A Deeper Understanding of the Quadratic Wasserstein Metric in Inverse Data Matching

Yunan Yang (NYU), Matt Dunlop (NYU), Björn Engquist (UT-Austin), Kui Ren (Columbia University)

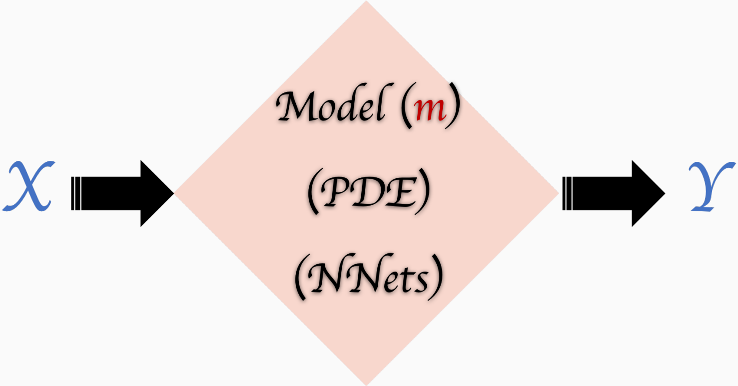
May 5, 2020

https://arxiv.org/abs/1911.06911
https://arxiv.org/abs/2004.03730

This work is partially supported by NSF DMS-1913129.

MSRI Workshop: Optimal Transport And Applications To Machine Learning And Statistics, May 4, 2020 - May 8, 2020

Inverse Data Matching Problem



Inverse data matching problems aim at finding m such that the predicted outputs $(X, Y(m))$ match given measured data (X, Y).

The Role of the Wasserstein Metric

- As an objective function measuring data, $W_p(Y(m), Y)$.
- The functional space for m (new gradient formula).
- Study the convergence of m_k as iteration number k increases (gradient flow, JKO, etc).
- Study the convergence of m_n (as an empirical measure) as (over)parameterization n increases (mean-field limit).
- etc.

The Role of the Wasserstein Metric

- As an objective function measuring data, $W_p(Y(m), Y)$.
- The functional space for m (new gradient formula).
- Study the convergence of m_k as iteration number k increases (gradient flow, JKO, etc).
- Study the convergence of m_n (as an empirical measure) as parameterization n increases (mean-field limit).
- etc.

The Model $F(m)$

Model (m)

(PDE)

(NNets)

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

1. Better Convexity (Optimization Landscape)

Important Components in the Deterministic Approach

Objective function

(convexity)

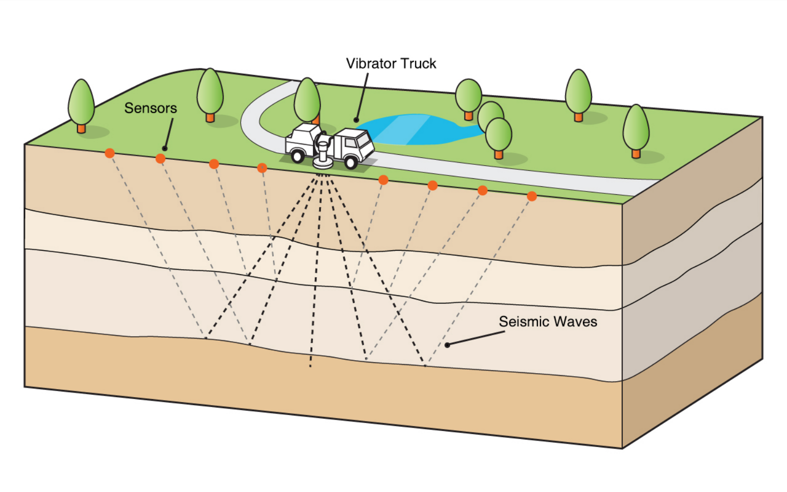
Regularization

Optimization algorithm

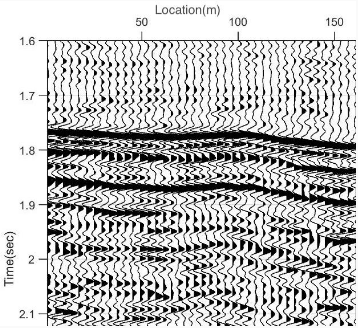
How much do we trust the solution?

Seismic Inversion: Earthquake Source, Hydrocarbons, etc.

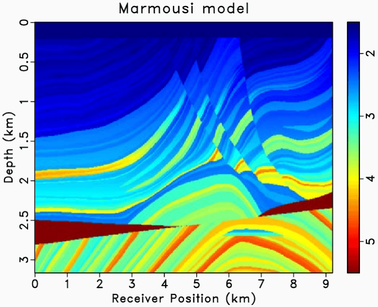
Seismic inversion is one of the inherently more difficult families of large-scale nonlinear inverse problems.



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.

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

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

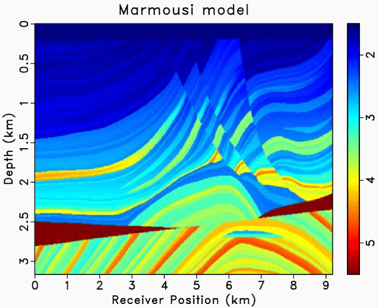
J is an objective function measuring the difference between f and g.

Seismic (Nonlinear) Inversion

Forward Wave Propagation

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

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



m

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]

A PDE-Constrained Optimization

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

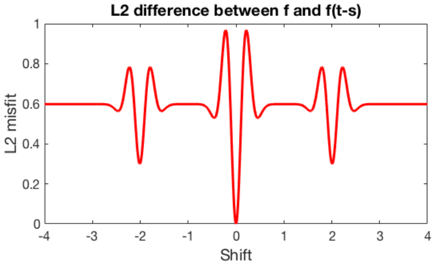
$$J(m) = \frac{1}{2}\sum_r \int |f(x_r, t; m) - g(x_r, t)|^2 \, dt, \qquad (1)$$

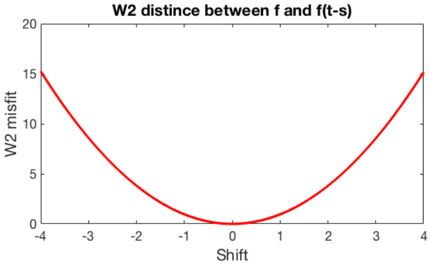
- observed data g,
- simulated data $f(m) = u|_\Gamma$,
- receiver x_r,
- the model parameter m,
- Regularization is often added in (1).

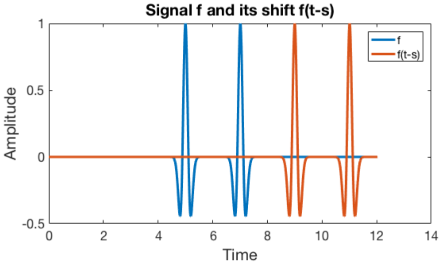
Main Challenges

1. Local minima trapping
2. Sensitive to noise

Motivation of Using the Wasserstein Distance (EMD)







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

The Quadratic Wasserstein Distance

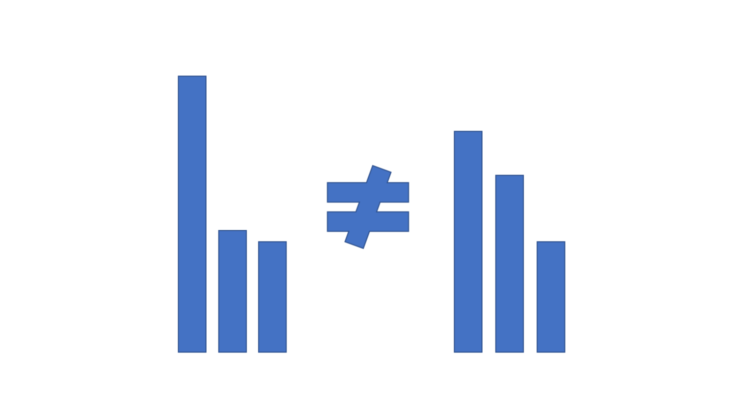
For $f, g \in \mathcal{P}(\Omega)$ ($f, g \geq 0$ and $\int f = \int g = 1$), the quadratic Wasserstein distance is formulated as

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

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

[Monge, 1781]

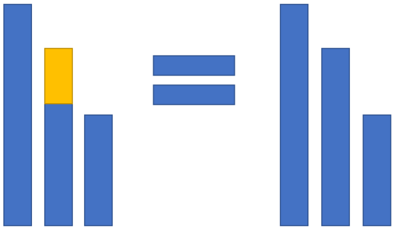
Optimal Transport



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

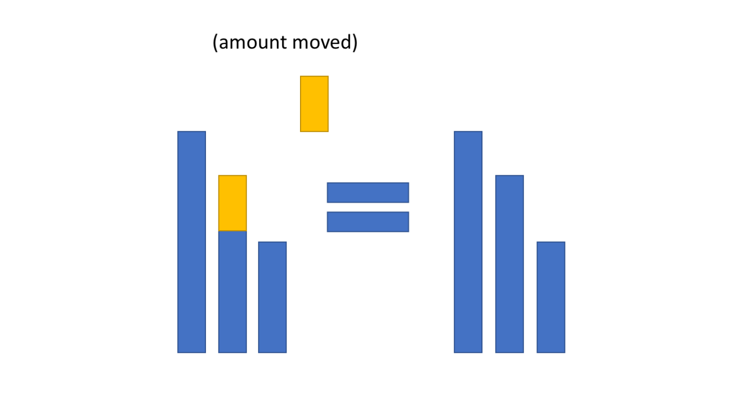
[Monge, 1781]

Optimal Transport



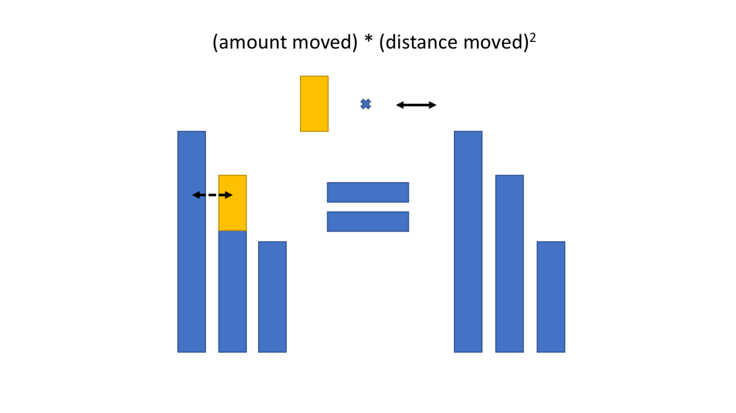
Synthetic data f (left) and observed data g (right)

Optimal Transport



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

Optimal Transport



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

Tackling Nonconvexity

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 $A = \text{diag}(1/\lambda_1, \ldots, 1/\lambda_d)$.
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]

Data Normalization: From Seismic Signal to Probability Density

- Absolute value scaling: $f_2 = |f_1|$
- Square scaling: $f_2 = f_1^2$
- Linear scaling: $f_2 = f_1 + a$
- Exponential scaling: $f_2 = \exp(af_1)$
- Soft-Plus: $f_2 = \log\left(\exp(af_1) + 1\right)$

$$f = \frac{f_2}{\int f_2}$$

[Engquist-Y, 2018],Engquist-Y, 2020]

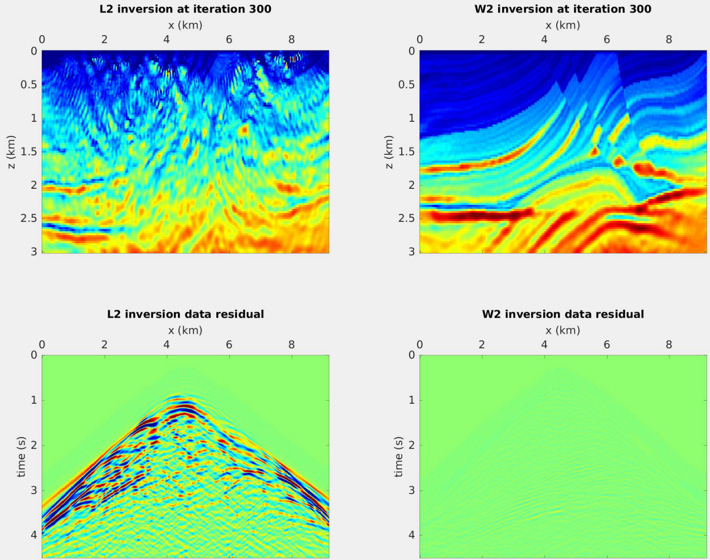
Data Normalization: From Seismic Signal to Probability Density

- Absolute value scaling: $f_2 = |f_1|$
- Square scaling: $f_2 = f_1^2$
- Linear scaling: $f_2 = f_1 + a$
- Exponential scaling: $f_2 = \exp(af_1)$
- Soft-Plus: $f_2 = \log\left(\exp(af_1) + 1\right)$

$$f = \frac{f_2}{\int f_2}$$

[Engquist-Y, 2018],Engquist-Y, 2020]

Tackling Nonconvexity



[Y-Engquist-Sun-Hamfeldt, 2016]

Convexity is not limted to the Wave Applications

- Inversion for Transport in homogeneous flow;
- Reconstruction from projections;
- Deconvolution of highly localized sources;
- Deconvolution from diffusive environment.

[Engquist-Ren-Y, 2019], https://arxiv.org/abs/1911.06911

2. Robustness w.r.t. Noise

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:

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

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

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

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

The Connection with the Weak Norm

Asymptotic Connection [Otto-Villani, 2000]

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 [R. Peyre, 2018]

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 (a smoothing operator) is diagonal in the Fourier domain:

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

We seek the solution by minimizing the objective functional

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

[Engquist-Ren-Y, 2019]

What do we lose?

If we can obtain the solution by direct solve (best-case scenario)

Theorem

Let R_c an approximation to A^{-1} defined through its symbol:

$$\widehat{R}_c(\boldsymbol{\xi}) \sim \begin{cases} \langle \boldsymbol{\xi} \rangle^{\alpha}, & |\boldsymbol{\xi}| < \xi_c \\ 0, & |\boldsymbol{\xi}| > \xi_c \end{cases}.$$

Let $\delta = \|g_\delta - g\|_{\mathcal{H}^s}$, $m_\delta^c := R_c g_\delta$ as the minimizer of $\Phi(m)_{\mathcal{H}^s}$.

$$\|m - m_\delta^c\|_{L^2} \lesssim \|m\|_{\mathcal{H}^\beta}^{\frac{\alpha - s}{\alpha+\beta-s}} \delta^{\frac{\beta}{\alpha+\beta-s}}. \tag{8}$$

Reconstruction based on $\mathcal{H}^s$ has an optimal spatial resolution

$$\varepsilon \sim \delta^{\frac{1}{\alpha+\beta-s}}. \tag{9}$$

[Engquist-Ren-Y, 2019]

What do we gain?

If the noise contains mainly the higher frequency components

The solution at frequency ξ is therefore

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

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

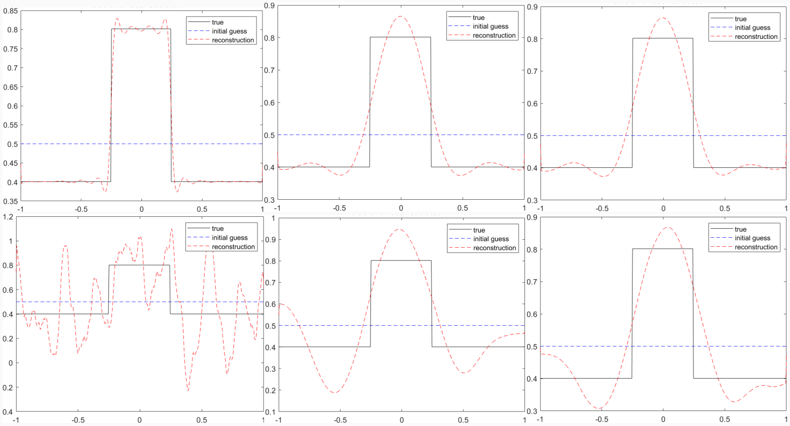
where the operator $(\mathcal{I} - \Delta)^{s/2}$ is defined through the relation

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

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

[Engquist-Ren-Y, 2019]

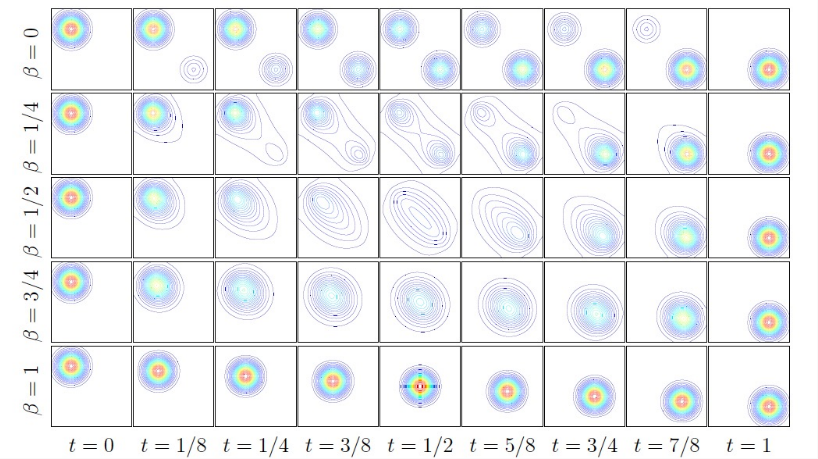
What do we gain and loss?



Deconvolution with the kernel $K_l(x) = \frac{1}{1+|x|}$ with the L^2 (left), $\mathcal{H}^{-1}$ (middle), and W_2 (right) metrics. Top row: with noise-free data; Bottom row: with data containing respectively 2%, 10%, and 10% random noise.

[Engquist-Ren-Y, 2019]

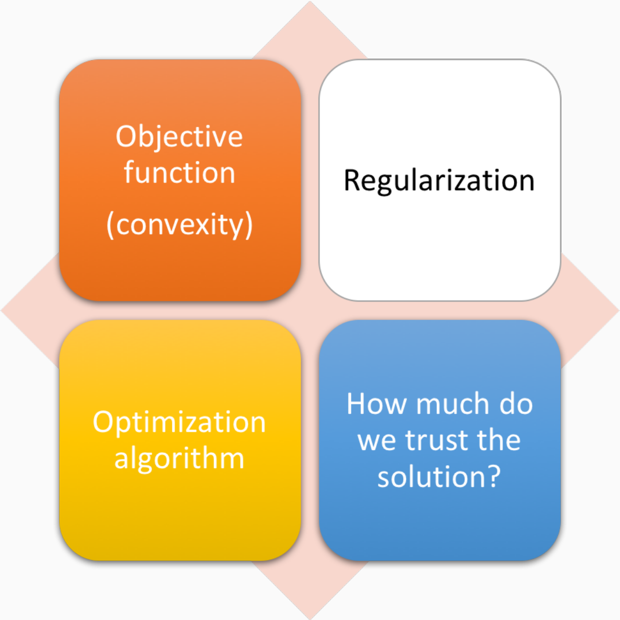
Differences Between W_2 and $\dot{H}^{-1}$ (the gradient flow)



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

[Papadakis-Peyré-Oudet, 2013]

Regularization



Implicit Regularization

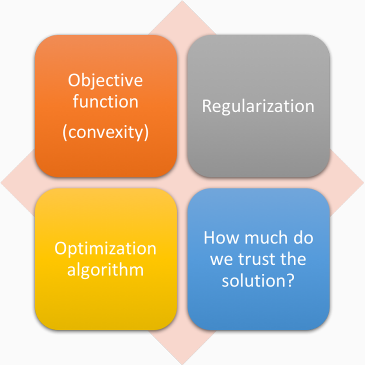
$$m^* = \underset{m}{\operatorname{argmin}} J(m) + R(m)$$

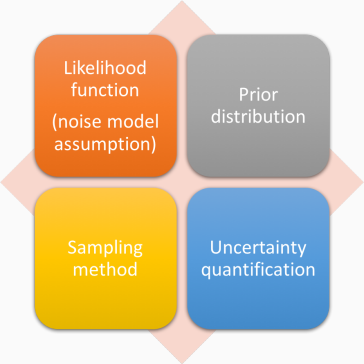
Regularization does not have to be in the form of $R(m)$.

- 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)

Deterministic & Bayesian





For Large-Scale inverse data matching problems

3. Wasserstein Metric as a Likelihood Function in Bayesian Inference

Data Normalization

One problem: $\mathcal{G}(u)$ and y are not probability density functions.
An potential solution: Data Normalization; [Engquist-Y, 2020].

Given a $\sigma : \mathbb{R} \to \mathbb{R}^+$, we define P_σ on functions $y : D \times T \to \mathbb{R}$ as

$$\widetilde{y} = (P_\sigma y)(x,t) = \frac{1}{Z_\sigma(x)} \sigma(y(x,t)), \quad Z_\sigma(x) = \int_T \sigma(y(x,t'))\, \mathrm{d}t'.$$

We only measure the T domain under the Wasserstein metric.

[Dunlop-Y,2020]

W_2 Likelihood Function

$$W_2\left(\widetilde{\mathcal{G}(u)}(x,\cdot),\widetilde{y}(x,\cdot)\right)^2 \approx \left\|\frac{\widetilde{\mathcal{G}(u)}(x,\cdot)-\widetilde{y}(x,\cdot)}{\widetilde{\mathcal{G}(u)}(x,\cdot)}\right\|^2_{\dot{H}^{-1}(\widetilde{\mathcal{G}(u)})}$$

which indicates the following noise model

$$\widetilde{y} = \eta\cdot\widetilde{\mathcal{G}(u)}, \quad \eta|u \sim N(1,\mathcal{L}(u))$$

where $\mathcal{L}(u) : D(\mathcal{L}(u)) \to L^2(D; L^2(T))$ is defined by

$$\mathcal{L}(u)\varphi = -\frac{1}{\underbrace{\widetilde{\mathcal{G}(u)}}_{\rho}}\nabla_T\cdot\left(\underbrace{\widetilde{\mathcal{G}(u)}}_{\rho}\nabla_T\varphi\right),$$

where $D(\mathcal{L}(u)) = \left\{\varphi\in L^2(D;H^2(T))\,|\,\int_T\varphi(\widetilde{\mathcal{G}(u)})\,\mathrm{d}t = 0\right\}$ and ∇_T is the gradient in the T domain.

[Dunlop-Y,2020]

Likelihood Function

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

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

measurement error is proportional to the size of the quantity, and the distribution depends on the model parameter.

[Dunlop-Y,2020]

Theorem (Existence)

Let π_0 be a Borel probability measure on X. Then for any choice $\Phi \in \{\Phi_{L^2}, \Phi_{H^{-1}}, \Phi_{W_2}, \Phi_M\}$,

$$Z_\Phi(y) = \int_X \exp(-\Phi(u; y))\, \pi_0(\mathrm{d}u)$$

is strictly positive and finite, and

$$\pi_\Phi^y(\mathrm{d}u) := \frac{1}{Z_\Phi(y)} \exp\left(-\Phi(u; y)\right)\, \pi_0(\mathrm{d}u)$$

defines a Radon probability measure on X.

[Dunlop-Y,2020]

W_2 Likelihood Function — Benefits

Theorem (Well-posedness)

Choose any $\Phi \in \{\Phi_{L^2}, \Phi_{H^{-1}}, \Phi_{W_2}, \Phi_M\}$. *Under mild assumptions, there exists* $C_\Phi(r) > 0$ *such that for all* $y, y' \in Y$ *with* $\|y\|_{L^\infty(D;L^\infty(T))}, \|y'\|_{L^\infty(D;L^\infty(T))} < r$,

$$d_{\mathrm{H}}(\pi_\Phi^y, \pi_\Phi^{y'}) \leq C_\Phi(r)\|y - y'\|_Y.$$

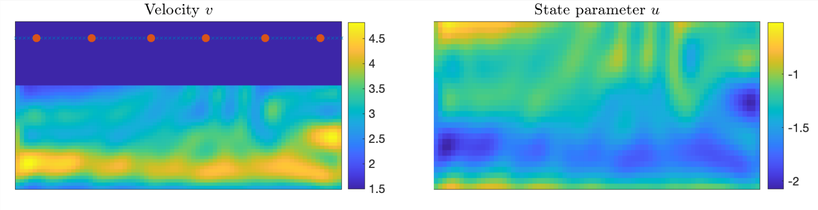
d_{H} *represents the Hellinger distance.*

$$d_{\mathrm{H}}(\pi_{\Phi_{W_2}}^y, \pi_{\Phi_{W_2}}^{y'}) \leq C_{W_2}\|y - y'\|_{H^{-1}}.$$
$$d_{\mathrm{H}}(\pi_{\Phi_{L^2}}^y, \pi_{\Phi_{L^2}}^{y'}) \leq C_{L^2}\|y - y'\|_{L^2}.$$

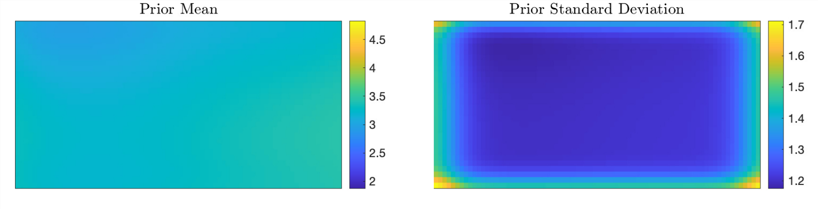
If $y - y' \approx \sin(kx)$, $\|y - y'\|_{H^{-1}} \approx \mathcal{O}(\frac{1}{k})$, while $\|y - y'\|_{L^2} \approx \mathcal{O}(1)$.

[Dunlop-Y,2020]

W_2 Likelihood Function — Example

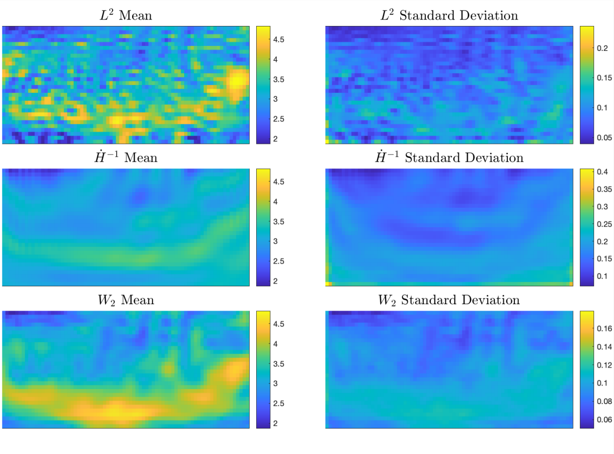


The true continuous velocity field v and the state parameter $u = F^{-1}(1/v^2)$.



[Dunlop-Y,2020] The prior mean $m_0(x)$ and standard deviation.

W_2 Likelihood Function — Example

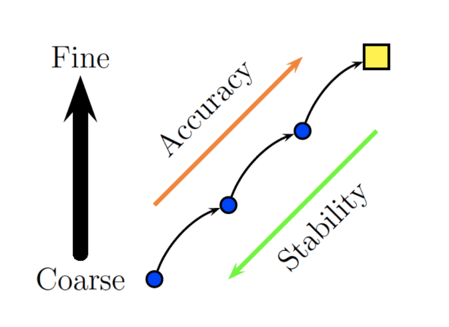


The means (left) and standard deviations (right) of the Laplace approximations.

[Dunlop-Y,2020]

Properties of the Wasserstein Metric in Inverse Data Matching

1. Better convexity (optimization landscape) as an objective function for certain problems.
2. Robust with respect to high-frequency noise.
3. As a likelihood function in Bayesian inference for better stability. (Well-posedness of the posterior is proved.)



From [Qiu, 2013]

Acknowledgment

All my collaborators.

NYU | COURANT
Mathematics

Thank you for the attention!