Application of optimal transport to data assimilation

Nelson Feyeux, Maëlle Nodet, Arthur Vidard

Application of optimal transportation to data assimilation

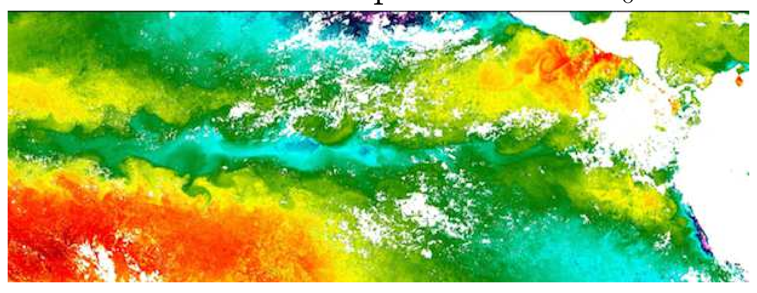
Nelson FEYEUX, Maëlle NODET, Arthur VIDARD

Example of oceanographic data assimilation

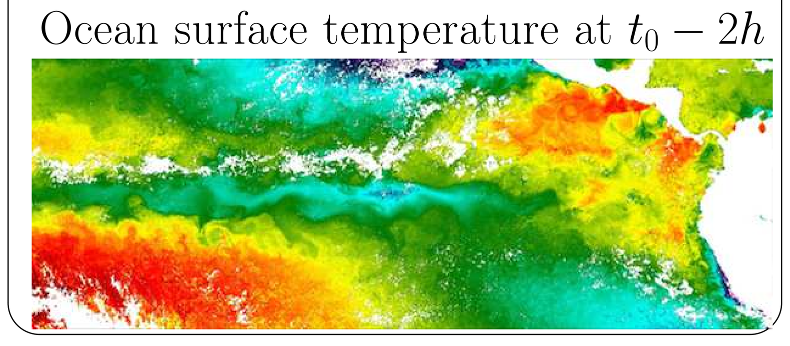
Given a state $\mathbf{x}$ whose evolution is simulated through a model $\mathcal{M}$, **data assimilation** is an inverse problem aiming at reconstructing an initial condition $\mathbf{x}_0$ of $\mathbf{x}$ given only several partial observations $(\mathbf{y}_i^o)_i$ of $\mathbf{x}$. For example, the state $\mathbf{x}(t,x,y,z)$ of the ocean gathers :

- Temperature $T(t,x,y,z)$
- Velocities $\mathbf{u}(t,x,y,z)$
- Salinity $S(t,x,y,z)$
- Water height $h(t,x,y)$

Ocean surface temperature at $t_0 - 4h$



Ocean surface temperature at $t_0 - 2h$



The question is then: What is the complete state of the ocean at $t=0$, $\mathbf{x}_0(x,y,z)$?

In variational data assimilation this is done by minimizing a cost function $\mathcal{J}$ defined as

$$\mathcal{J}(\mathbf{x}_0) = \underbrace{\frac{1}{2}\sum_i d\Big(\mathcal{H}_i(\mathbf{x}_0), \mathbf{y}_i^o\Big)^2}_{\approx d\left(\mathbf{x}_0, \mathbf{x}_0^t\right)^2} + \frac{\gamma}{2} d\Big(\mathbf{x}_0, \mathbf{x}_0^b\Big)^2, \tag{1}$$

It is common for the distance d to be a weighted $\mathcal{L}_2$ distance. Yet, with this $\mathcal{L}_2$ distance, variational data assimilation cannot well cope with displacement errors. By choosing distances d that take into account the datas more in its entirety, we hope get a more realistic variational data assimilation. That is why we are interested in the Wasserstein distance $d = \mathcal{W}_2$.

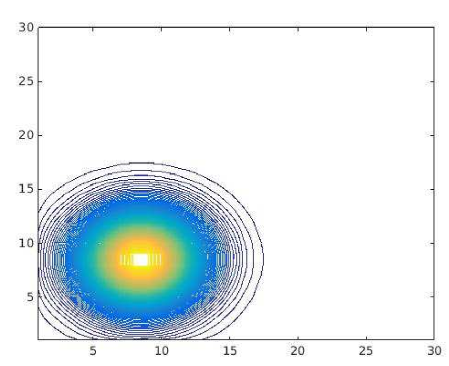
Advantages of optimal transportation in case of displacement error

Knowing two densities $\rho_0(x)$ and $\rho_1(x)$, the **Wasserstein distance** $\mathcal{W}_2(\rho_0,\rho_1)$ is defined as

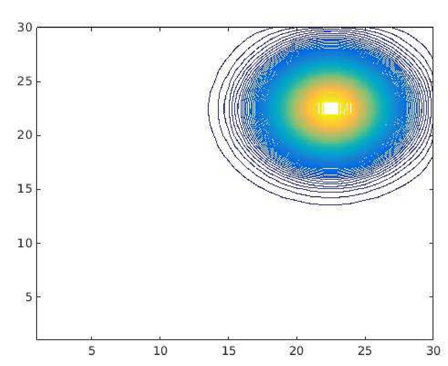
$$\mathcal{W}_2^2(\rho_0,\rho_1) := \inf_{\substack{(\rho(t,x),v(t,x)) \\ \partial_t\rho + \mathrm{div}(\rho v) = 0 \\ \rho(0,x)=\rho_0(x),\, \rho(1,x)=\rho_1(x)}} \iint_{[0,1]\times\Omega} \rho(t,x)|v(t,x)|^2 \,\mathrm{d}t\mathrm{d}x.$$

Example of interpolation using Wasserstein and $\mathcal{L}_2$ distances. (Interpolation in the sense of minimizing $d(\rho,\rho_0)^2 + d(\rho,\rho_1)^2$).

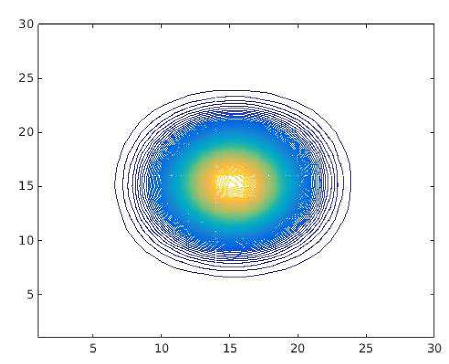
ρ_0



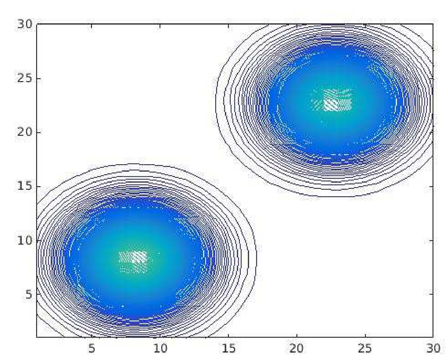
ρ_1



$\mathcal{W}_2$ interpolation



$\mathcal{L}_2$ interpolation



For Wasserstein distance to be defined, one needs $\rho_0 \geq 0$, $\rho_1 \geq 0$ and $\int_\Omega \rho_0 = \int_\Omega \rho_1$.

Gradient descent with the Wasserstein distance

Assuming $\mathbf{x}_0$, $\mathcal{H}_i(\mathbf{x}_0)$ and $\mathbf{y}_i^o$ are probability measures, the cost function $\mathcal{J}_W$ is

$$\mathcal{J}_W(\mathbf{x}_0) = \frac{1}{2}\sum_i \mathcal{W}_2^2\Big(\mathcal{H}_i(\mathbf{x}_0), \mathbf{y}_i^o\Big)$$

The gradient descent algorithm for minimizing $\mathcal{J}_W$ consists in using the following iterative algorithm

$$\mathbf{x}_0^{n+1} = \mathbf{x}_0^n - \alpha\, \mathrm{grad}\mathcal{J}_W(\mathbf{x}_0^n)$$

so that $\mathcal{J}_W(\mathbf{x}_0^{n+1}) < \mathcal{J}_W(\mathbf{x}_0^n)$. To get the gradient, the **differential** must be computed and also an **inner product**. Indeed, $\mathrm{grad}\mathcal{J}(\mathbf{x}_0)$ is such that for all η, $(\eta, \mathrm{grad}\mathcal{J}(\mathbf{x}_0)) = D\mathcal{J}[\mathbf{x}_0].\eta$.

The differential:

Let's differentiate $\mathcal{J}_W$ by computing $\mathcal{J}_W(\mathbf{x}_0 + \epsilon\eta)$:

- $\mathcal{H}_i(\mathbf{x}_0+\epsilon\eta) = \mathcal{H}_i(\mathbf{x}_0) + \epsilon\, L[t_i,\mathbf{x}_0].\eta$ with L the tangent model.
- $\frac{1}{2}\mathcal{W}_2^2(\mathbf{y}+\epsilon\mu, \mathbf{y}') = \epsilon\langle\mu,\psi\rangle_2$ with ψ the **Kantorovich potential** of optimal transportation between $\mathbf{y}$ and $\mathbf{y}'$.

Then,

$$\mathrm{D}\mathcal{J}_W[\mathbf{x}_0].\eta = \sum_i \Big\langle L[t_i,\mathbf{x}_0].\eta\ ,\ \psi_i \Big\rangle_2$$

with ψ_i the Kantorovich potential of the optimal transportation between $\mathcal{H}_i(\mathbf{x}_0)$ and $\mathbf{y}_i^o$.

The inner product:

For convergence reason, instead of using the $\mathcal{L}_2$ inner product, we use the $\mathcal{W}_2$ inner product defined in $\mathbf{x}_0$ by

$$(\eta,\eta') = \int_\Omega \mathbf{x}_0 \nabla\Phi\cdot\nabla\Phi',$$

$$\text{with } \Phi, \Phi' \text{ s.t.} \qquad \eta = -\mathrm{div}(\mathbf{x}_0\nabla\Phi), \qquad \eta' = -\mathrm{div}(\mathbf{x}_0\nabla\Phi')$$

Finally, the gradient of $\mathcal{J}_W$ w.r.t. $\mathcal{W}_2$ inner product is, (L^* is the **adjoint model**),

$$\mathrm{grad}\mathcal{J}_W(\mathbf{x}_0) = -\mathrm{div}\left(\mathbf{x}_0\nabla\left(\sum_i L^*[t_i,\mathbf{x}_0].\psi_i\right)\right).$$

Assimilation of non-probability measure variables

In the case where $\mathcal{H}_i(\mathbf{x}_0)$ and $\mathbf{y}_0$ are probability measures, but *not* $\mathbf{x}_0$, it is not possible to write the background term $\mathcal{J}^b$ with $\mathcal{W}_2$. For example, for the Shallow-water system

$$\begin{cases} \partial_t h + \mathrm{div}(hu) = 0 \\ \partial_t u + (u\cdot\nabla)u + g\nabla h = \mu\Delta u \end{cases} \tag{2}$$

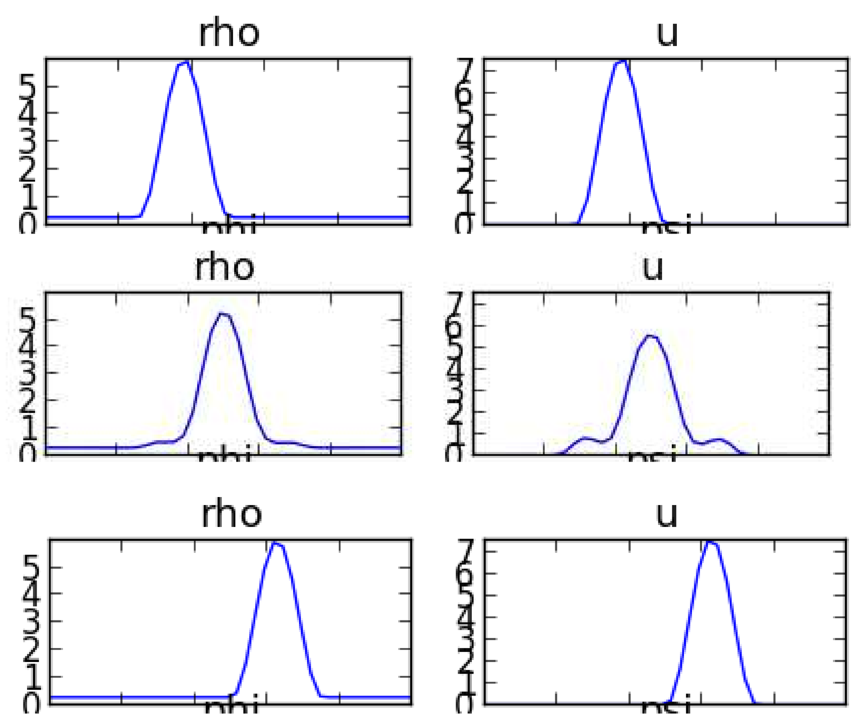
where (h_0,u_0) are to be assimilated using observations of h only. The cost function writes

$$\mathcal{J}(h_0,u_0) = \frac{1}{2}\sum_i \mathcal{W}_2^2\Big(\mathcal{H}_i(h_0,{\color{red}u_0}), h_i^o\Big) + \gamma\mathcal{J}^b(h_0,u_0).$$

As it is impossible to write $\mathcal{W}_2^2(u_0,u_0^b)$, we rather use the following background term

$$J^b(h_0,u_0) = \inf_{\substack{\partial_t\rho + \mathrm{div}(\rho v) = 0 \\ \partial_t u + \mathrm{div}(\rho w) = 0 \\ \rho(0,x)=h_0,\, \rho(1,x)=h_0^b \\ u(0,x)=u_0,\, u(1,x)=u_0^b}} \iint_{[0,1]\times\Omega} (|v|^2+|w|^2)\,\rho\,\mathrm{d}t\mathrm{d}x, \tag{3}$$

Using this, the interpolation of two shifted (h_0,u_0) and (h_1,u_1) is in-between:



The inner product to be chosen for computing the gradient will be

$$\left(\begin{pmatrix}\eta \\ v\end{pmatrix}, \begin{pmatrix}\eta' \\ v'\end{pmatrix}\right) = \int_\Omega h_0\nabla\Phi\cdot\nabla\Phi' + \int_\Omega h_0\nabla\Psi\cdot\nabla\Psi'$$

with Φ, Φ', Ψ, Ψ' such that

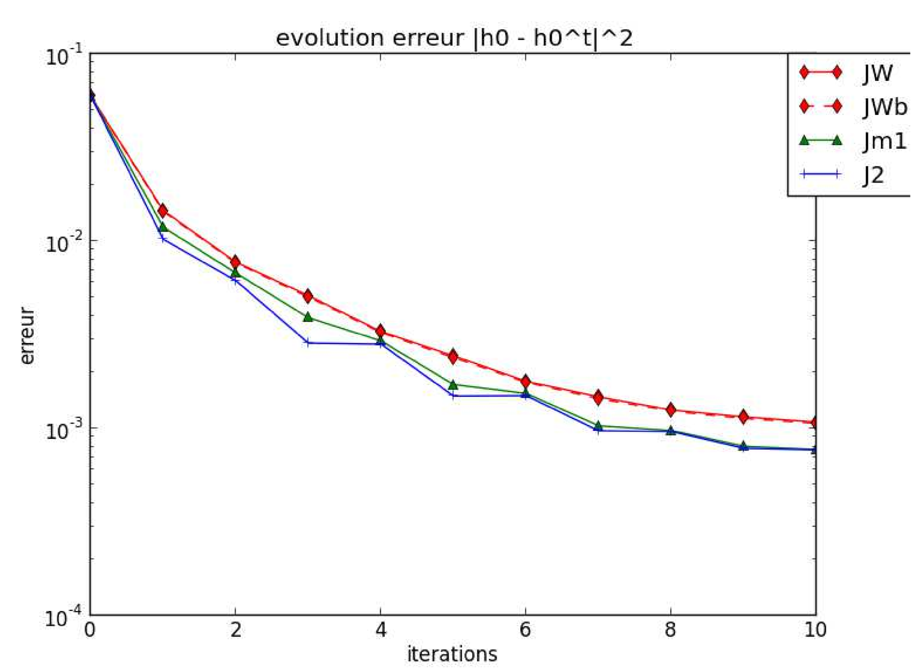
$$\begin{aligned} \eta &= -\mathrm{div}(h_0\nabla\Phi), & \eta' &= -\mathrm{div}(h_0\nabla\Phi') \\ v &= -\mathrm{div}(h_0\nabla\Psi), & v' &= -\mathrm{div}(h_0\nabla\Psi'). \end{aligned}$$

Difficulties of using the Wasserstein distance

- The Wasserstein distance is only defined for probability measures.
- When $\rho_0,\rho_1 \approx 1$, the $\mathcal{W}_2$ interpolation looks like $\mathcal{L}_2$ interpolation...
- When $\mathcal{J}(h_0^n) \to \min_{h_0}\mathcal{J}(h_0)$, one only has $h_0^n \rightharpoonup \arg\min_{h_0}\mathcal{J}(h_0)$.
- The computing time is larger for $\mathcal{W}_2$ than for $\mathcal{L}_2$ [Peyré, Papadakis, Oudet, 2013].

Results and propects

With some tests on the assimilation of (2), we compare the error $\|h_0 - h_0^t\|_2^2$ by using $d = \mathcal{L}_2$ and $d = \mathcal{W}_2$ in the cost function. The behaviors seem correct.



The background term (3) is still to be implemented.

This work has been supported by the region Rhône-Alpes.