Pour postuler : Patrick.Gallinari@lip6.fr, mathelin@limsi.fr

Thèse de Doctorat

Simulation numérique augmentée par Machine Learning

Sujet de thèse

La simulation numérique représente aujourd'hui un outil indispensable dans la conception des systèmes physiques, grâce notamment au gain qu'elle permet de réaliser sur le coût global de conception. Les ingénieurs disposent principalement de deux sources de modélisation : les méthodes de modélisation physique telles que les équations aux dérivées partielles et leur résolution numérique, et les essais réels réalisés sur le système physique considéré. Ces essais offrent une masse de données parfois considérable qui pourra être exploitée de diverses manières : pour trouver les meilleurs paramètres du modèle physique, pour améliorer la qualité des décisions prises sur la base des méthodes de modélisation numériques, pour remplacer complètement les-dites méthodes.

Ce sujet de thèse se situe au carrefour des méthodes de modélisation de la physique et de celle d'apprentissage automatique. L'objectif est de développer de nouvelles approches permettant l'hybridation des méthodes d'apprentissage statistique avec les méthodes classiques de calcul scientifique, afin de permettre une meilleure prédiction de la simulation tout en réduisant le coût de calcul nécessaire. On se placera dans le cas usuel où on dispose de deux types de données pour représenter un phénomène physique :

- des données qui proviennent d'un, ou de plusieurs, modèle de simulation. Ce modèle peut être peu fidèle à la réalité, mais peu coûteux « en temps de calcul » et donc disponible en nombre très important, ou au contraire coûteux mais relativement fidèle,
- des données qui proviennent des essais réels qui représentent plus « finement » la réalité, mais présentant un coût d'obtention très élevé rendant la taille de ce deuxième type de données très limitée.

Hybridation solveurs PDE et DNN

Une question clé à la base de la construction de systèmes hybrides est la combinaison des deux sources de connaissances : la Physique et les données. Il existe plusieurs cas pour lesquels cette approche est judicieuse. Le paradigme de l'Apprentissage Automatique (*Machine Learning*, ML) pourrait venir en complément des modèles basés sur la physique : ils permettent par exemple de prendre en compte des informations non fournies par le modèle, d'intégrer des informations fournies par des données d'observation comme une alternative aux méthodes d'assimilation. Du point de vue des réseaux neuronaux profonds (*Deep Neural Networks*, DNN), le contexte physique constitue une connaissance préalable qui guide et contraint le processus d'apprentissage. L'équilibre entre les deux sources de connaissances est également extrêmement important et devrait être une question clé du processus d'apprentissage. Nous considérerons ici un processus général où il existe une source de connaissances physiques sous la forme d'Equations aux Dérivées Partielles (EDP) qui représente une information partielle sur le processus sous-jacent, et nous analyserons différents schémas pour compléter cette information avec des modèles basés sur des données pures. Cela soulève plusieurs questions ouvertes comme la caractérisation des solutions de tels systèmes et leur cohérence, la dérivation d'un cadre d'apprentissage nous permettant de combiner des solveurs différentiables et des modules DNN et le développement des algorithmes correspondants. Sur le plan expérimental, nous examinerons différents cas de problèmes où un modèle physique partiellement connu est complété par des modules ML. L'ensemble de ces développements seront étendus au cas où plusieurs modèles de connaissance physique, de qualité de prédiction et de coût variables, sont disponibles. Il s'agira alors de combiner l'ensemble des sources de connaissances, modèles de la Physique et données, dans un cadre multi-niveaux multi-fidélité.

Extrapolation / Robustesse / Modèle sans échelle

Les modèles physiques explicites sont assortis de certaines garanties et peuvent être utilisés dans tout contexte où le modèle est valable. Ce n'est pas le cas pour les DNN, et nous n'avons aucune garantie qu'ils pourront être extrapolés à des situations ou des conditions initiales inconnues. La généralisation est au cœur du ML depuis plusieurs années, mais les cadres développés sont inutiles pour la modélisation de données complexes comme celles considérées ici. L'extrapolation et l'élimination des biais de données ont récemment motivé l'élaboration de nouveaux cadres (Arjovsky *et al*., 2017), mais là encore, ils ne correspondent pas à la complexité des données. Nous proposons ici de nous attaquer au problème en nous inspirant des résultats de stabilité de l'analyse numérique. Les paramètres DNN (poids et

architecture) déterminent les propriétés PDE associées comme l'ordre, le type ou la stabilité. On pourrait proposer différentes architectures DNN inspirées de schémas numériques stables associés à différentes familles d'EDP. Cette analogie motive également l'utilisation de termes de régularisation spécifiques pour augmenter la stabilité. Ceci a été initié par Ruthotto & Haber (2019) qui proposent différents schémas associés à des équations paraboliques ou hyperboliques, Haber *et al.*, 2019. Par ailleurs, on s'intéressera à une formulation Lagrangienne en coordonnées généralisées de façon à prendre en compte naturellement les invariants physiques du système dans l'apprentissage du DNN, *cf*. Cranmer *et al*., 2020.

Quantification de l'incertitude

Enfin, il faut pouvoir quantifier la confiance que l'on peut avoir dans les prédictions du modèle complet (hybridé), un aspect crucial pour les applications. On réfléchira à la quantification de l'incertitude de la prédiction, étant donnés les modèles physiques sous-jacents, un modèle d'erreur et de bruit sur les données d'entraînement et l'incertitude des entrées.

Développement et validation

Comme cas de développement, nous considérons un problème physique simple décrit par l'équation de Kuramoto-Sivashinsky (KS) (Kuramoto, 1980) une PDE déterministe initialement proposée pour décrire la dynamique des fronts de flamme en combustion. Cette équation peut présenter des régimes de solution 2-D (1 dimension d'espace, 1 dimension de temps) aussi bien périodiques que chaotiques, en fonction de la valeur de ses paramètres. Les développements seront ensuite étendus au cas de PDE stochastiques avec le modèle de Kardar–Parisi–Zhang, Kardhar (2007).

Les travaux de la thèse seront par la suite appliqués sur un cas d'usage industriel de simulation en mécanique des fluides (écoulement turbulent). Les modèles de simulation ainsi que les données des essais réels seront fournis comme base pour l'application des approches développées dans le cadre de la thèse et déjà validées sur le cas de développement (KS).

Références

ARJOVSKY M., CHINTALA S. & BOTTOU L., *Wasserstein Generative Adversarial Networks*, Proceedings of the 34th International Conference on Machine Learning, PMLR 70:214-223, 2017.
CRANMER M., GREYDANUS S., HOYER S., BATTAGLIA P., SPERGEL D. & HO S., *Lagrangian Neural Networks*, ArXiV 2003.04630, 2020.
HABER E., LENSINK K., TRIESTER E. & RUTHOTTO L., *IMEXnet: A Forward Stable Deep Neural Network*, ICML, 2019.
KARDAR M., *Statistical physics of fields*, Cambridge University Press, 2007.
KURAMOTO Y., *Instability and turbulence of wavefronts in reaction-diffusion systems*, Progress of Theoretical Physics, **63**(6), p. 1885–1903, 1980.
RUTHOTTO L. & HABER E., *Deep Neural Networks Motivated by Partial Differential Equations*, J. Math. Imaging Vis., 2019

Contexte de la thèse

Au sein de l'Institut de Recherche Technologique SystemX, situé au cœur du campus scientifique d'excellence mondiale de Paris-Saclay, vous prendrez une part active au développement d'un centre de recherche technologique de niveau international dans le domaine de l'ingénierie numérique des systèmes. Adossé aux meilleurs organismes de recherche français du domaine et constitué par des équipes mixtes d'industriels et d'académiques, ce centre a pour mission de générer de nouvelles connaissances et solutions technologiques en s'appuyant sur les percées de l'ingénierie numérique et de diffuser ses compétences dans tous les secteurs économiques.

Plus particulièrement au sein de l'IRT SystemX, le doctorant sera rattaché à l'axe scientifique « Sciences des données et Interaction ». Le sujet de thèse a été défini par le consortium réuni dans le cadre du projet « Hybridation Simulation-Apprentissage » (HSA).

La direction de la thèse sera assurée par Patrick Gallinari du Laboratoire d'Informatique de Paris 6 (LIP6) et Lionel Mathelin du Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI) à Saclay. La thèse sera inscrite à l'école doctorale STIC de Paris-Saclay.

Le poste est basé à l'IRT SystemX - Gif sur Yvette.

Profil recherché

Étudiant BAC +5 (Ingénieur et/ou Master), dans les domaines Mathématiques appliquées / Informatique / Apprentissage statistique

Connaissances et savoir-faire essentiels :

Maîtrise des méthodes d'apprentissage statistique, d'optimisation et de calcul scientifique.

Bonne maîtrise de Python - la connaissance d'une librairie d'apprentissage profond sera un plus certain.

Le goût pour les expérimentations numériques, et l'analyse détaillée et en profondeur des résultats de ses expériences est essentiel.

Qualités professionnelles :

- Capacités d'analyse, forte autonomie et esprit d'équipe ;
- Organisé et rigoureux ;
- Aptitude à communiquer aussi bien à l'oral qu'à l'écrit en français et en anglais.

- **Pour candidater**, merci d'envoyer à Patrick.Gallinari@lip6.fr et mathelin@limsi.fr
- un curriculum vitae détaillé,
- le relevé de notes de master 2 ou de dernière année d'école d'ingénieur
- une lettre de recommandation du responsable de master.