Distributed Deep Learning for
in situ processing of Large-Scale Physical Simulations
Internship + PhD offer

Supervision:

- Thomas Moreau, MIND, Inria (thomas.moreau@inria.fr)
- Hadrien Hendrikx, Thoth, Inria (hadrien.hendrikx@inria.fr)

Context. While artificial intelligence is growing at a fast pace, the bulk of the world's computing power remains targeted at modeling and predicting physical phenomena, such as climate models, weather forecasting, or nuclear physics. These simulations are run on highly parallel supercomputers on which both the hardware and the software are optimized for the task at hand. While the computing power of each processing unit is still increasing, the communication networks and the storage capabilities in these clusters do not follow such fast trends. As a result, computing nodes produce outputs faster than what can be stored or sent to process elsewhere: These **simulations are IO bound**.

To reduce the communication burden, a promising venue is **in situ computations**, meaning that most of the data is processed locally by the nodes, and only meaningful aggregates are stored or sent over the network. However, this is a difficult problem since meaningful information for the global simulation depends on the other nodes' output. This internship aims to **leverage machine learning techniques to bypass IO bottlenecks in the context of physics simulation on high-performance computing (HPC) clusters**. Thus, this work is placed in a broader "Machine Learning for Science" context, which aims to use ML to solve key problems in traditional sciences. Machine learning techniques will identify relevant information, detect anomalies, or compress the data for specific analysis. Driven by the impossibility of storing the data for post-mortem model training, we will investigate distributed learning techniques to learn machine learning models as the simulation runs.

Methods. To focus on an unsupervised task that can be easily evaluated, we will start the work focusing on dimensionality reduction and data-driven compression techniques. These methods are one way to impact this communication vs. computation trade-off. To fit the requirements imposed by the HPC setting, we will consider distributed incremental dimensionality reduction methods, ranging from PCA [7] to autoencoders [5]. The key complexity here is to efficiently train deep learning models using streaming data split over many computing nodes. An important consideration in our context is that, unlike classical data stream, the data is not *i.i.d.* on the nodes, but stems from the domain partitioning imposed by the physic of the problem and by the temporal dynamic imposed by the simulation. The two main objectives are the following:

- **Benchmark existing methods:** This will require a thorough state-of-the-art review, as well as defining the relevant metrics for evaluating data compression in physics simulations (communication/computation time/cost, quality of the solution...). The benchmark will be realized with `benchopt` [2] and will benefit from the distributed coding expertise of both supervisors.
- **Designing new efficient methods:** To account for the structure of physic simulations, we propose to investigate how to efficiently leverage the inter-node communication to improve over existing distributed optimization methods, with a first focus on PCA [4, 3]. Tight convergence analyses of the proposed methods will be investigated.

This project will investigate several advanced compression methods, in particular with spatial compression [1], mesh-based wavelets [6], or auto-encoders [5].

Skills developed by the candidate. One key element of this project is access to state-of-the-art hardware and real data. By the end of the PhD, the candidate will have developed a strong experience in leveraging state-of-the-art deep learning models to accelerate scientific simulations and will have gained both a strong practical experience and a thorough theoretical understanding of the methods. Given the increasingly important role of machine learning in scientific simulation and of HPC technologies in large model training, this PhD will give the candidate a relevant background for research positions in both industry and academia.

Environment. The internship will take place at Inria Grenoble, in the Thoth team for at most 6 months (April-September 2024). This is a large team focused on machine learning, and in particular computer vision. Particular topics of interest include visual comprehension, hyperspectral imaging, numerical and parallel optimization, and unsupervised learning. A particular emphasis is put on interdisciplinary projects. The internship will include frequent visits to the MIND team, at Inria Saclay. Full-time in Saclay can also be discussed, but the preferred location is Inria Grenoble. The two supervisors are young Inria researchers, with a strong track record in optimization and machine learning.

This project also takes place in the PEPR NumPEx, an initiative to improve the use of supercomputers for physical simulations. This internship provides the unique opportunity to discuss with scientists from other fields and to improve their workflows through AI research. Interaction with scientists developing computational simulations in various fields will be encouraged, in particular with the Gysela code, which is part of the ITER project.

Requirements. We seek candidates strongly motivated by challenging research topics in machine learning for science. Applicants should have a strong mathematical background with knowledge of numerical optimization and machine learning. With regards to software engineering, proficiency in `Python` is expected and preliminary experience in a distributed computation library is a plus. Note that funding for PhD is already available, so we are seeking a strongly motivated candidate who would like to continue on this topic with a PhD.

References

[1] Milan Klöwer, Miha Razinger, Juan J. Dominguez, Peter D. Düben, and Tim N. Palmer. Compressing atmospheric data into its real information content. *Nature Computational Science*, 1(11):713–724, November 2021.

[2] Thomas Moreau, Mathurin Massias, Alexandre Gramfort, and benchopt contributors. Benchopt: Reproducible, efficient and collaborative optimization benchmarks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, New-Orleans, LA, USA, November 2022. Curran Associates, Inc.

[3] Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal convergence rates for convex distributed optimization in networks. *Journal of Machine Learning Research*, 20:1–31, 2019.

[4] Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pages 1000–1008. PMLR, 2014.

[5] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy Image Compression with Compressive Autoencoders. In *International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.

[6] S. Valette and R. Prost. Wavelet-based progressive compression scheme for triangle meshes: Wavemesh. *IEEE Transactions on Visualization and Computer Graphics*, 10(2):123–129, March 2004.

[7] Xiaolu Wang, Yuchen Jiao, Hoi-To Wai, and Yuantao Gu. Incremental aggregated riemannian gradient method for distributed pca. In *International Conference on Artificial Intelligence and Statistics*, pages 7492–7510. PMLR, 2023.