Latent space Data Assimilation by using Deep Learning

M. Peyron, A. Fillion, S. Gürol, V. Marchais, S. Gratton, P. Boudier, G. Goret
ANITI/Atos, UFTMIP, CERFACS, ANITI, INPT, NVIDIA, Atos

Email : mathis.peyron@atos.net

Mots Clés : Assimilation de données, Apprentissage profond, espace latent, auto-encodeur, surrogate model, Lorenz 96

Biographie – Diplômé de l'ENSEEIHT en 2020 - spécialité *HPC* & *Big Data* -, j'ai réalisé mon stage de fin d'études au sein de l'ANITI (Artificial and Natural Intelligence Toulouse Institute). Enrichi de cette expérience et dans le *continuum* de mon projet de recherche, je me suis engagé dans une thèse CIFRE avec Atos sur le sujet suivant : "Utilisation conjointe du machine learning et de l'assimilation de données pour un apprentissage plus efficace des dynamiques physiques".

Resumé : Considérons un système dynamique $\boldsymbol{x} \in \mathbb{R}^n$, un modèle $\mathcal{M}$ (entaché d'une erreur $\boldsymbol{\varepsilon}_k$ à l'instant k) tel que $\boldsymbol{x}_{k+1} = \mathcal{M}(\boldsymbol{x}_k) + \boldsymbol{\varepsilon}_k$ et un vecteur d'observations $\boldsymbol{y} \in \mathbb{R}^p$ lui-même erroné de part une acquisition imparfaite : l'assimilation de données consiste à corriger la prédiction modèle $\boldsymbol{x}_k^f$ faite à l'instant k par la prise en compte de la donnée observée $\boldsymbol{y}_k$. La confiance respective qui peut être faite dans la prédiction $\boldsymbol{x}_k^f$ et dans l'observation $\boldsymbol{y}_k$ est centrale pour l'analyse.

La théorie de l'assimilation de données est particulièrement bien adaptée aux phénomènes climatiques pour lesquels d'une part certains processus physiques sont méconnus ou mal représentés, et d'autre part le maillage utilisé lors des simulations ne permet pas de capturer les dynamiques de faible échelle. À ceci, s'ajoute le caractère parfois chaotique de dynamiques propres au phénomène étudié. En météorologie, le système de Lorenz 96 synthétise cette dernière difficulté et fait donc l'objet d'une attention particulière dans la conception d'algorithmes d'assimilation de données.

Bien qu'élégante et au coeur des modélisations actuelles les plus pointues en météorologie, en géosciences ou encore en chimie, l'assimilation de données souffre de deux manquements majeurs : une grande difficulté à traiter des données volumineuses ($n > 10^7, p > 10^6$) et un coût d'application du modèle $\mathcal{M}$ parfois prohibitif. On peut également mentionner l'hypothèse forte selon laquelle les erreurs modèle et d'observation suivent des distributions gaussiennes.

Émergeant depuis une dizaine d'années et aujourd'hui en pleine expansion, l'apprentissage profond répond à ces deux difficultés qui entravent le développement plus efficace de l'assimilation de données. En effet, il rend accessible la recherche d'une représentation latente - sous réserve qu'elle existe - de la dynamique d'intérêt : si les calculs s'avèrent excessivement coûteux dans l'espace physique $\mathbb{R}^n$, la construction d'un espace latent dans $\mathbb{R}^\ell$ (avec $\ell << n$) est alors légitime voire indispensable. Néanmoins, celle-ci n'est envisageable que si le phénomène physique observé possède une dynamique latente de faible dimension ℓ. Dès lors, le recours aux auto-encodeurs qui bénéficient de transformations non-linéaires offre des performances largement accrues comparativement à celles de l'Analyse en Composantes Principales (ACP).

Une nouvelle assimilation de données plus rapide, dite "latente", est alors réalisable à condition de disposer d'un modèle propagatif latent analogue à $\mathcal{M}$: ici encore, on peut tirer parti des réseaux de neurones afin d'apprendre une telle propagation latente. Deux bénéfices majeurs sont alors escomptés : réduction du coût calculatoire et de la sollicitation mémoire d'une part ; l'assimilation de données latente est susceptible de produire des résultats plus précis d'autre part.

Cette démarche qui a été la nôtre est une approche nouvelle dans la résolution de problèmes physiques, à la croisée de l'assimilation de données et de l'apprentissage machine. Étant donné

un système dynamique $\boldsymbol{x} \in \mathbb{R}^n$, nous nous plaçons sous l'hypothèse selon laquelle $\boldsymbol{x}$ possède une représentation latente $\boldsymbol{z} \in \mathbb{R}^\ell$ avec $\ell << n$. Ainsi la variable $\boldsymbol{z}$ constitue l'essentiel de l'information contenue dans $\boldsymbol{x}$ et cette seule connaissance latente suffit à générer une approximation fidèle de la variable physique $\boldsymbol{x}$.

Dans ce contexte, il est possible d'entraîner conjointement deux réseaux de neurones, un auto-encodeur et un surrogate : l'objectif du premier est d'engendrer un espace latent de dimension ℓ dans lequel le second sera à même de réaliser la propagation temporelle latente. Si l'on note $\mathcal{E}$, $\mathcal{D}$ et $\mathcal{S}$ l'encodeur, le décodeur et le surrogate respectivement, alors ce double entrainement est rendu possible au travers d'une fonction de coût adéquate :

$$\boldsymbol{\mathcal{L}}\left(\boldsymbol{x}_{k:k+C}\right) = \frac{1}{C}\sum_{c=1}^{C} \mathrm{MSE}\left(\mathcal{D}\left(\mathcal{E}\left(\boldsymbol{x}_{k+c}\right), \boldsymbol{x}_{k+c}\right)\right) + \rho \times \mathrm{MSE}\left(\mathcal{D}\left(\mathcal{S}^c\left(\mathcal{E}\left(\boldsymbol{x}_k\right)\right)\right), \boldsymbol{x}_{k+c}\right)$$

où MSE désigne la *Mean Square Error*, C le nombre d'itérations d'application du surrogate $\mathcal{S}$ et ρ représente le poids relatif du second terme dans le coût global. La première MSE contraint la performance de l'auto-encodeur - mesure de la qualité de la reconstruction - alors que la seconde impose l'apprentissage d'une propagation temporelle latente. Celle-ci est itérative afin d'assurer la stabilité du réseau : en effet, il y a un forçage sur la capacité du surrogate à produire les éléments $\boldsymbol{x}_{k+1}, \boldsymbol{x}_{k+2}, \ldots, \boldsymbol{x}_{k+C}$ à partir de l'unique donnée $\boldsymbol{x}_k$.

Dès lors, on dispose des outils nécessaires à la réalisation d'une assimilation de données latente. L'algorithme retenu est une adaptation de l'ETKF-Q (Ensemble Transform Kalman Filter with Model error) proposé par [1] : cette nouvelle méthode est ici désignée par le sigle ETKF-Q-L. Il s'agit d'un algorithme ensembliste présentant une étape de correction de l'erreur modèle et s'appuyant sur les éléments entrainés $\mathcal{E}, \mathcal{D}$ et $\mathcal{S}$.

Fortement motivée par les considérations théoriques précédentes, cette nouvelle approche requiert toutefois une justification pratique. Aussi avons-nous considéré un système de Lorenz dit "*augmenté*" construit sur la base des équations de Lorenz en dimension 40 : après avoir généré une telle dynamique de Lorenz sur 500 pas de temps, celle-ci a été multipliée par une matrice orthonormale $O \in \mathbb{R}^{400\times 40}$. Puis, chaque élément de cette nouvelle matrice s'est vu appliquer une transformation polynomiale de degré 3 inversible. Ce système augmenté possède par construction une représentation latente décrite par les équations de Lorenz 96.

L'approche proposée a été comparée à d'autres méthodes plus simples faisant intervenir une ACP ou une régression linéaire. Les critères de comparaison retenus sont la RMSE - *Root Mean Square Error* - et le temps de calcul. L'approche ETKF-Q-L ressort comme étant la plus performante.

Name	RMSE	Inflation	$\boldsymbol{\sigma_Q/\sigma_{Q_t}}$
ETKF-Q	0.194	1.12	0.07
ETKF-Q-P	0.217	1.08	0.1
ETKF-Q-L	**0.168**	**1.004**	$\mathbf{5.10^{-5}}$
PCA-S-P	0.383	1.145	0.5
PCA-S-L	0.383	1.13	0.6
PCA-LinReg-P	0.429	1.24	0.7
PCA-LinReg-L	0.426	1.2	0.9

Name	GPU		CPU	
	Avg. Time	Std	Avg. Time	Std
ETKF-Q	$16.32s$	$0.20s$	$16.23s$	$0.31s$
ETKF-Q-P	$13.60s$	$0.19s$	$17.12s$	$0.84s$
ETKF-Q-L	$\mathbf{6.52s}$	$\mathbf{0.12s}$	$\mathbf{6.89s}$	$\mathbf{0.62s}$
PCA-S-P	$13.22s$	$0.18s$	$12.18s$	$0.31s$
PCA-S-L	$\mathbf{5.93s}$	$\mathbf{0.10s}$	$\mathbf{5.35s}$	$\mathbf{0.16s}$
PCA-LinReg-P	$12.09s$	$0.29s$	$11.62s$	$0.27s$
PCA-LinReg-L	$\mathbf{5.02s}$	$\mathbf{0.12s}$	$\mathbf{4.94s}$	$\mathbf{0.15s}$

Davantage de précisions sur la méthode et les résultats obtenus sont disponibles en ligne [2].

Références

[1] Anthony Fillion, Marc Bocquet, Serge Gratton, Selime Gürol, and Pavel Sakov. An iterative ensemble kalman smoother in presence of additive model error. *SIAM/ASA Journal on Uncertainty Quantification*, 8(1):198–228, 2020.

[2] Mathis Peyron, Anthony Fillion, Selime Gürol, Victor Marchais, Serge Gratton, Pierre Boudier, and Gael Goret. Latent space data assimilation by using deep learning, 2021.