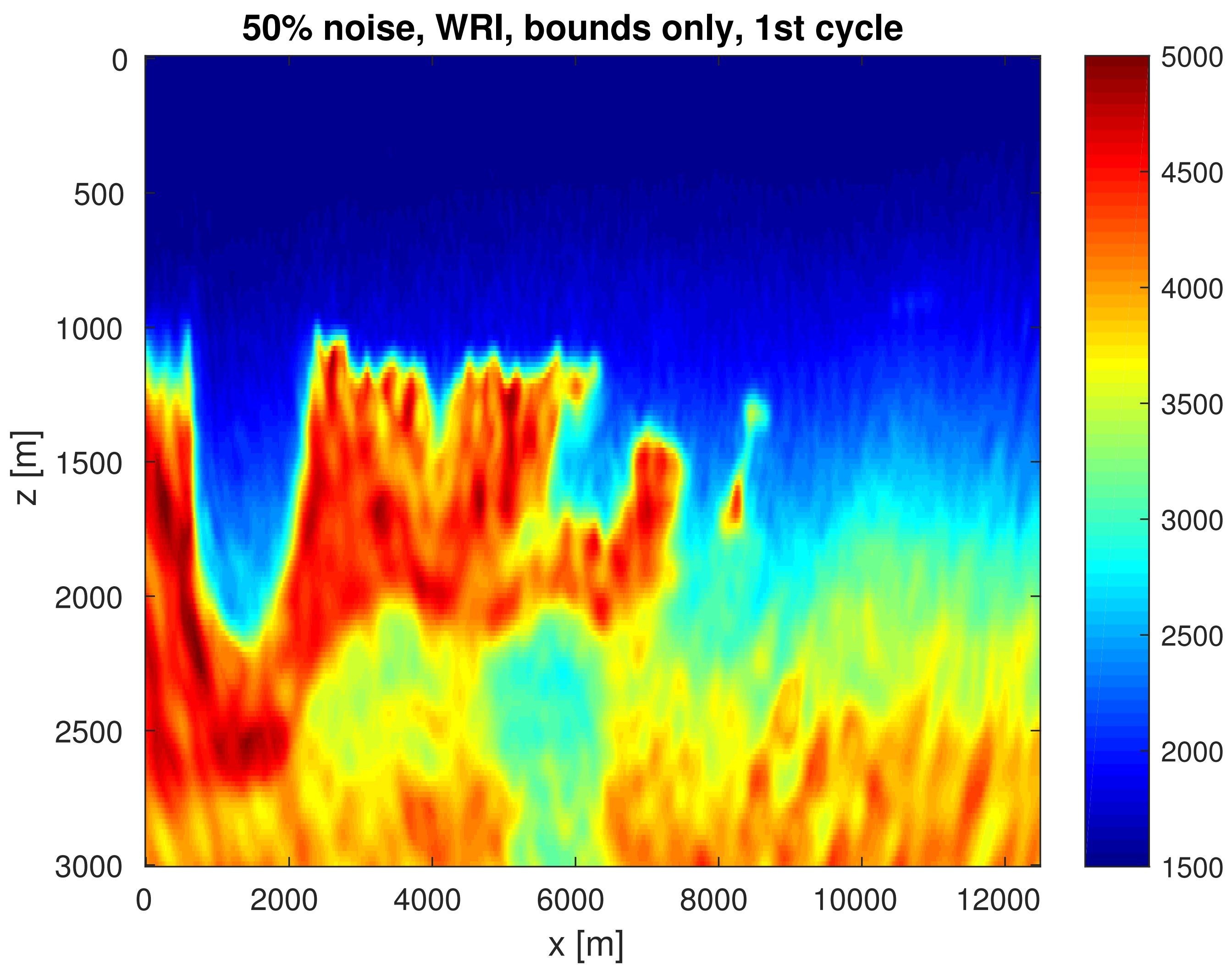


1st cycle cycle

$$\|\text{noise}\|_2 / \|\text{signal}\|_2 = 0.5$$

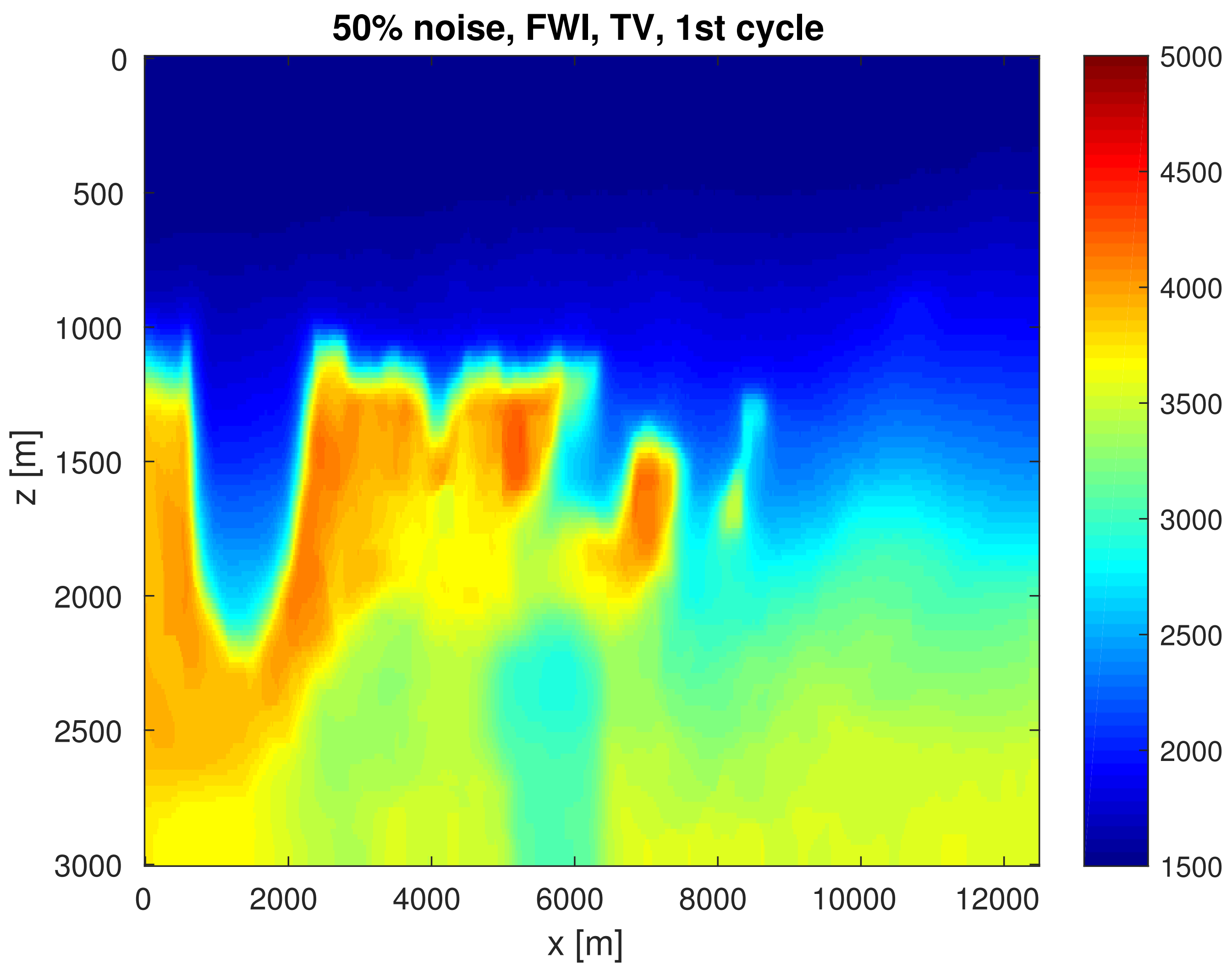
bounds only

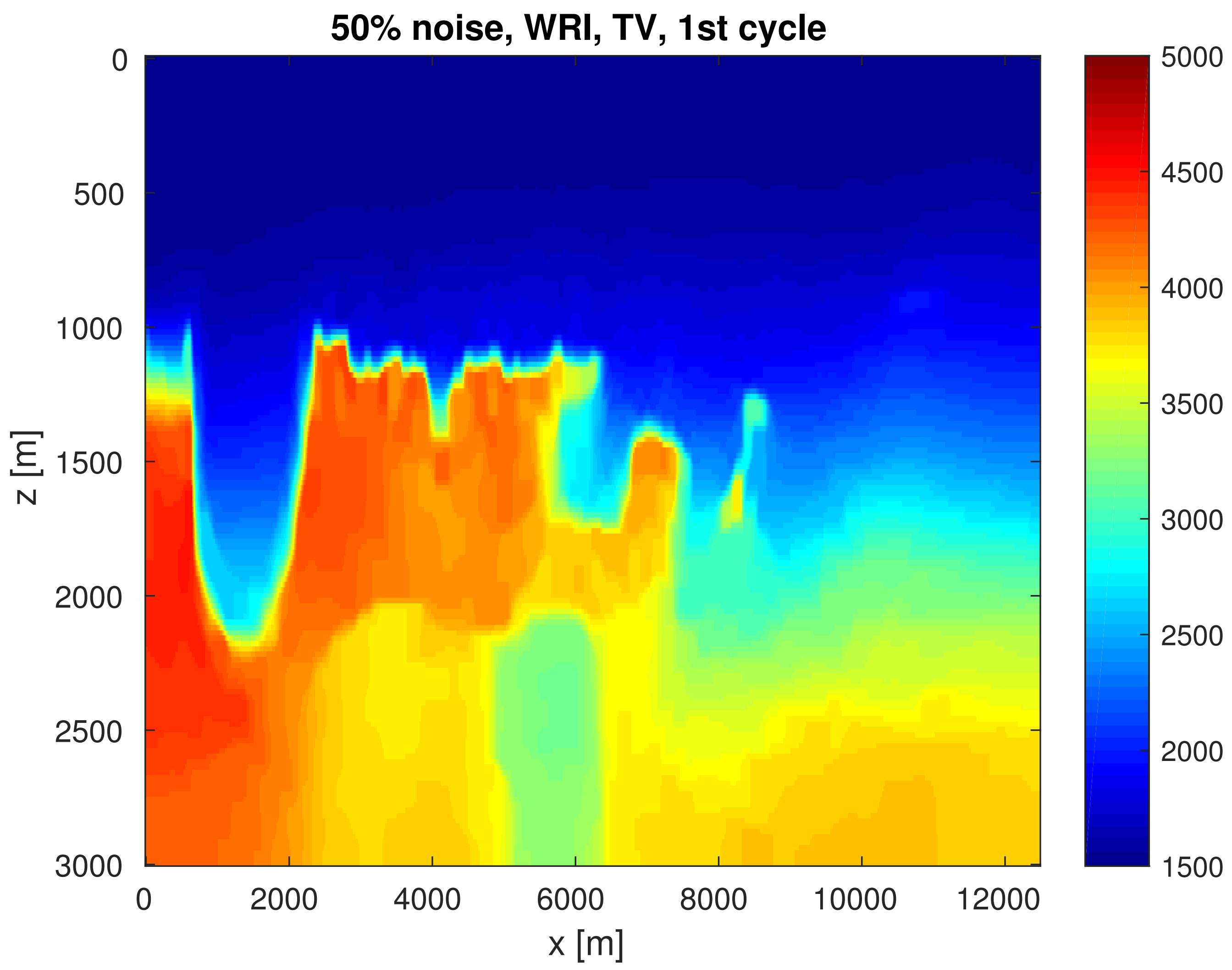




FWI **WRI**

bounds & TV

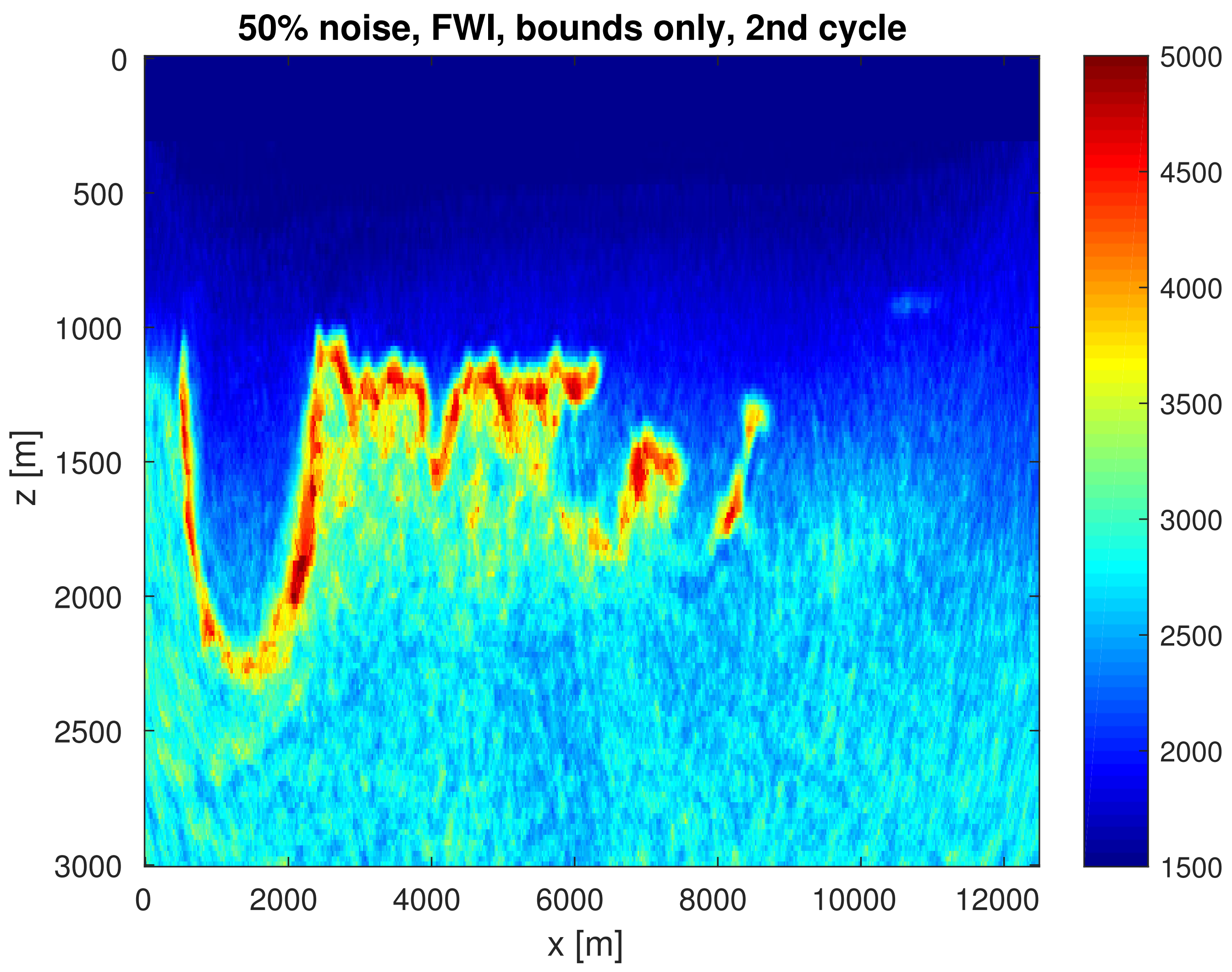


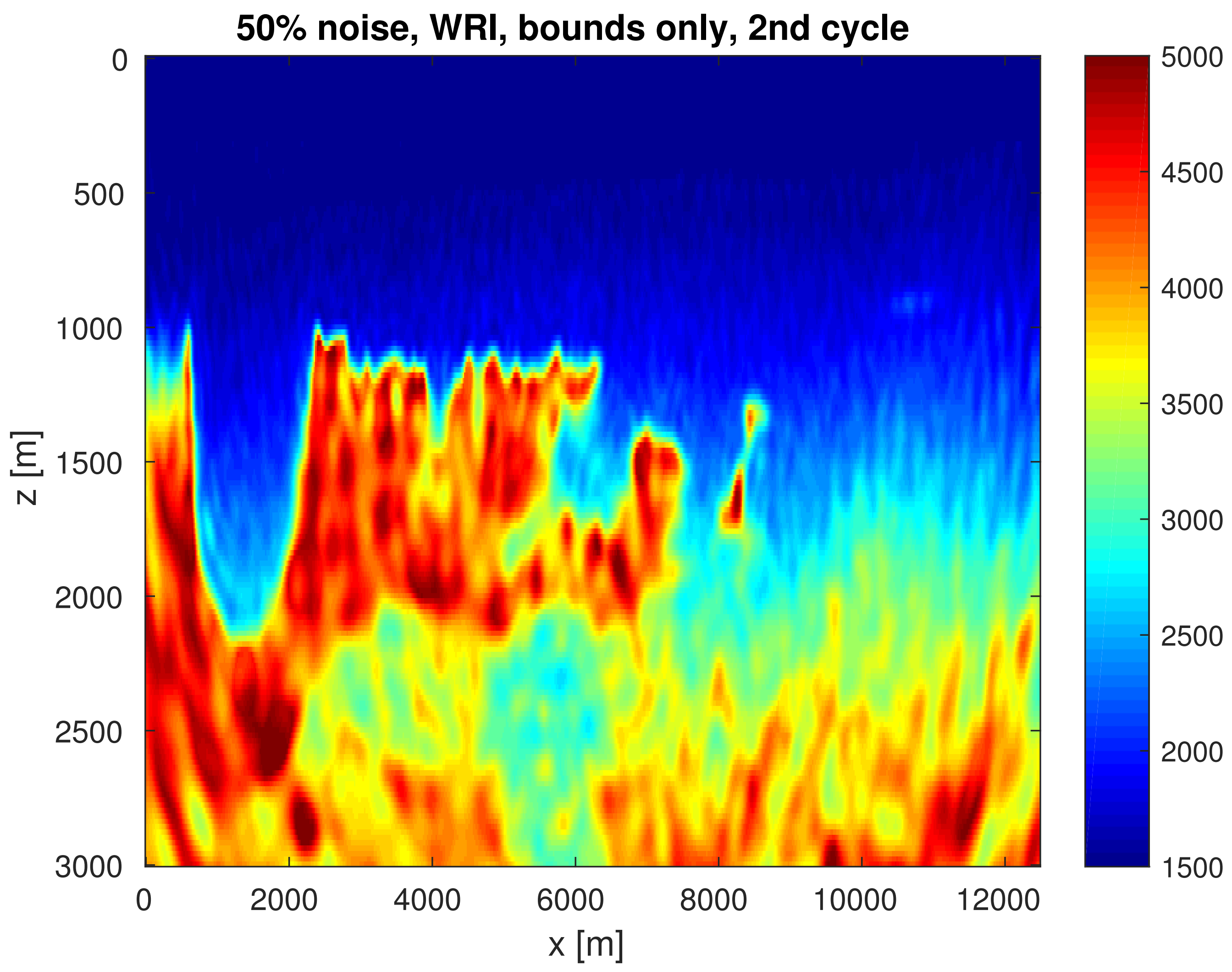


2nd cycle

$$\|\text{noise}\|_2 / \|\text{signal}\|_2 = 0.5$$

bounds only

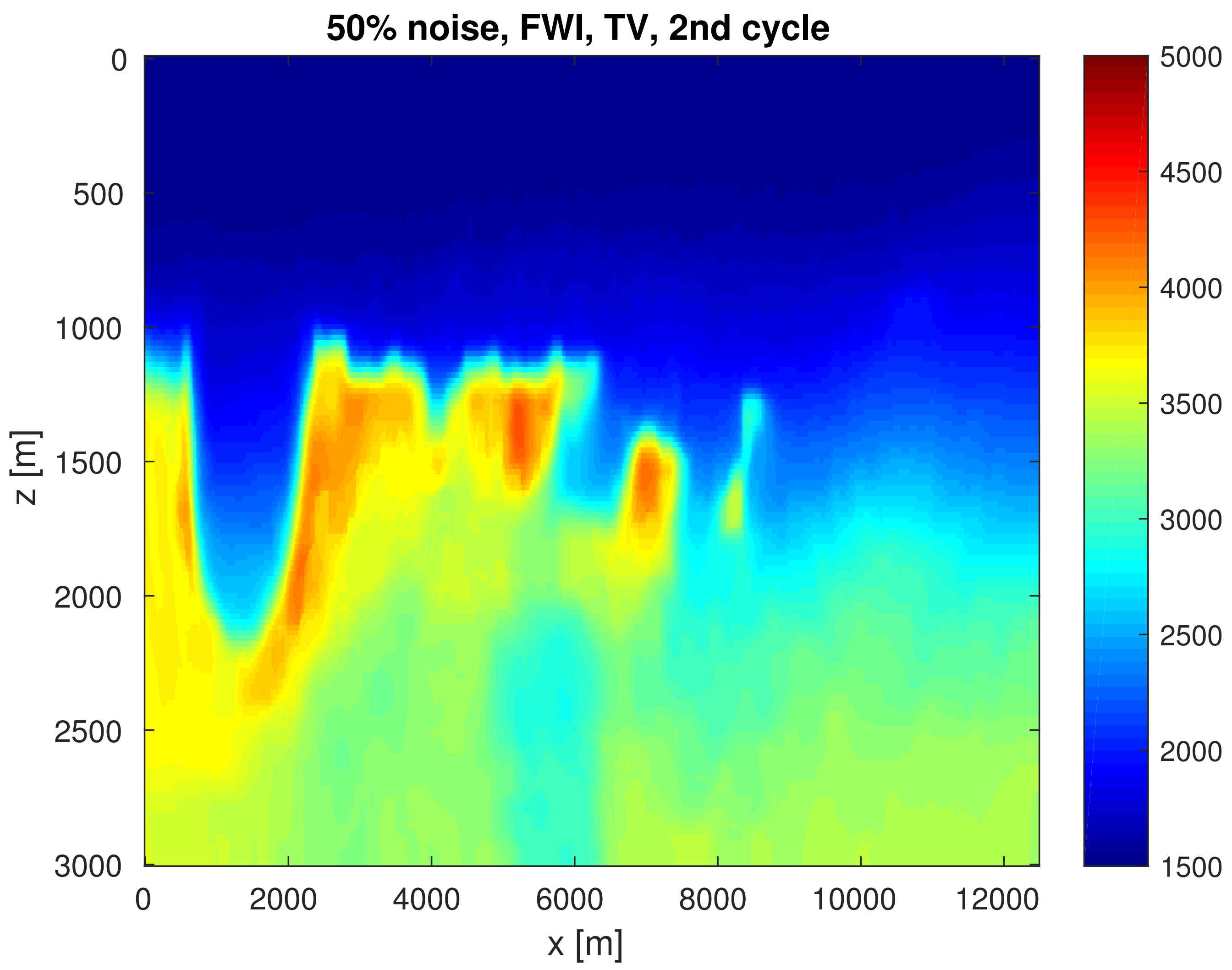


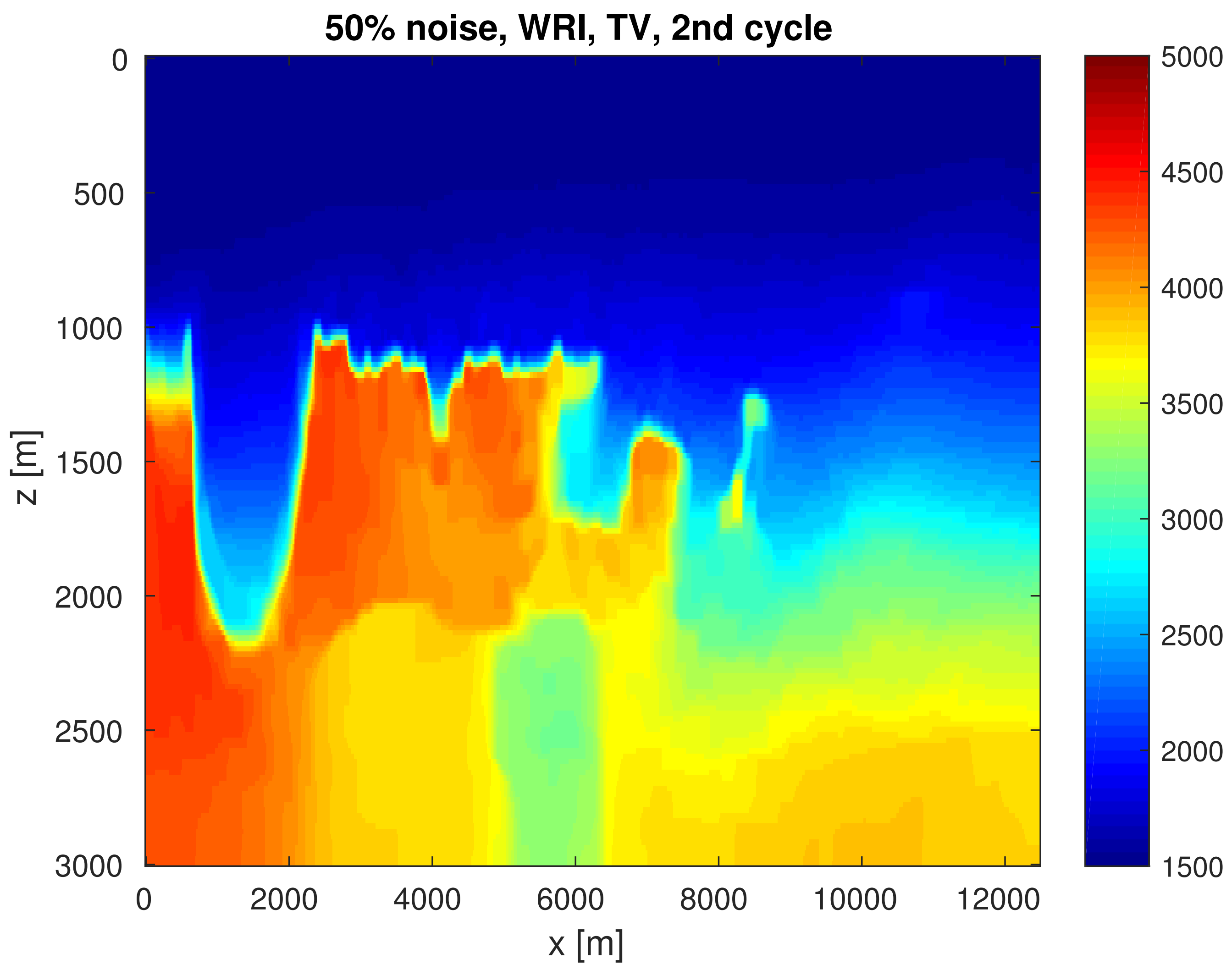


FWI

WRI

bounds & TV

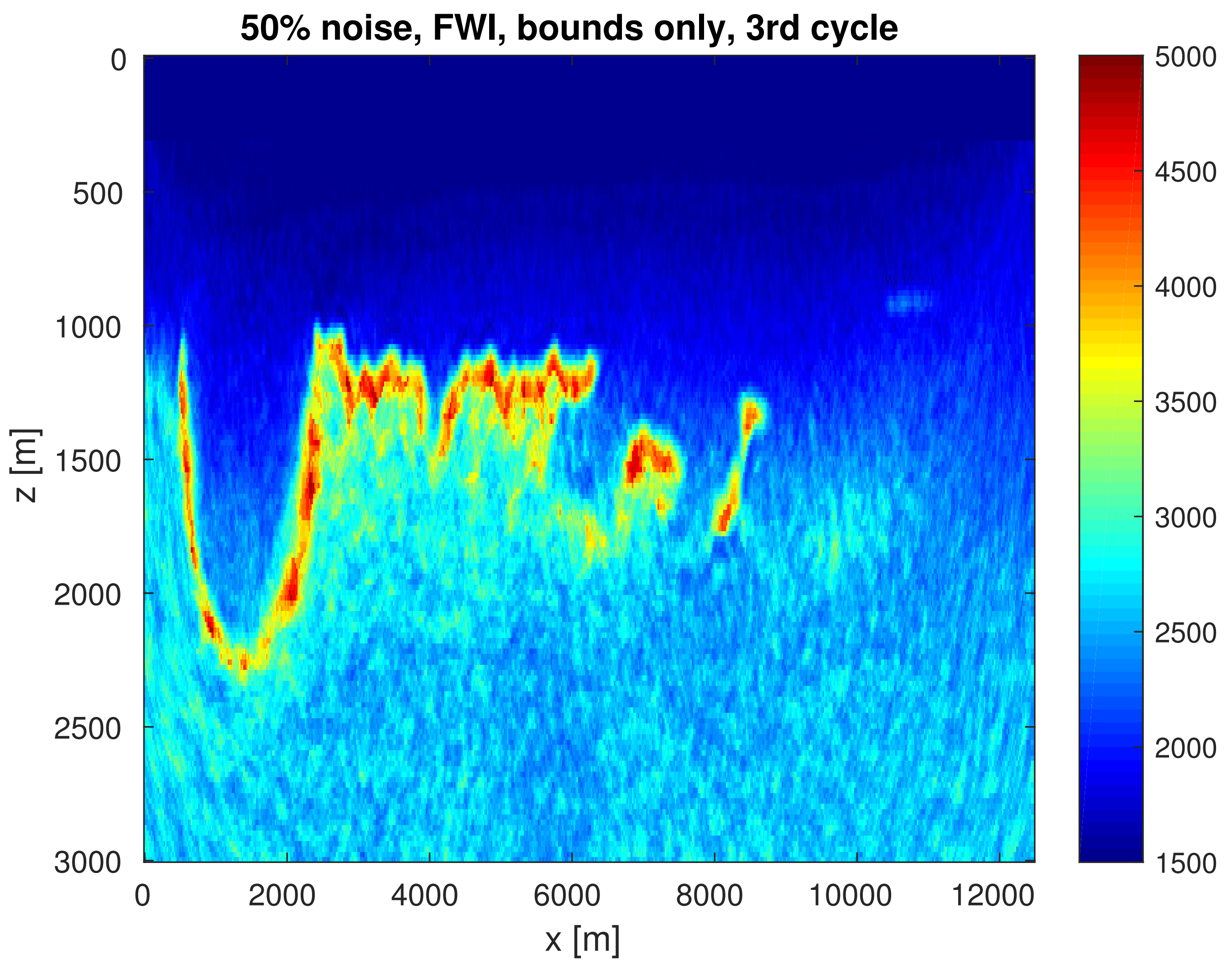


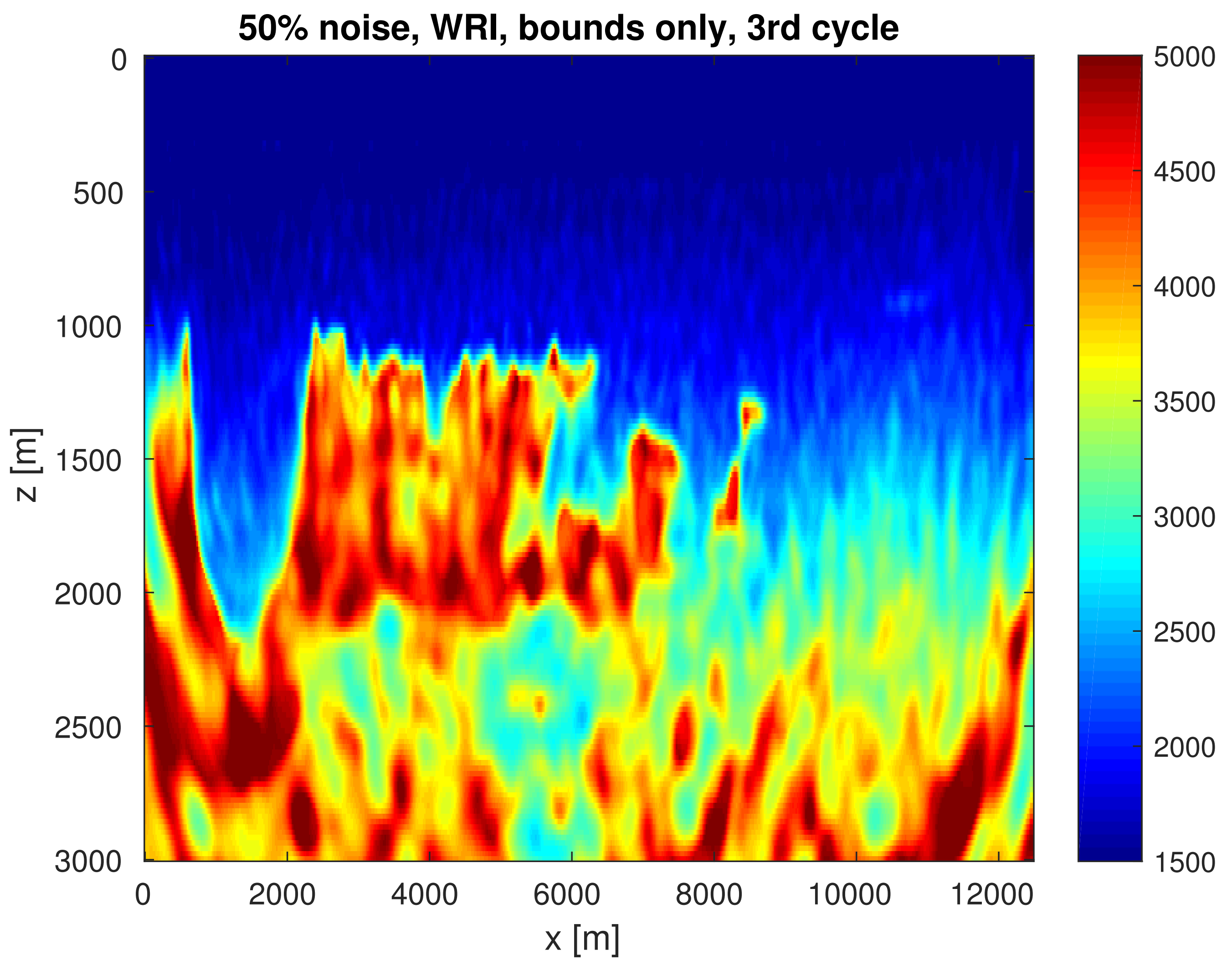


3rd cycle

$$\|\text{noise}\|_2/\|\text{signal}\|_2 = 0.5$$

bounds only

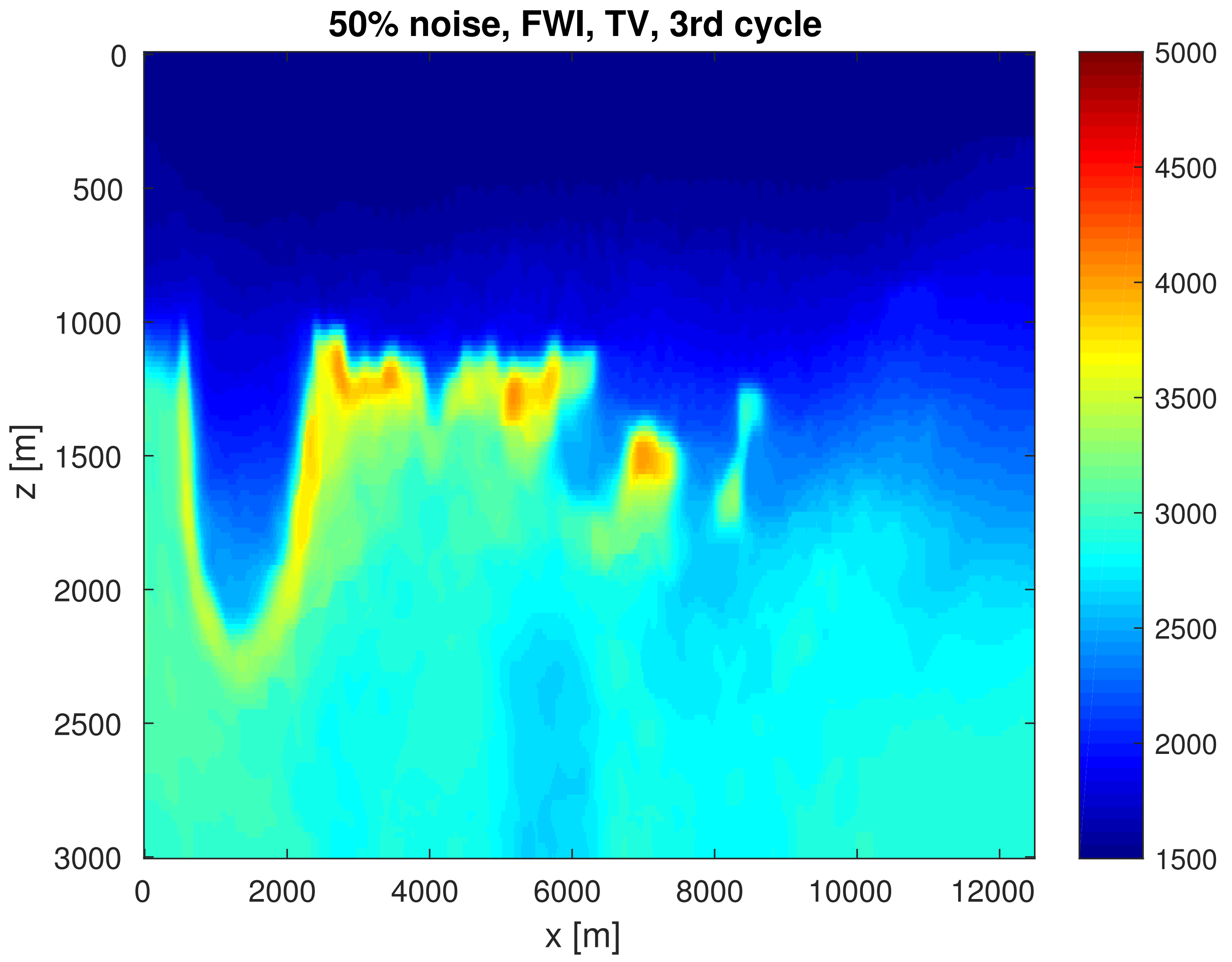


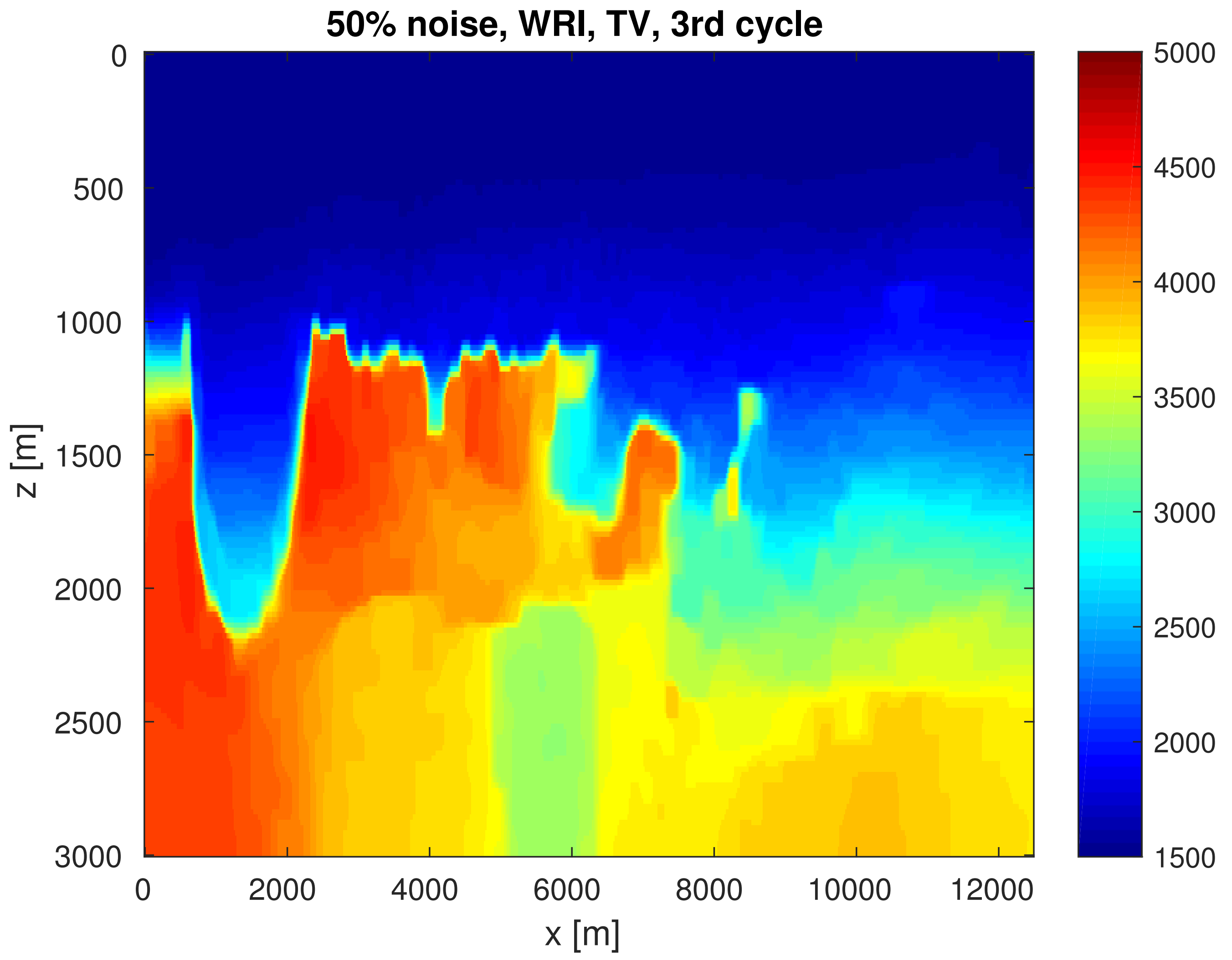


FWI

WRI

bounds & TV





Conclusions & generalizations

Adding constraints to inversion:

- leaves gradient (and Hessian) untouched
- is robust & behaves predictably w/ model iterates that remain feasible
- intuitive parameterizations of prior information in >2 constraints
- can be accelerated w/ quasi-Newton & parallelized projections
- “black box” solution that work w/ any implementation for FWI/WRI

Extensions paired w/ constraints are a powerful combination!

Acknowledgements

This research was carried out as part of the SINBAD project with the support of the member organizations of the SINBAD Consortium.