The Joint Power of Deterministic and Bayesian Approaches for (Seismic) Inverse Problems

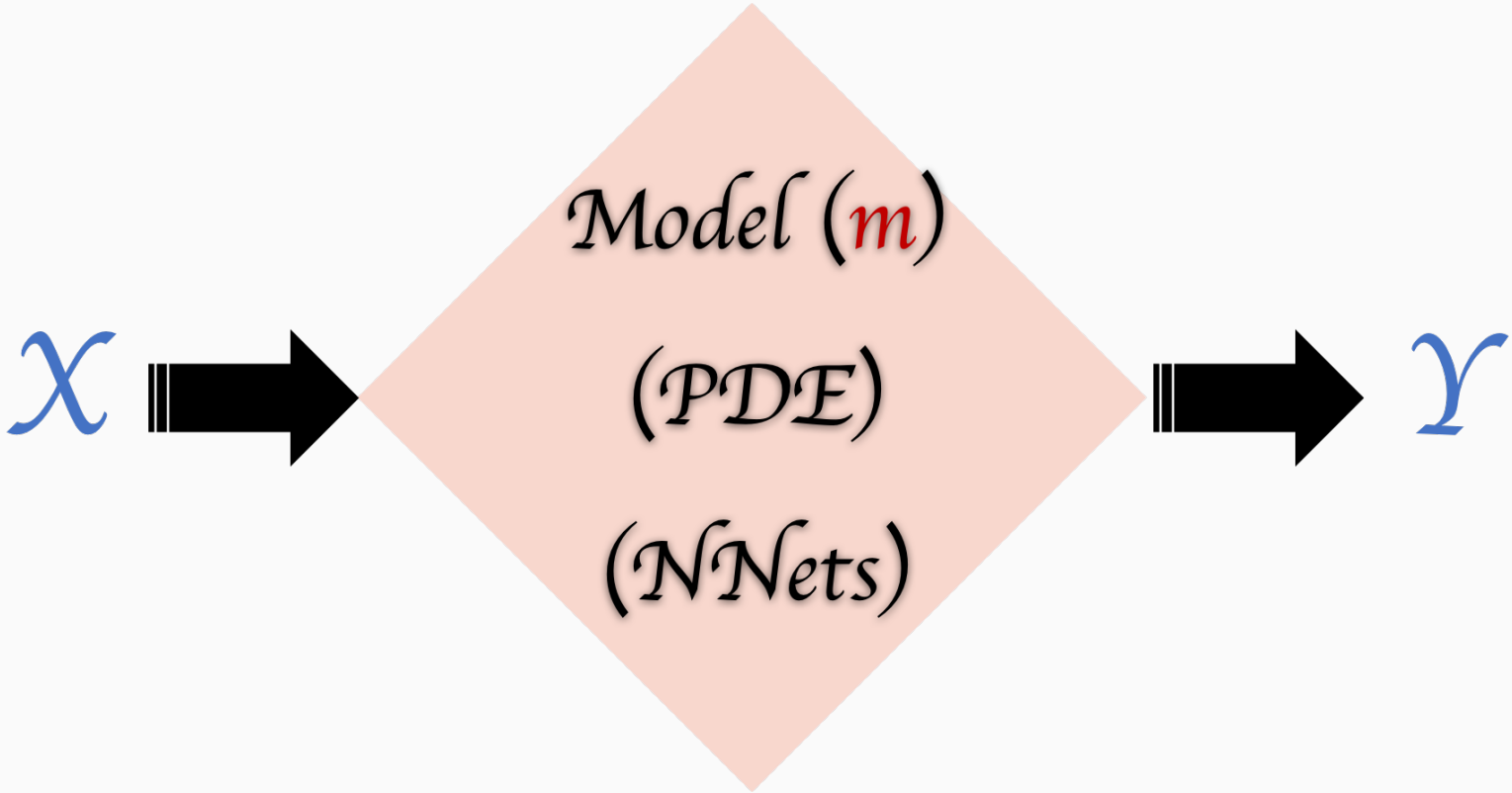
Yunan Yang, Daniel Appelö (CU Boulder), Matt Dunlop (NYU), Björn Engquist (UT-Austin), Kui Ren (Columbia), Alex Townsend (Cornell)

April 27, 2020

This work is partially supported by NSF DMS-1913129.

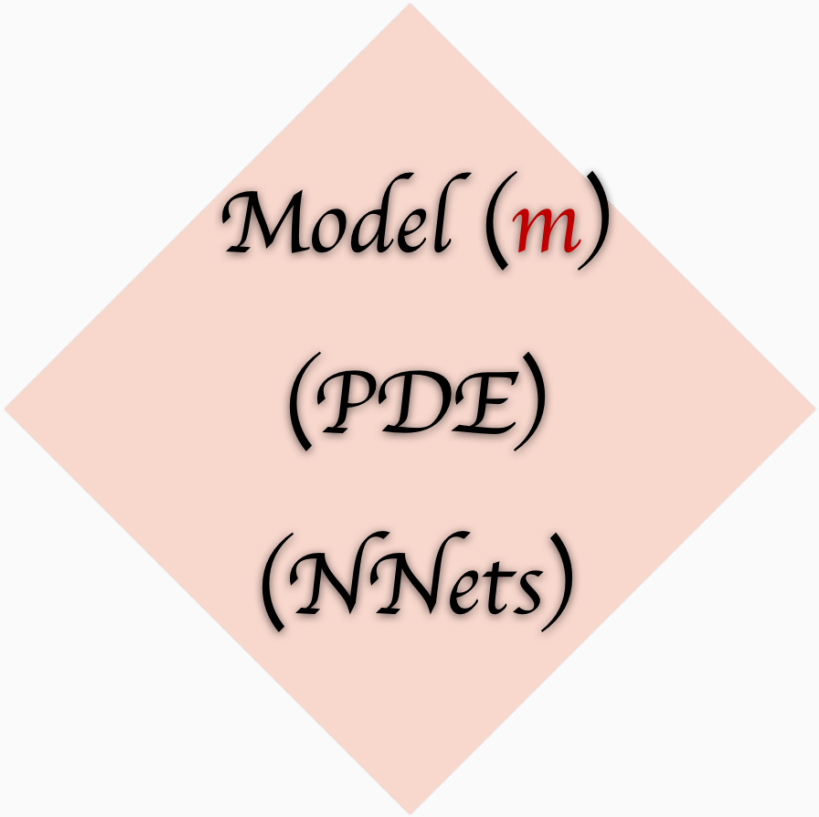
Cornell SCAN Seminar over Zoom

General Large-Scale Inverse Problem



Given (many) $\{X, Y\}$, find m.

The Model $F(m)$



F is given; we just find m (e.g., PDEs).

OR

F is not known; m depends on F .

The Model $F(m)$

F is given; we just find m (e.g., PDEs).

- Pro: We know the best (exact) forward problem!
- Con: The forward and inverse problems are so nonlinear!

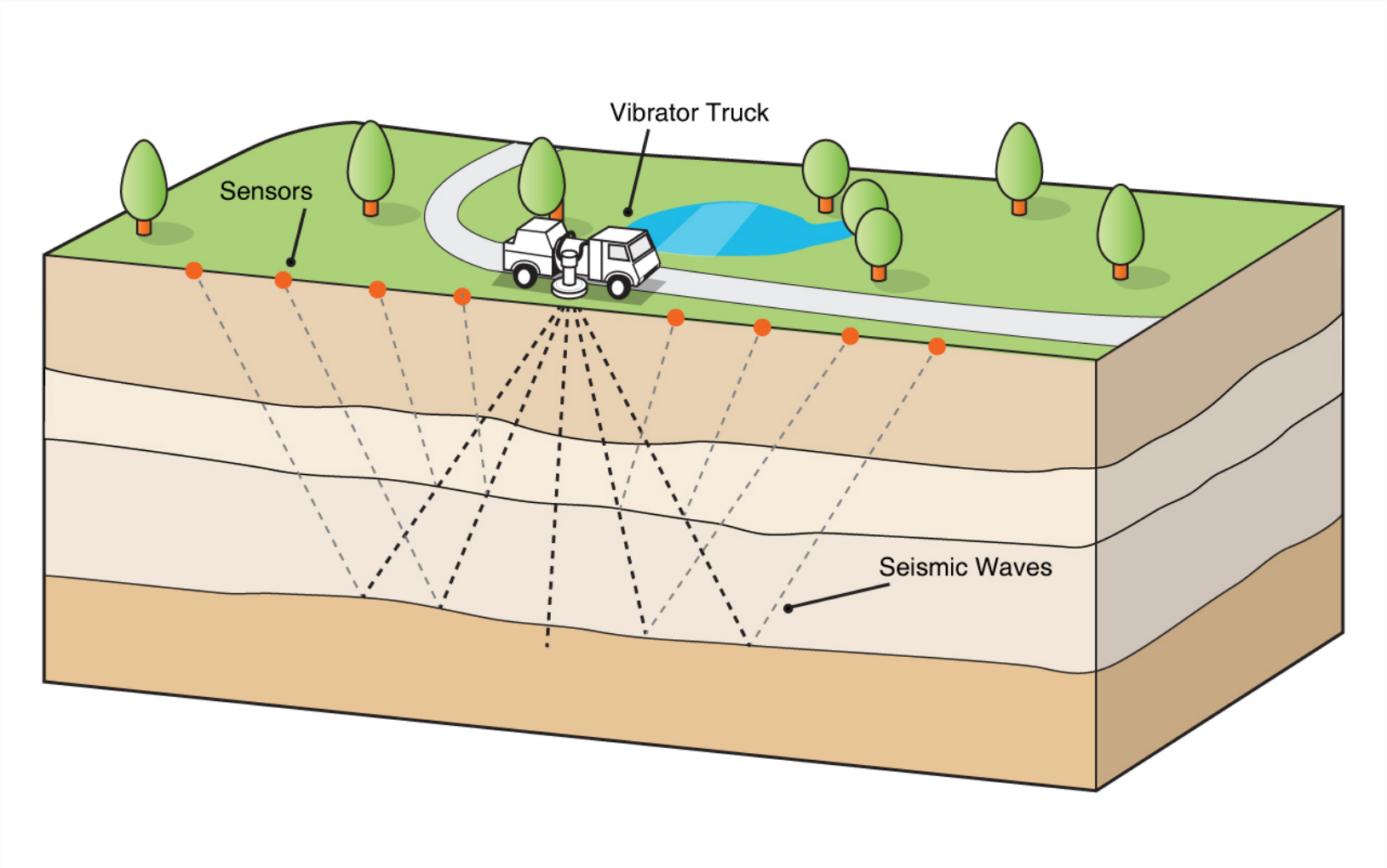
OR

F is not known; we are free to choose (e.g., XXX-net).

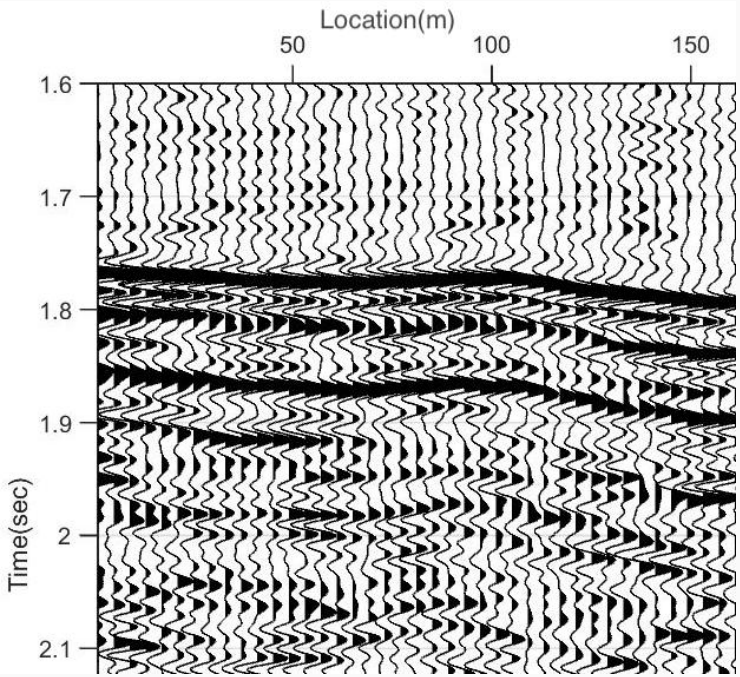
- Pro: The freedom to modify it to a “better” map (Over-Parametrization, ReLu)
- Con: Trial and error to build the model

Seismic Inversion: Earthquake Source, Hydrocarbons, etc.

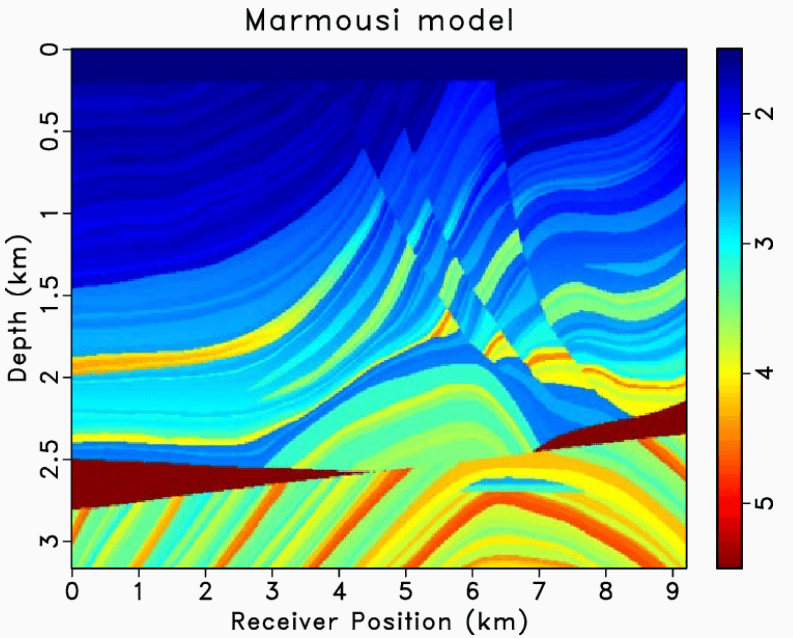
Seismic inversion is one of the inherently more difficult families of large-scale nonlinear inverse problems.



Seismic Inversion



Invert



Waveform measurements from receivers at the surface

Subsurface properties (i.e. wave velocity or material density)

PDE-Constrained Optimization

Forward Problem

$\mathcal{F} : m \to u|_\Gamma, \Gamma \subseteq \partial\Omega$ or Ω

Inverse Problem

$\mathcal{G} : u|_\Gamma \to m$

$\mathcal{F}$ and $\mathcal{G}$ are often nonlinear.

$$m^* = \underset{m}{\operatorname{argmin}}\, J(f(m), g)$$

$$f(m) = u|_\Gamma$$

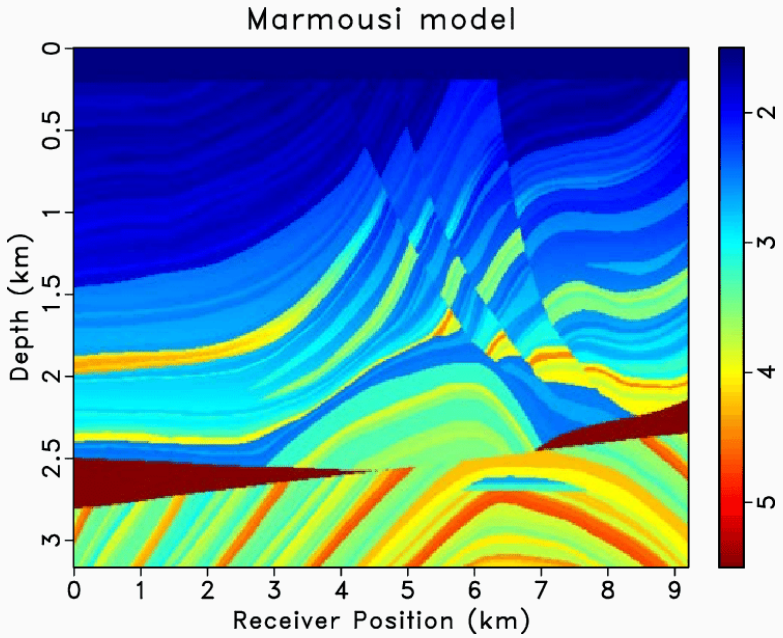
J is an objective function measuring the difference between f and g.

Seismic Inversion

Forward Wave Propagation

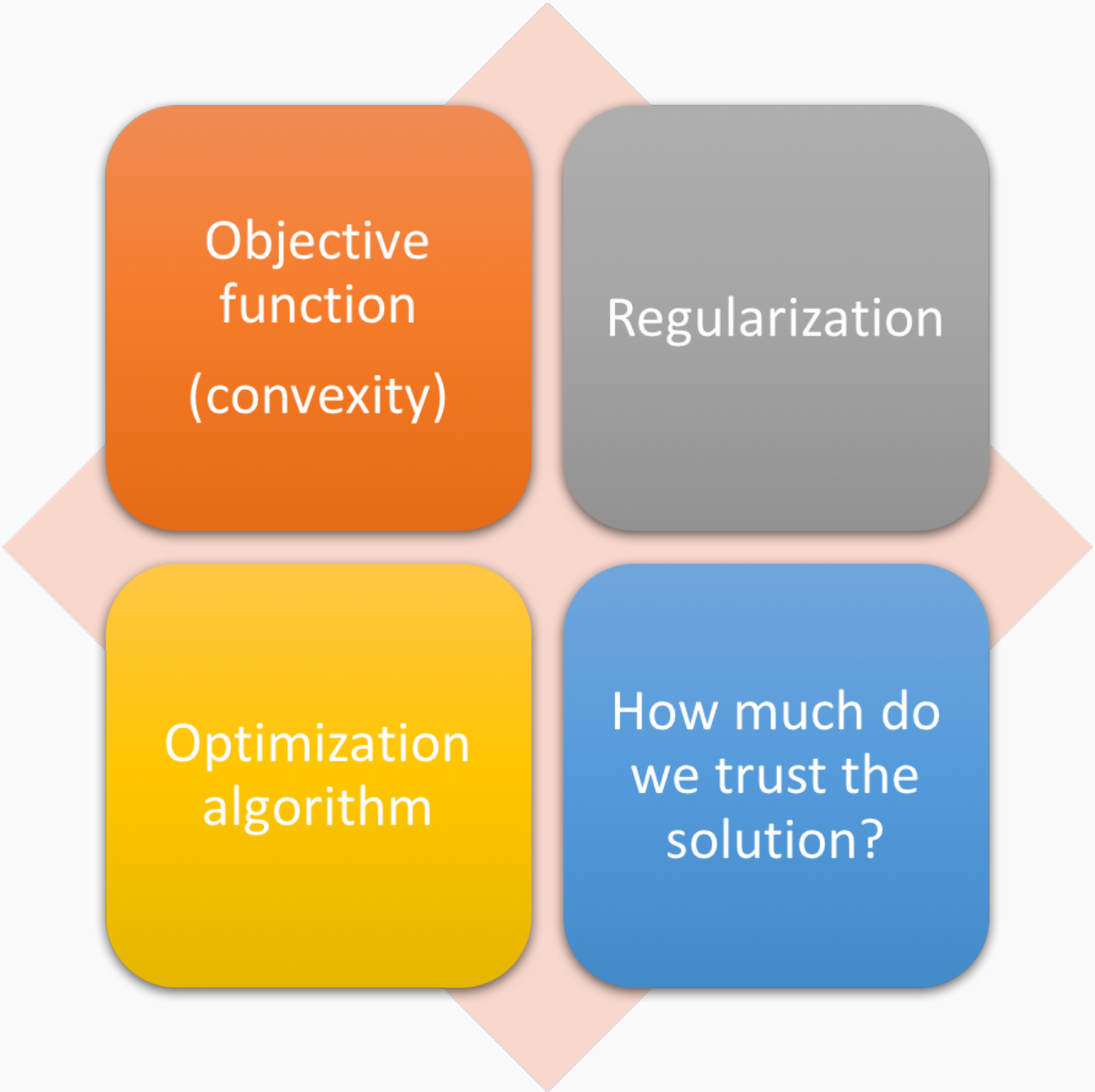
$$\begin{cases} m(\mathbf{x})\dfrac{\partial^2 u(\mathbf{x},t)}{\partial t^2} - \triangle u(\mathbf{x},t) = s(\mathbf{x},t) \\ \text{Zero i.c. in half-space } \Omega \\ \text{Neumann b.c. on } \partial\Omega \end{cases}$$

$m(\mathbf{x}) = \dfrac{1}{c(\mathbf{x})^2}$, $c(\mathbf{x})$ is the wave velocity

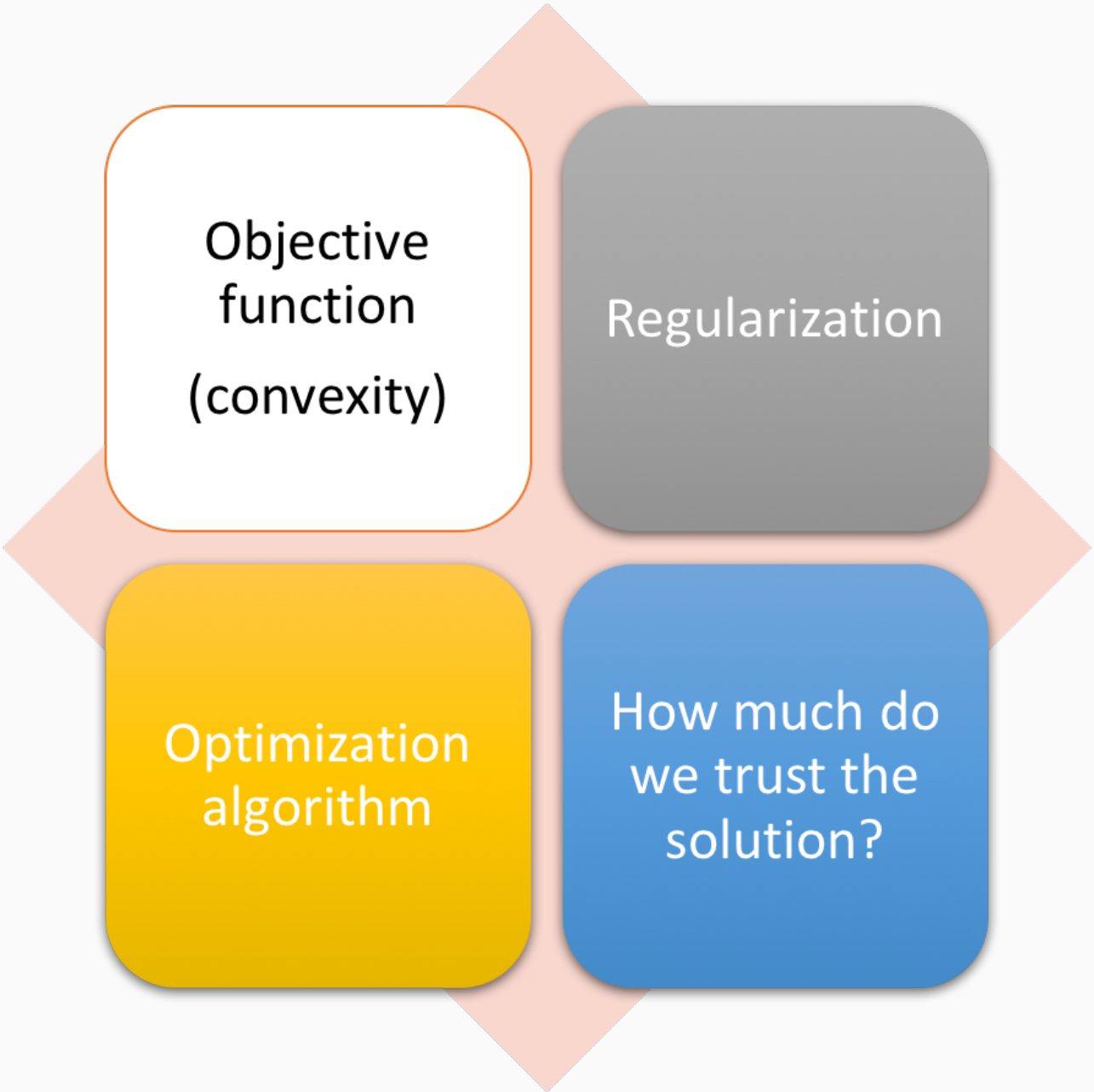


m

Important Components in the deterministic approach



The objective function



A PDE-Constrained Optimization

Traditional Least-Squares (L^2 norm) Objective Function

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

- observed data g,
- simulated data $f(m) = u|_\Gamma$,
- receiver x_r,
- the model parameter m,
- Regularization is often added in (1).

Main Challenges

1. Local minima trapping
2. Sensitive to noise

Typical Effects from Variations in Wave Speed

The shift and dilation are typical effects from variations in velocity parameter $m(x) = m$ (constant). For example:

$$\begin{cases} m\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}, & x > 0, t > 0, \\ u = 0, \quad \frac{\partial u}{\partial t} = 0, & x > 0, t = 0, \\ u = f(t), & x = 0, t > 0. \end{cases}$$

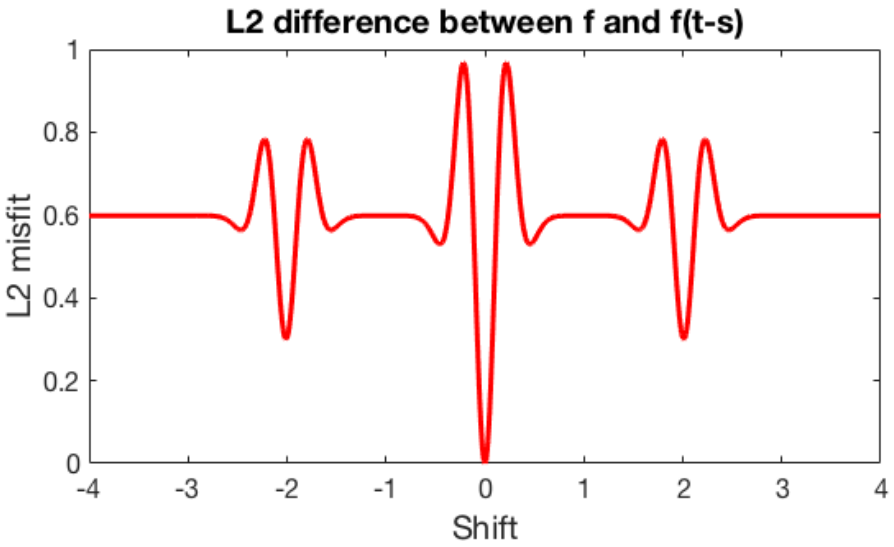
The solution to the equation is $u(x, t; m) = f(t - \sqrt{m}x)$.

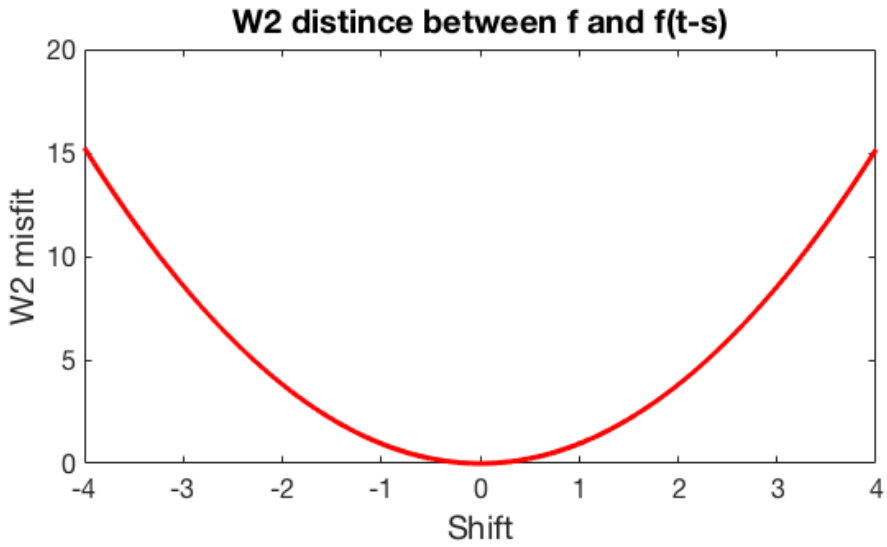
For fixed x, variation in m relates **shifts** in the signal.

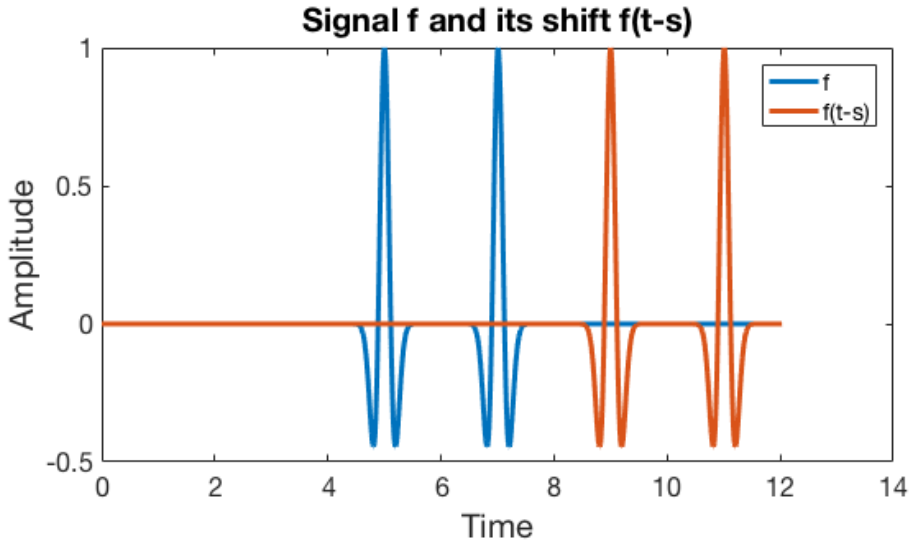
For fixed t, variation in m generates the **dilation** in data.

[Engquist, Froese & Y, 2016]

Motivation of Using the Wasserstein Distance (EMD)







[Engquist-Froese, 2014] [Engquist, Froese & Y, 2016]

The Quadratic Wasserstein Distance

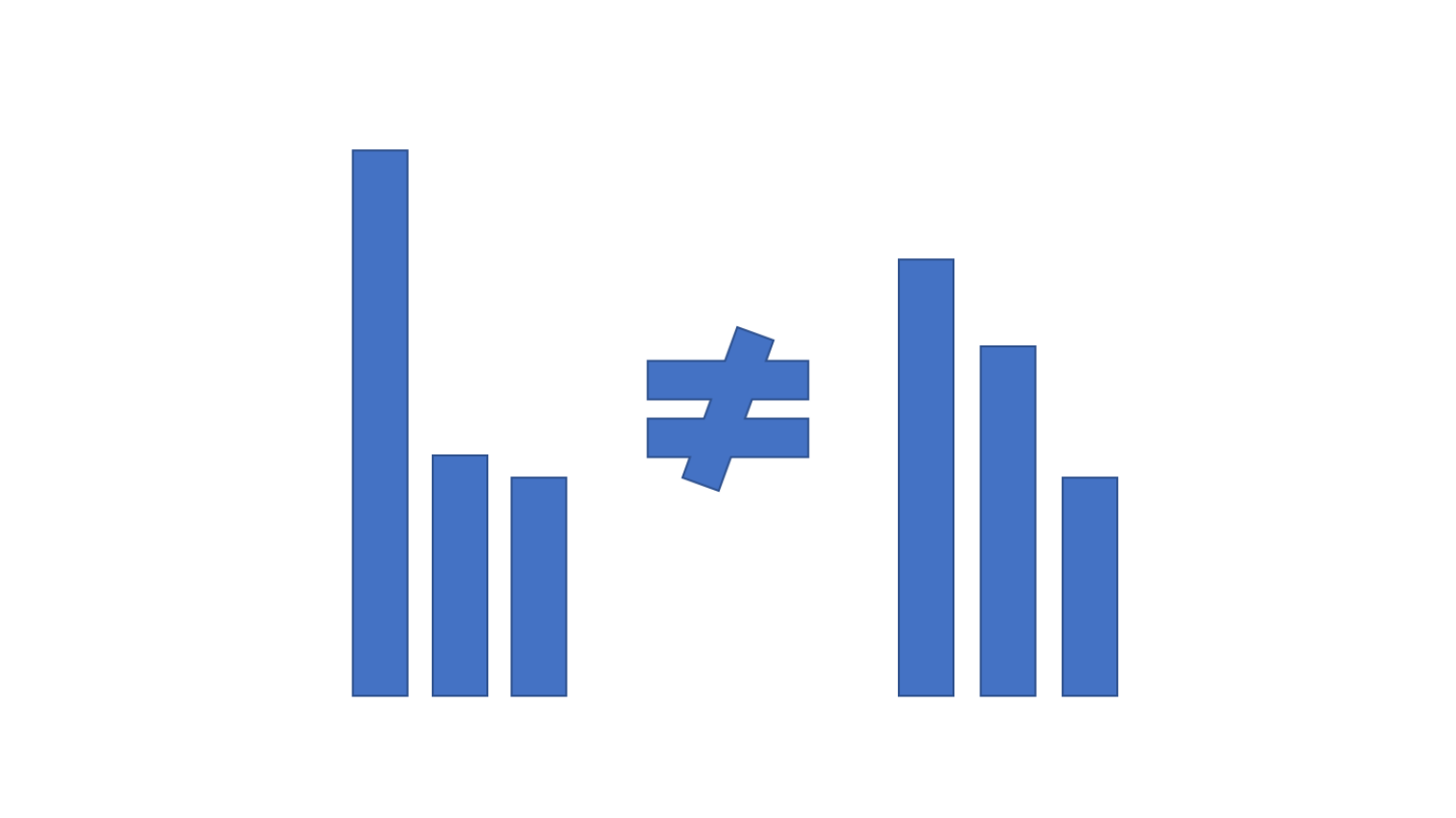
For $f, g \in \mathcal{P}(\Omega)$ ($f, g \geq 0$ and $\int f = \int g = 1$), the quadratic Wasserstein distance is formulated as

$$W_2(f, g) = \left(\inf_{T \in \mathcal{M}} \int |x - T(x)|^p f(x) dx \right)^{\frac{1}{2}} \tag{2}$$

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

[Monge, 1781]

Optimal Transport



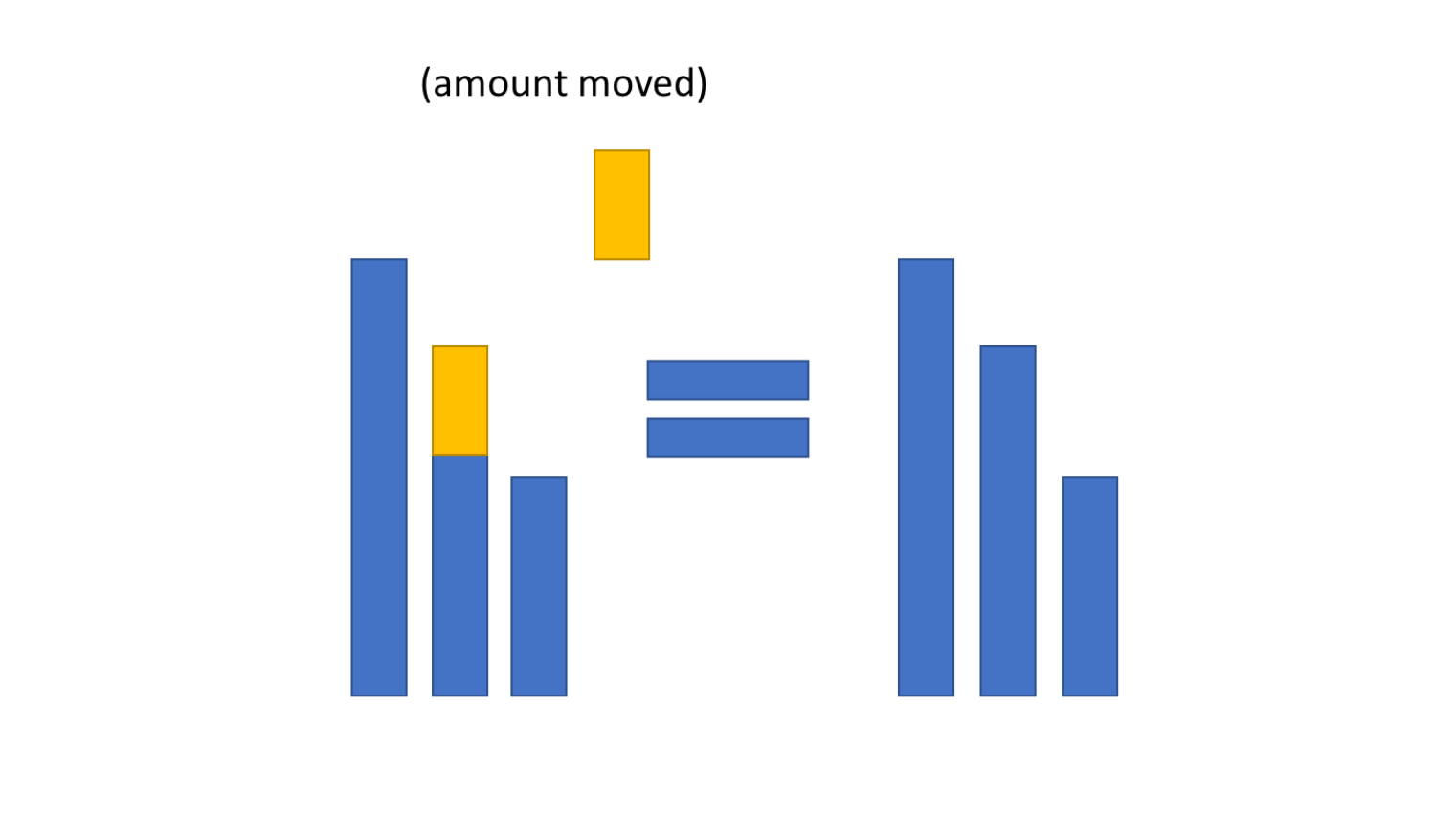
Synthetic data f (left) and observed data g (right)

[Monge, 1781]

Optimal Transport

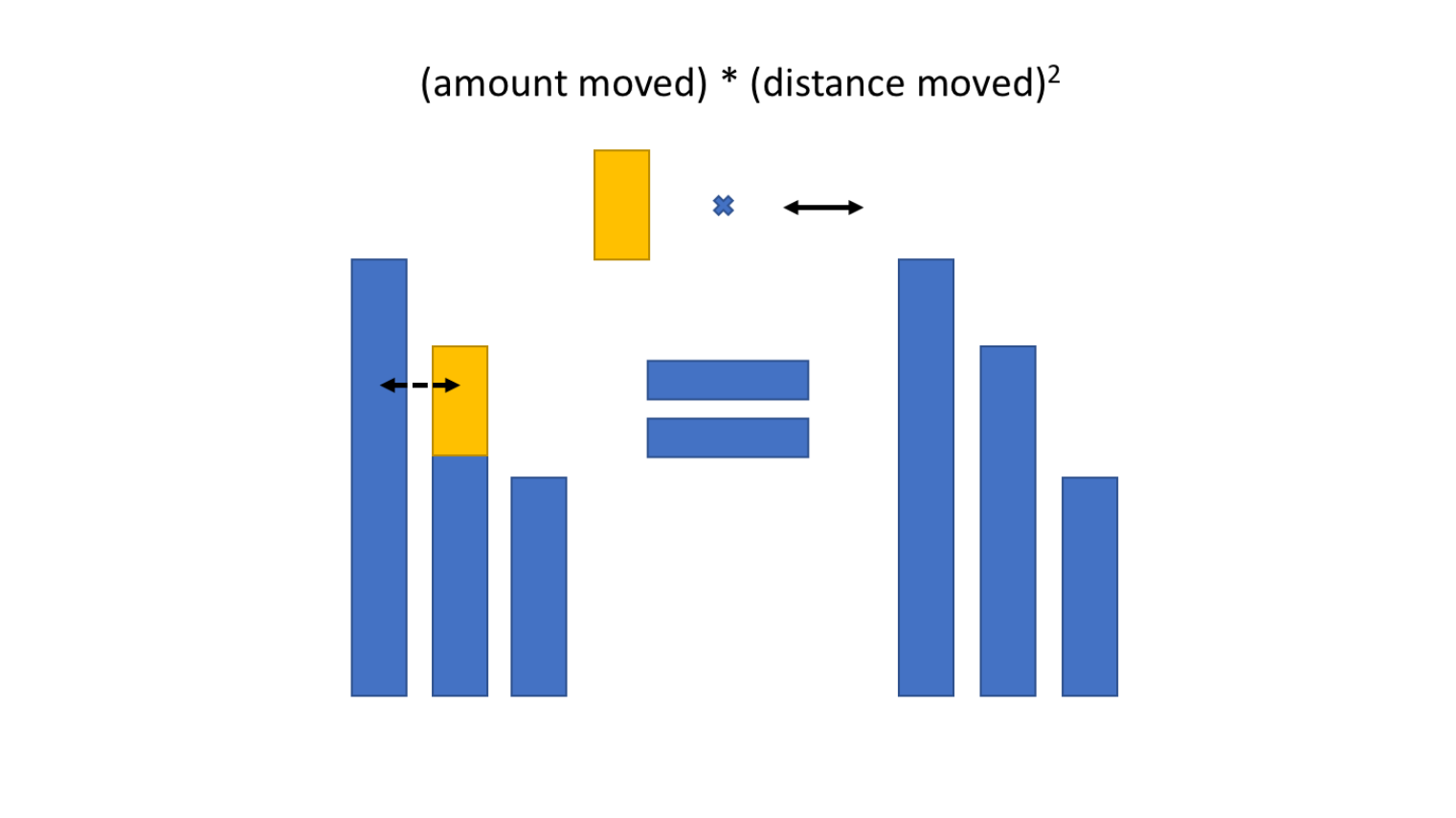
Synthetic data f (left) and observed data g (right)

Optimal Transport



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

Optimal Transport



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

1. Inversion with Local Minima (Optimization Landscape)

Let $\{e_k\}_{k=1}^d$ be standard basis of the Euclidean space $\mathbb{R}^d$.
Assume $s_k \in \mathbb{R}, \lambda_k \in \mathbb{R}^+, k = 1, \ldots, d$ and $A = \mathrm{diag}(1/\lambda_1, \ldots, 1/\lambda_d)$.
We define f_Θ as jointly the translation and dilation of g:

$$f_\Theta(x) = \det(A) g(A(x - \sum_{k=1}^{d} s_k e_k)), \Theta = \{s_1, \ldots, s_d, \lambda_1, \ldots, \lambda_d\}.$$

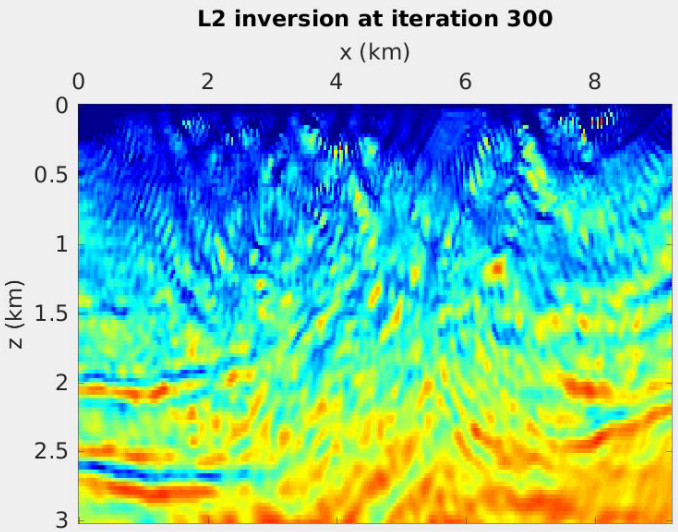
Theorem (Convexity of W_2 in translation and dilation)

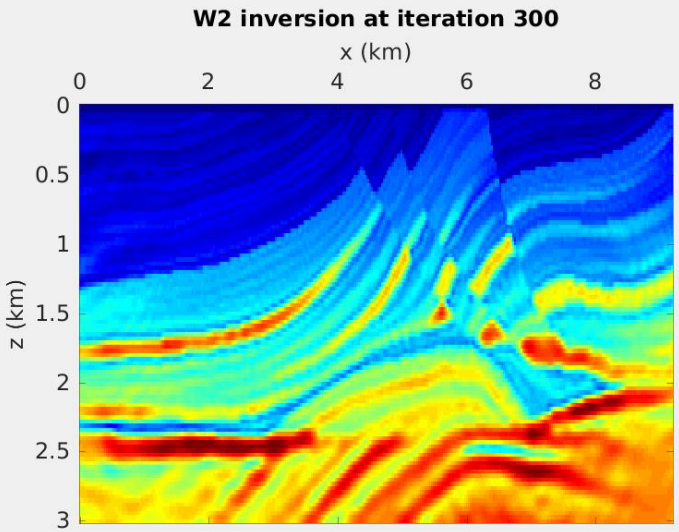
The optimal map between $f_\Theta(x)$ and $g(y)$ is $y = T_\Theta(x)$ where $\langle T_\Theta(x), e_k \rangle = \frac{1}{\lambda_k}(\langle x, e_k \rangle - s_k), k = 1, \ldots, d.$

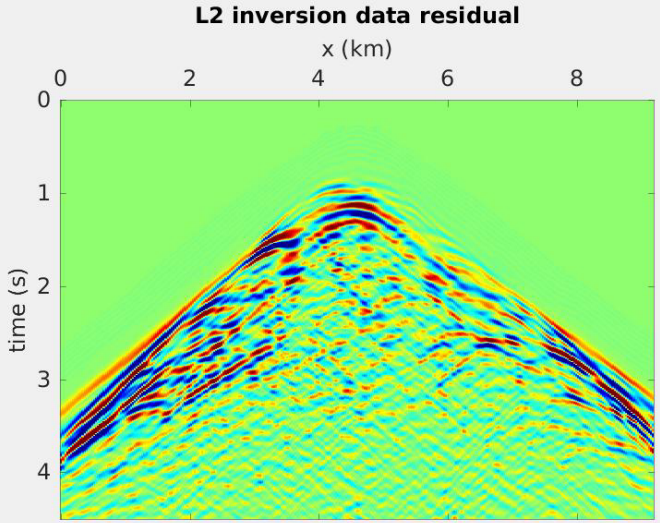
Moreover, $I(\Theta) = W_2^2(f_\Theta(x), g)$ is a convex function of Θ.

[Y, 2019]

Tackling Nonconvexity







[Y-Engquist-Sun-Hamfeldt, 2016]

2. More Robust w.r.t. Noise (Perturbation)

Given strictly positive probability density $f = d\nu$, we can define a Laplace-type linear operator

$$L = -\Delta + \nabla(-\log f) \cdot \nabla$$

which satisfies the fundamental integration by parts formula:

$$\int_{\mathbb{R}^d} (Lh_1)h_2 d\nu = \int_{\mathbb{R}^d} h_1(Lh_2)d\nu = \int_{\mathbb{R}^d} \nabla h_1 \cdot \nabla h_2 d\nu.$$

$$\|h\|^2_{L^2(f)} = \int_{\mathbb{R}^d} h^2 d\nu, \quad \|h\|^2_{\dot{H}^1(f)} = \int_{\mathbb{R}^d} |\nabla h|^2 d\nu,$$

$$\|h\|^2_{\dot{H}^{-1}(f)} := \sup\left\{ \int_{\mathbb{R}^d} h\varphi d\nu \;\middle|\; \|\varphi\|^2_{\dot{H}^1(f)} \le 1 \right\} = \int_{\mathbb{R}^d} h(L^{-1}h)d\nu.$$

If $f = 1$, we reconstruct the unweighted $\dot{\mathcal{H}}^{-1}_{(\mathbb{R}^d)}$ seminorm.

The Connection with Weak Norm

Asymptotic Connection

If μ is the probability measure and $d\pi$ is an infinitesimal perturbation that has zero total mass, then

$$W_2(\mu, \mu + d\pi) = \|d\pi\|_{\dot{\mathcal{H}}^{-1}_{(d\mu)}} + o(d\pi). \tag{3}$$

Non-Asymptotic Connection

If both $f = d\mu$ and $g = d\nu$ are bounded from below and above by constants $\mathfrak{c}_1$ and $\mathfrak{c}_2$, we have the following *non-asymptotic* equivalence between W_2 and $\dot{\mathcal{H}}^{-1}_{(d\mu)}$:

$$\frac{1}{\mathfrak{c}_2}\|\mu - \nu\|_{\dot{\mathcal{H}}^{-1}_{(\mathbb{R}^d)}} \leq W_2(\mu, \nu) \leq \frac{1}{\mathfrak{c}_1}\|\mu - \nu\|_{\dot{\mathcal{H}}^{-1}_{(\mathbb{R}^d)}}, \tag{4}$$

H^s-based inverse matching

A linear inverse problem of finding m from noisy data g_δ

$$Am = g_\delta. \tag{5}$$

A is diagonal in the Fourier domain:

$$\widehat{A}(\boldsymbol{\xi}) \sim \langle \boldsymbol{\xi} \rangle^{-\alpha}. \tag{6}$$

We seek the solution by minimizing the objective functional

$$\mathcal{O}_{\mathcal{H}^s}(m) \equiv \frac{1}{2}\|f(m) - g\|^2_{\mathcal{H}^s} := \frac{1}{2}\int_{\mathbb{R}^d} \langle \boldsymbol{\xi} \rangle^{2s} |\widehat{f}(m)(\boldsymbol{\xi}) - \widehat{g}(\boldsymbol{\xi})|^2 d\boldsymbol{\xi}, \tag{7}$$

[Engquist-Ren-Y, 2019]

The solution at frequency ξ is therefore

$$\widehat{m}(\xi) = \left(\widehat{A}^*(\xi)\left(\langle\xi\rangle^{2s}\widehat{A}\right)\right)^{-1}\widehat{A}^*(\xi)\left(\langle\xi\rangle^{2s}\widehat{g_\delta}(\xi)\right).$$

We can then perform an inverse Fourier transform to find the solution in physical space. The result is

$$m = \left(A^*PA\right)^{-1}A^*Pg_\delta, \qquad P := (\mathcal{I}-\Delta)^{s/2},$$

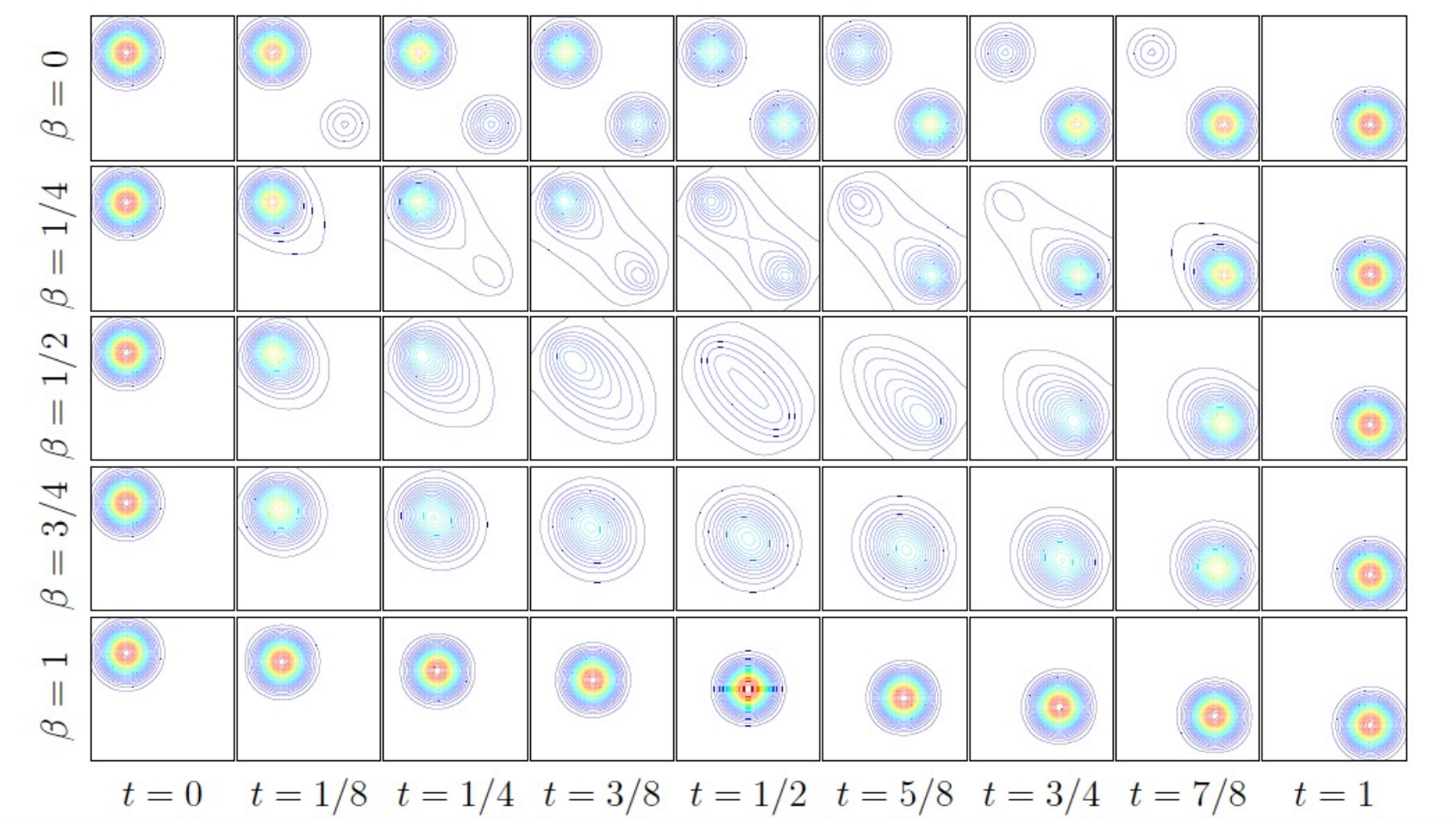
where the operator $(\mathcal{I}-\Delta)^{s/2}$ is defined through the relation

$$(\mathcal{I}-\Delta)^{s/2}m = \mathcal{F}^{-1}\left(\langle\xi\rangle^s\widehat{m}\right),$$

$s = 0, s > 0, s < 0.$

[Engquist-Ren-Y, 2019]

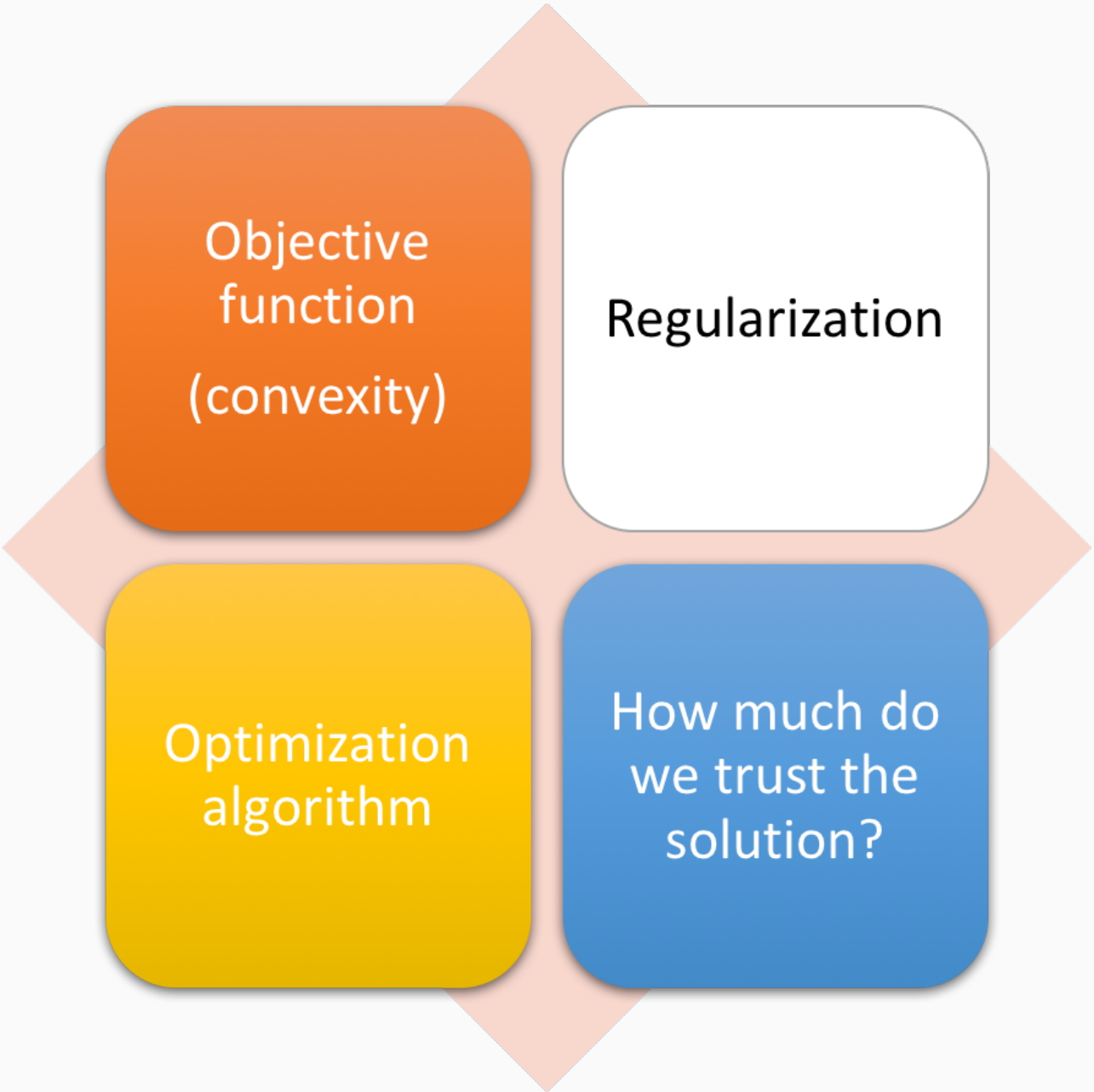
Differences Between W_2 and $\dot{H}^{-1}$



Top row: Geodesics in the $\dot{H}^{-1}$ space
Bottom row: Geodesics in the W_2 space

[Papadakis-Peyré-Oudet, 2013]

Regularization



Implicit Regularization

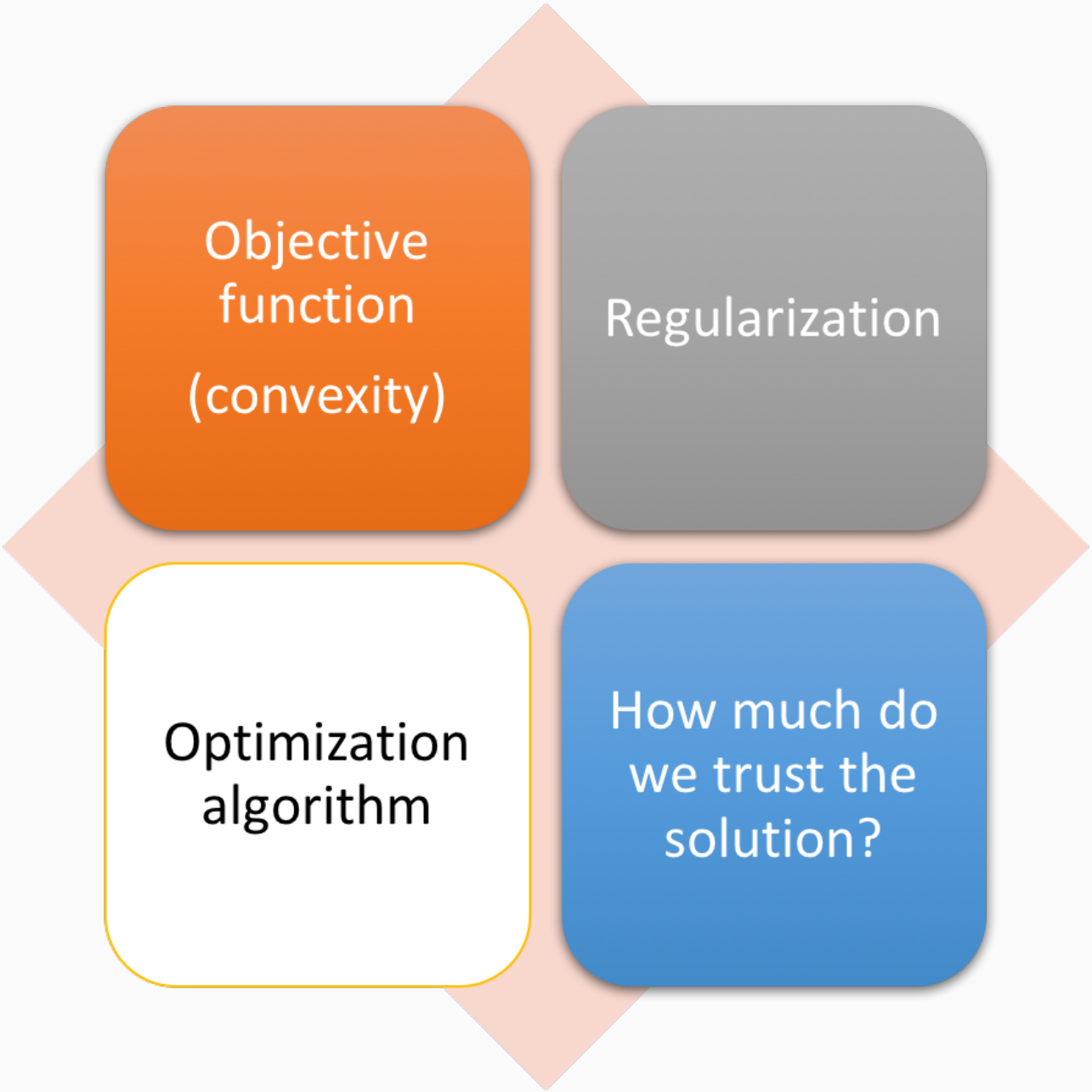
$$m^* = \underset{m}{\operatorname{argmin}}\, J(m) + R(m)$$

Regularization does not have to be in the form of $R(m)$.

- The choice of the objective function
- The choice of the data

  e.g., low-frequency data recovers low-wavenumber model
- The choice of numerical discretization
- The optimization algorithm (fixed step size)

Acceleration Methods



Acceleration Methods

Treat Gradient Descent as a Fixed-Point Iteration

$$p_{k+1} = p_k - \eta \left.\frac{\partial J}{\partial p}\right|_{p=p_k} = G(p_k).$$

$$p^* = G(p^*).$$

Steepest descent $\Leftrightarrow$ Picard Iteration applied to G

Can we do better?

Anderson Acceleration (Also known as: Pulay Mixing, DIIS)

Input: Given p_0 and $m \geq 1$. G is a given fixed-point operator.

Set $p_1 = G(p_0)$.

for k = 0, 1,2,... **do**

Step 1: Set $m_k = \min(m, k)$ and $F_k = (f_{k-m_k}, \ldots, f_k)$, where $f_i = G(p_i) - p_i$ is the residual.

Step 2: Find $\alpha^{(k)} = (\alpha_0^{(k)}, \ldots, \alpha_{m_k}^{(k)})^T$, $\sum_{i=0}^{m_k} \alpha_i^{(k)} = 1$ to minimize the sum of the weighted residual $\|F_k \alpha^{(k)}\|$.

Step 3: Update p_{k+1} according to

$$p_{k+1} = \sum_{i=0}^{m_k} \alpha_i^{(k)} G(p_{k-m_k+i}).$$

end for

[D. G. Anderson, 1965]

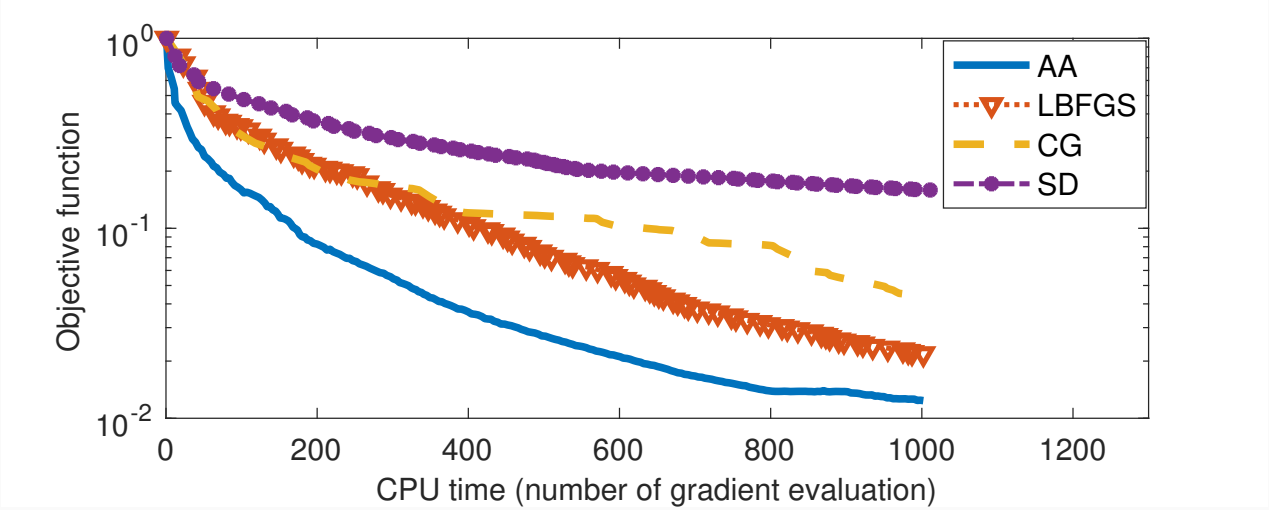
Anderson Acceleration (Also known as: Pulay Mixing, DIIS)

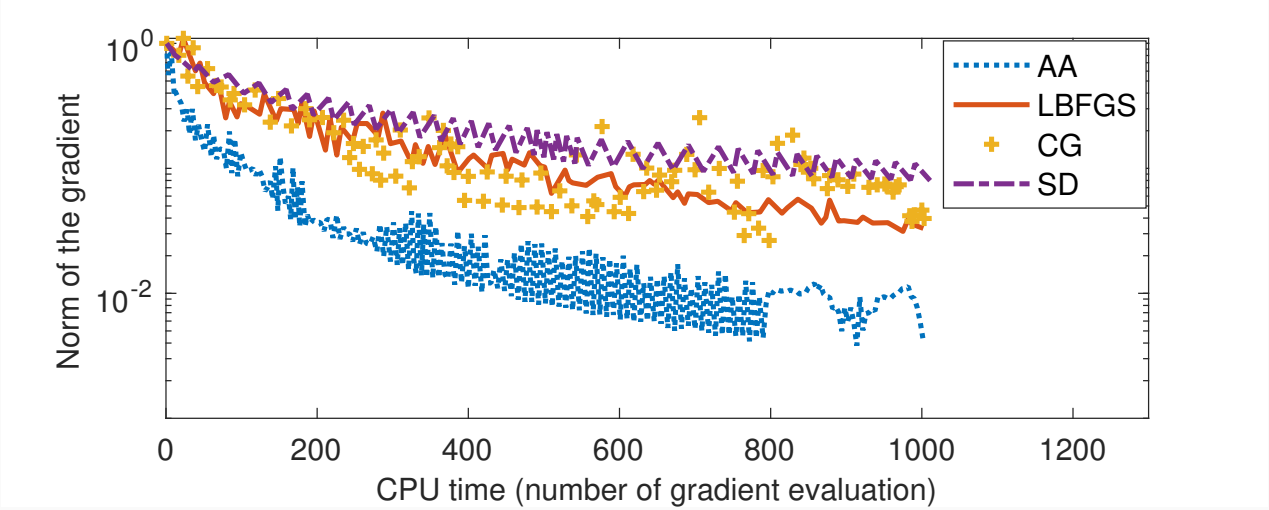
Interesting connections with:

- GMRES ($m = \infty$)
- Multi-secant methods
- Krylov-space methods
- Momentum method
- Superlinear w/o approximating inverse Hessian
- Noncontractive (nonlinear) operator convergence

[D. G. Anderson, 1965]

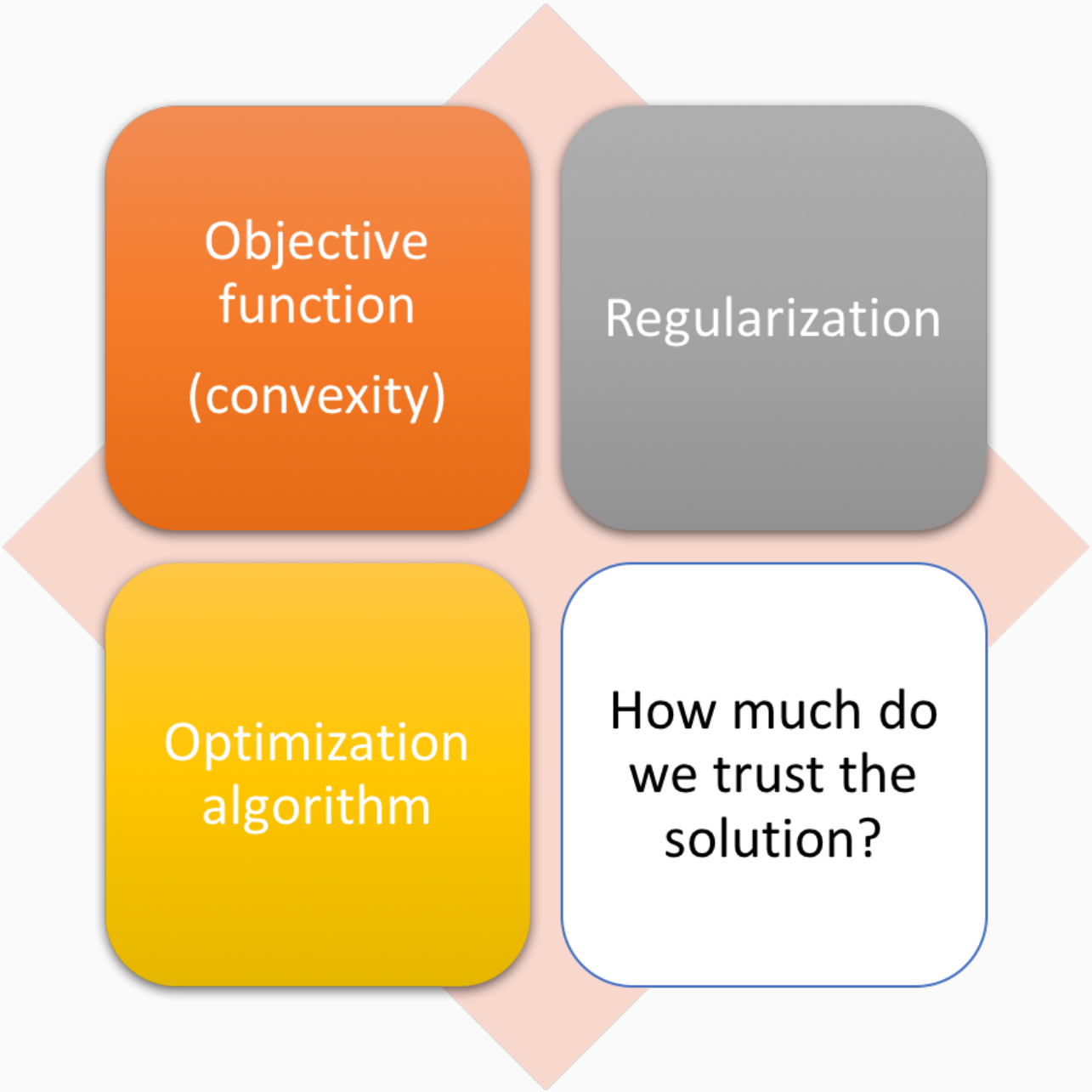
Anderson Acceleration for Seismic Inversion



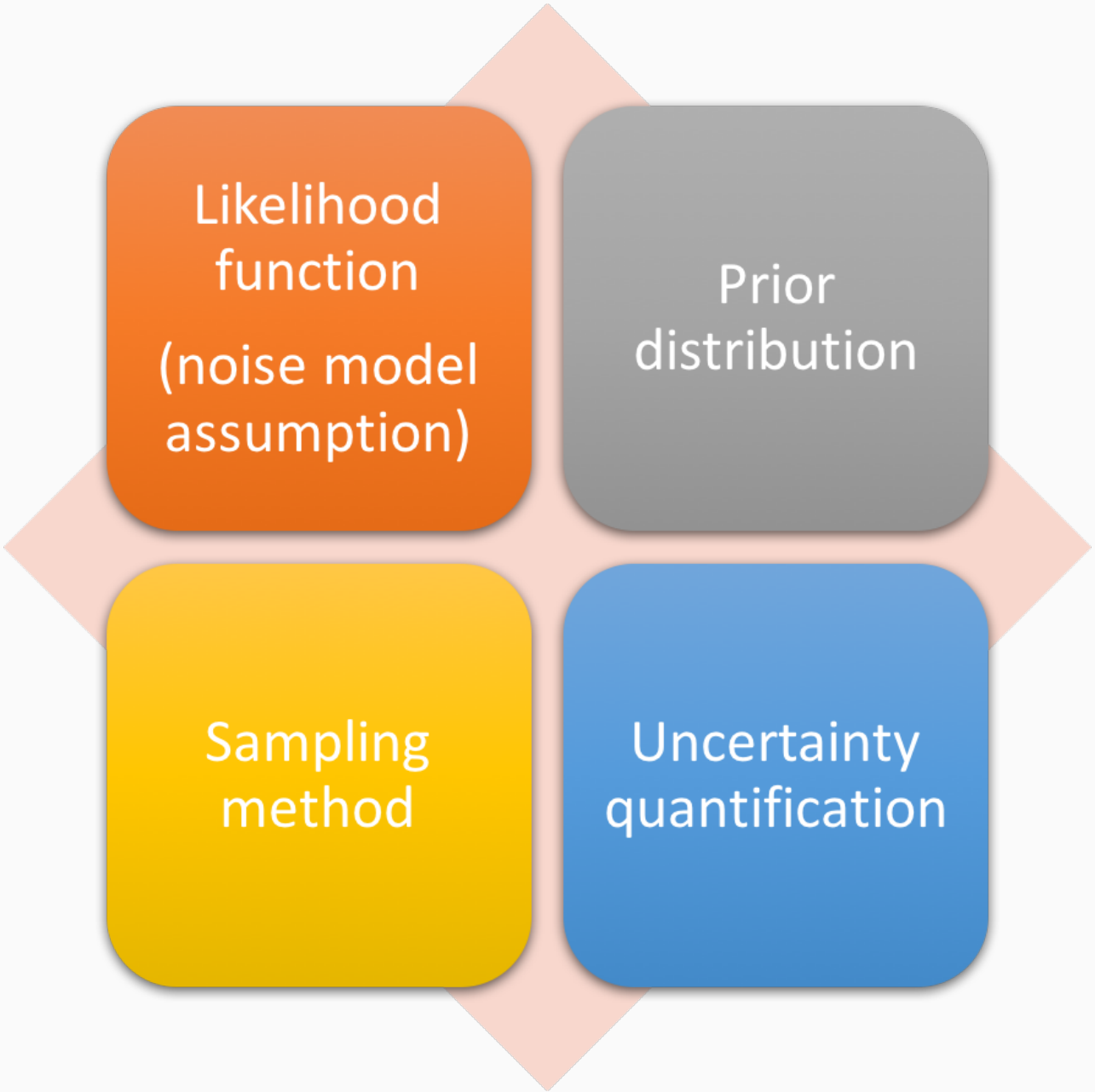


[Y-Townsend-Appeö,2020],[Y,2020]

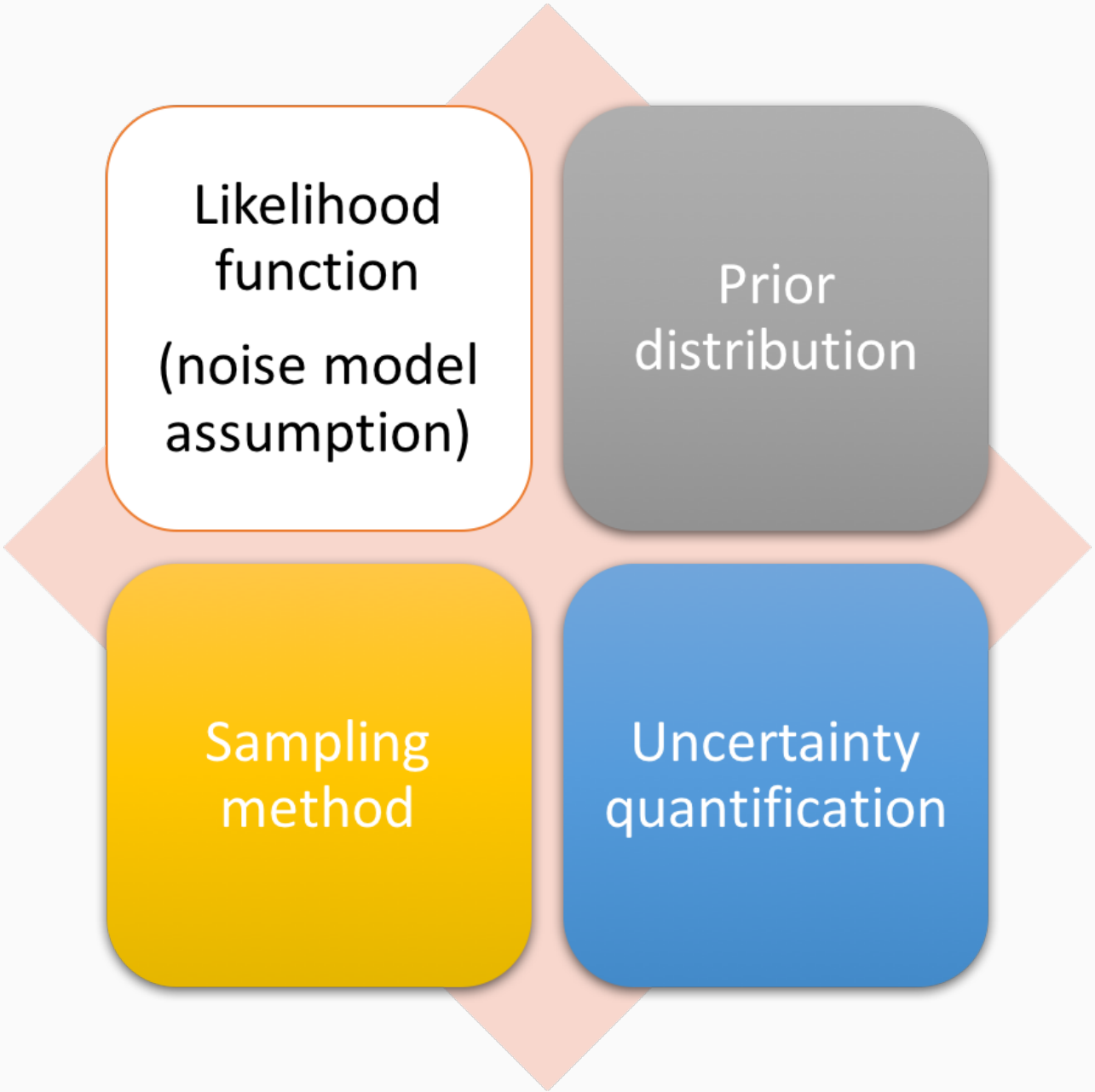
Important Components in the deterministic approach



Important Components in the Bayesian approach



Likelihood Function



Likelihood Function

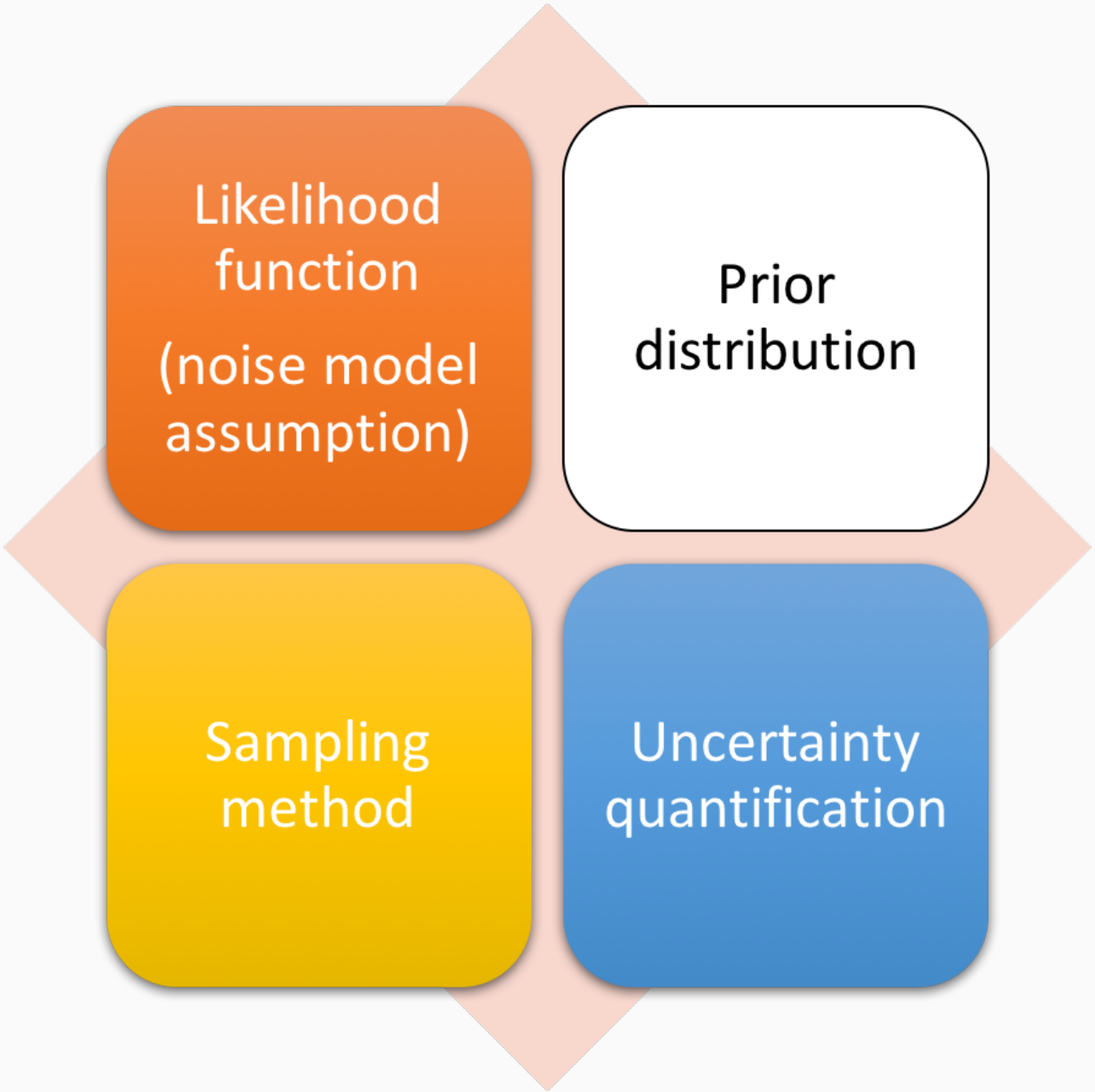
Likelihood function	Noise model assumption
$\int_D \|\mathcal{G}(u)(x,\cdot) - y(x,\cdot)\|^2_{L^2(T)}\, \lambda(\mathrm{d}x)$	$y = \mathcal{G}(u) + \eta,\ \eta \sim N(0, I)$
$\int_D \|\mathcal{G}(u)(x,\cdot) - y(x,\cdot)\|^2_{\dot{H}^{-1}(T)}\, \lambda(\mathrm{d}x)$	$y = \mathcal{G}(u) + \eta,\ \eta \sim N(0, -\Delta_T)$
$\int_D W_2^2\left(\widetilde{\mathcal{G}(u)}(x,\cdot), \widetilde{y}(x,\cdot)\right)\, \lambda(\mathrm{d}x)$	$\widetilde{y} = \eta \cdot \widetilde{\mathcal{G}(u)},\ \eta \| u \sim N(1, \mathcal{L}(u))$
$\int_D \left\|\frac{\mathcal{G}(u)(x,\cdot) - y(x,\cdot)}{(y)(x,\cdot)}\right\|^2_{L^2(T)}\, \lambda(\mathrm{d}x)$	$y = \eta \cdot \mathcal{G}(u),\ 1/\eta \sim N(1, I)$

The W_2 metric can be regarded as asymptotically from **the state-dependent multiplicative noise data model**:

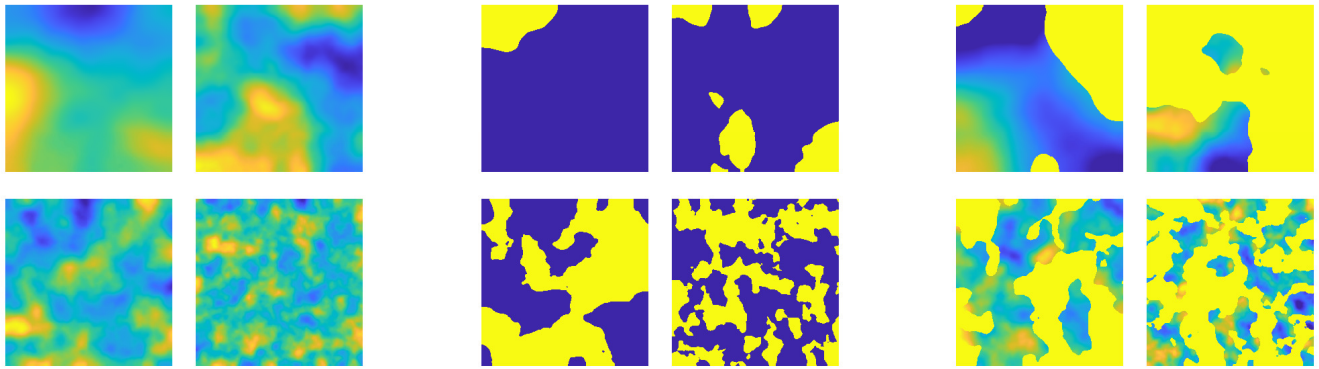
measurement error is proportional to the size of the quantity, and the distribution is affected by the model parameter.

[Dunlop-Y,2020]

Prior Distribution



Level-Set Prior



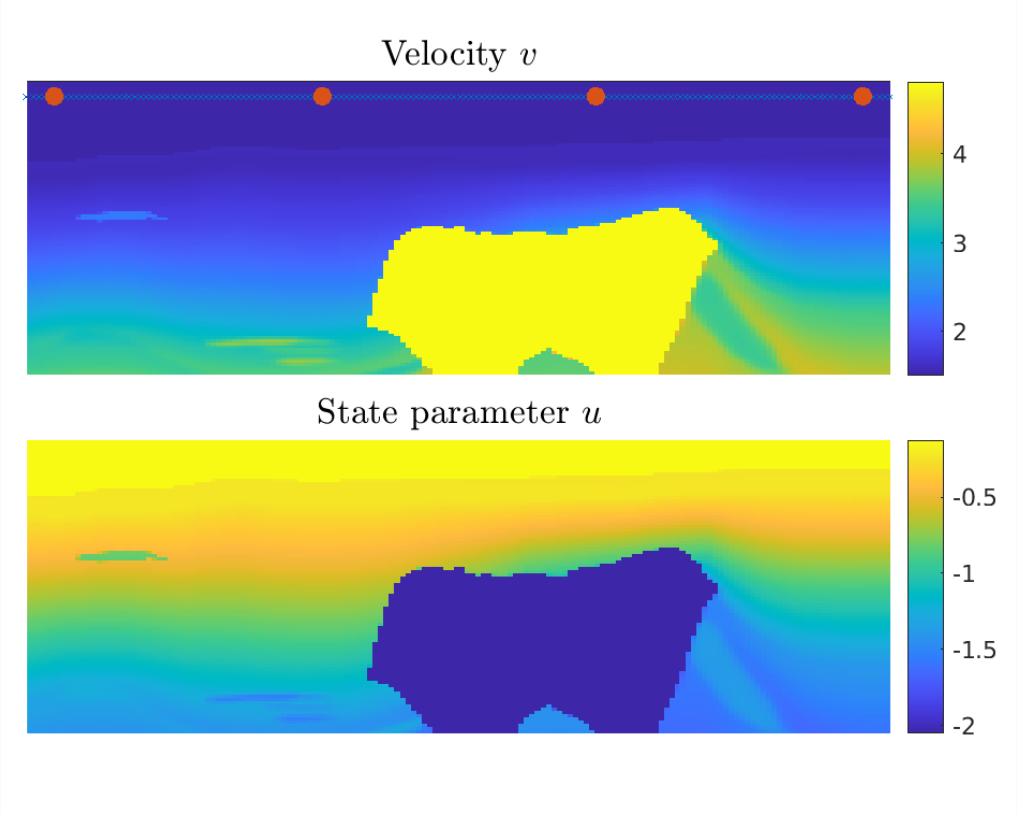
Given $\nu_0 = N(m_1, C_1) \times N(m_2, C_2)$,

$$\pi_0 = F^\sharp \nu_0, \quad F(v, w)(x) = u_+ \mathbb{1}_{v(x)>0} + w(x)\mathbb{1}_{v(x)\leq 0}.$$

We don't have to know u^+ a priori.

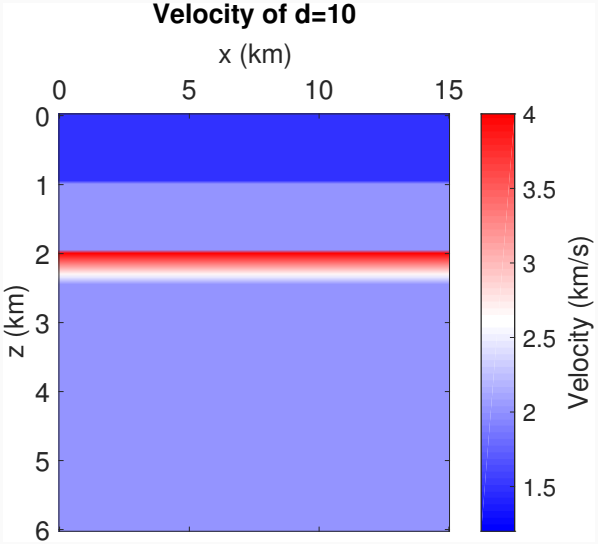
[Dunlop-Iglesias-Stuart, 2016], [Iglesias-Lu-Stuart, 2016]

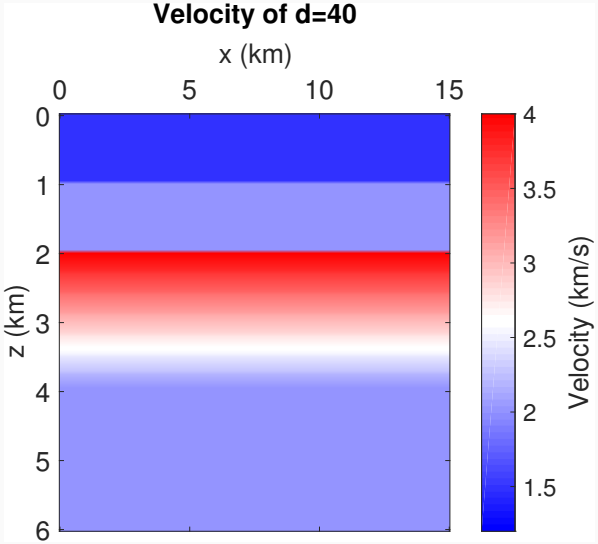
Level-Set Prior for Waveform Inversion

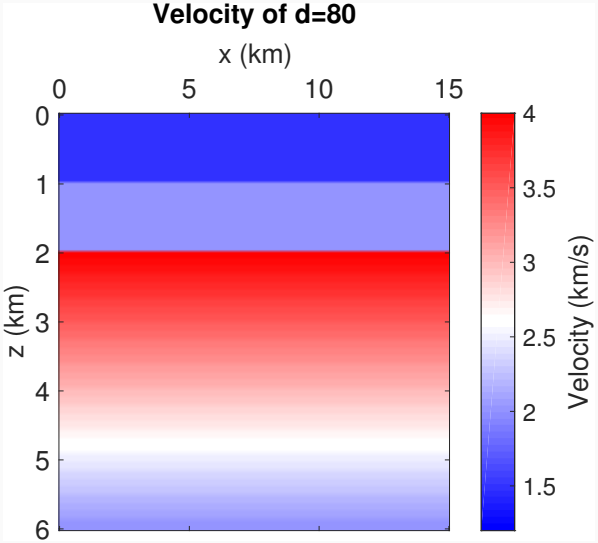


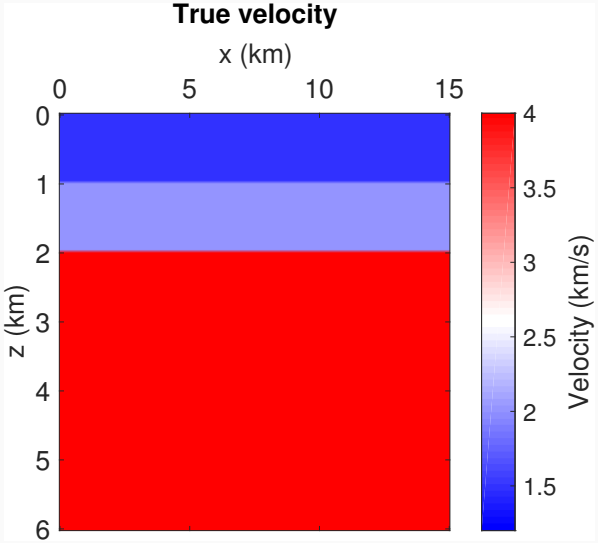
[Dunlop-Y,2020]

Why Model Below Reflector Is Hard to Recover by Reflections?



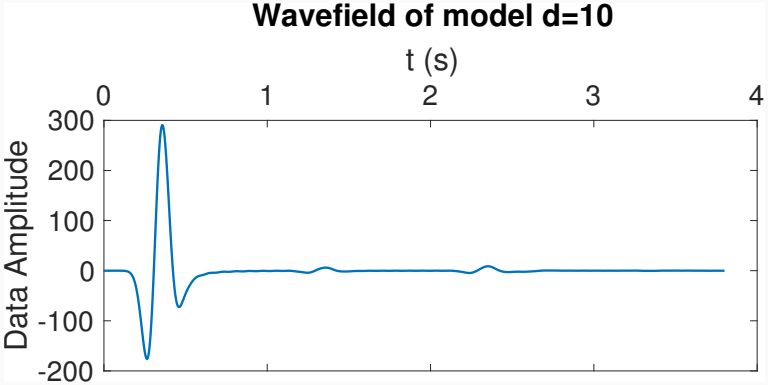


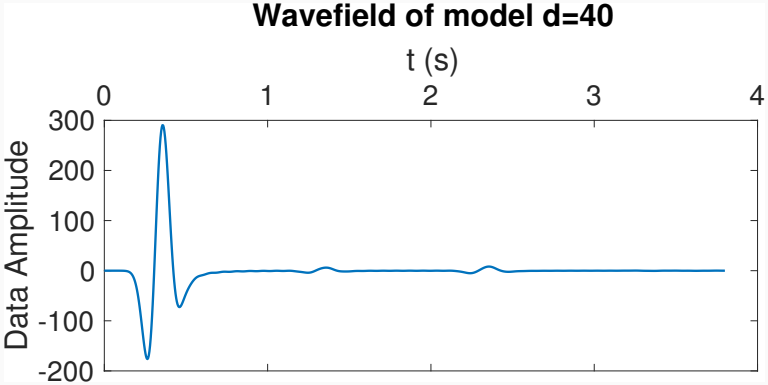


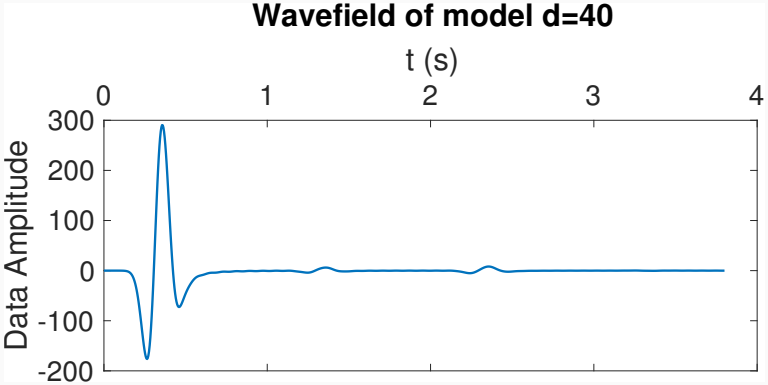


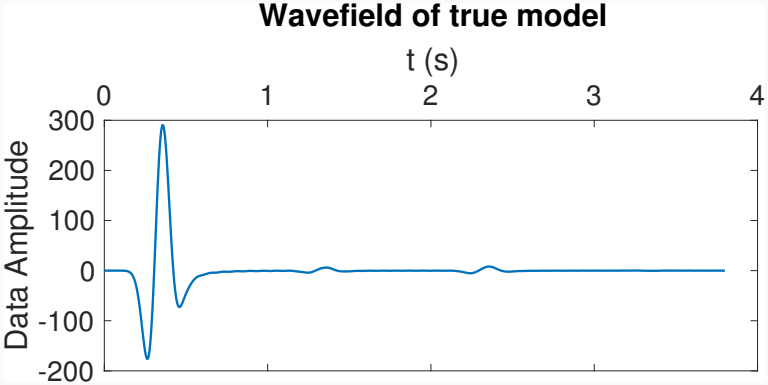
[Y-Engquist,2018]

Why Model Below Reflector Is Hard to Recover by Reflections?



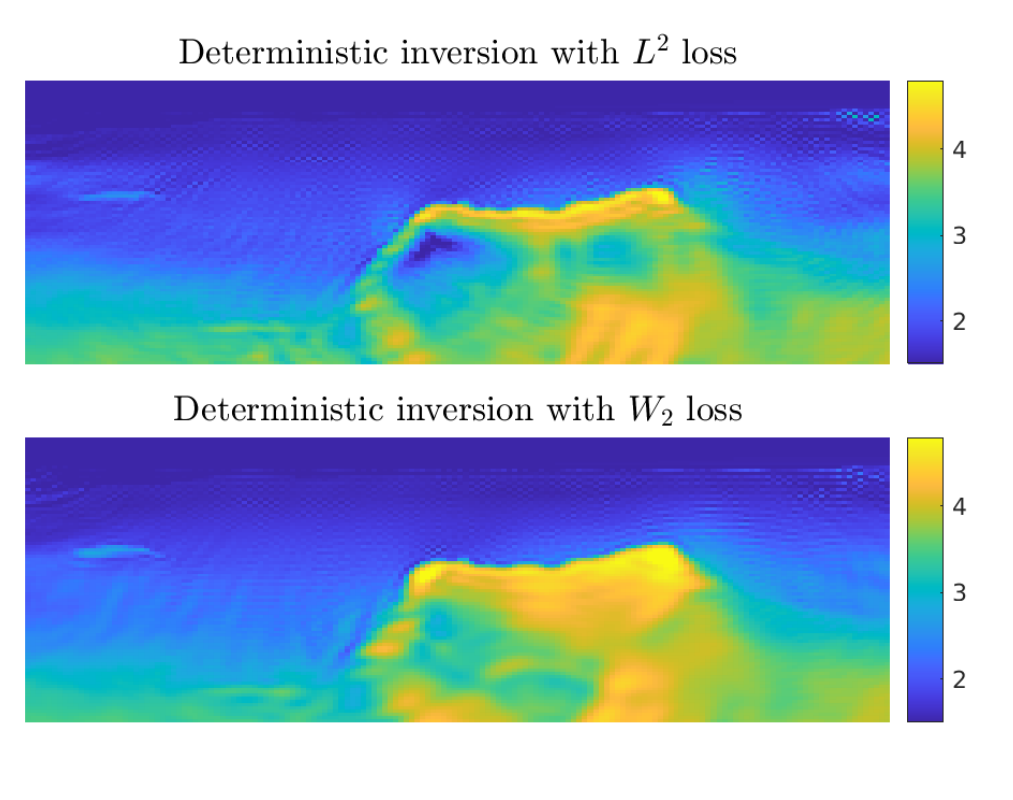






[Y-Engquist,2018]

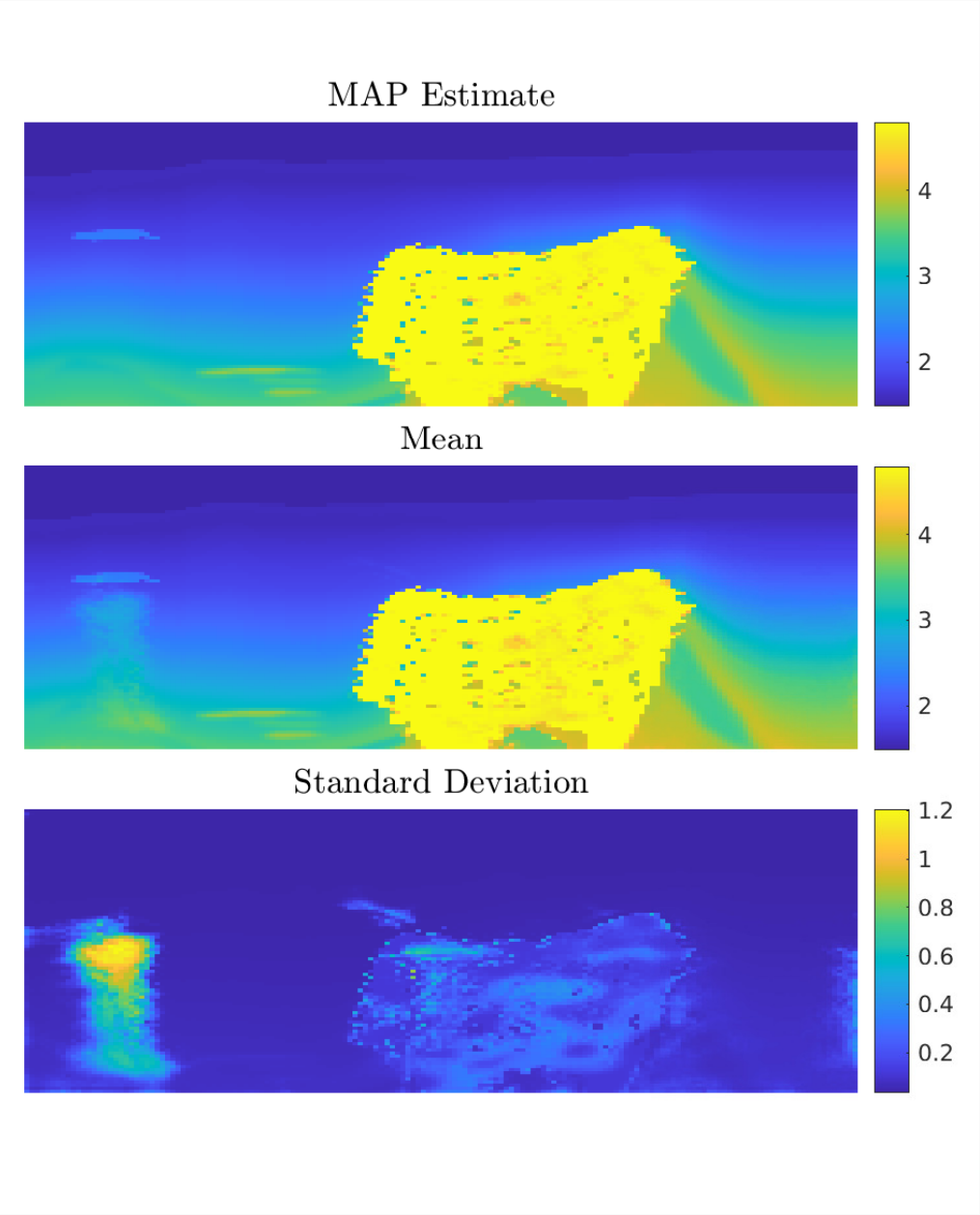
Level-Set Prior for Waveform Inversion



Reconstructions arising from deterministic inversion

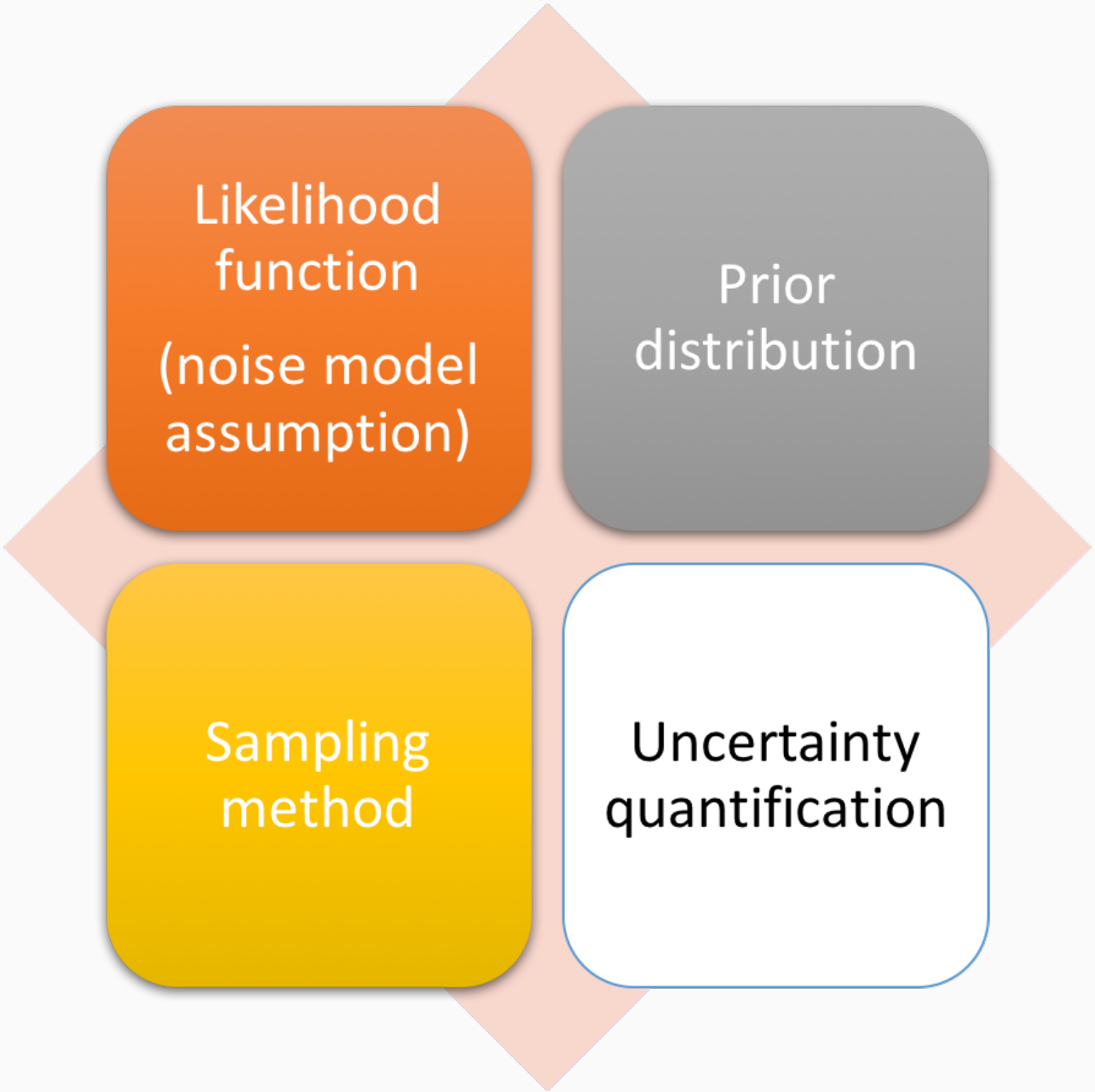
[Dunlop-Y,2020]

Level-Set Prior for Waveform Inversion

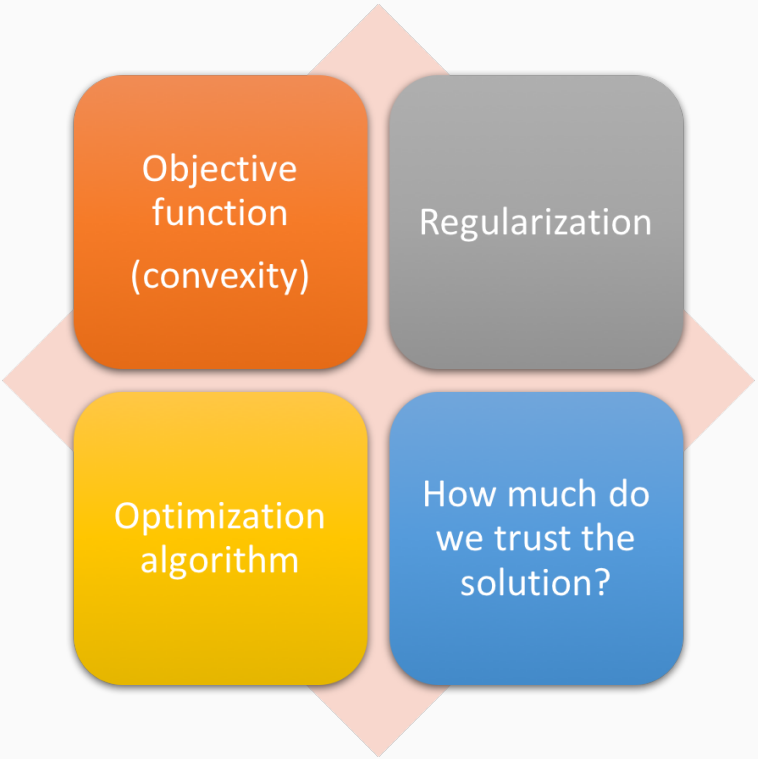


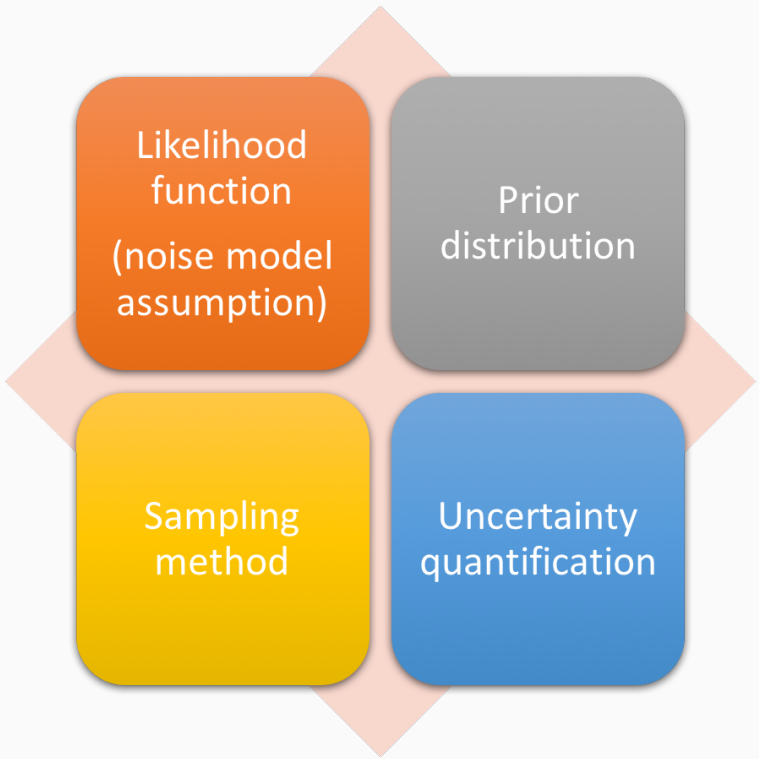
[Dunlop-Y,2020]

Uncertainty Quantification



Deterministic & Bayesian





For Large-Scale inverse problems

Acknowledgment

All my collaborators.

Thank you for the attention!