Inria

Offer #2024-07754

PhD Position F/M Distributed Dimensionality Reduction for Large-Scale Physical Simulations

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Fonction : PhD Position

About the research centre or Inria department

The Inria Grenoble research center groups together almost 600 people in 23 research teams and 7 research support departments.

Staff is present on three campuses in Grenoble, in close collaboration with other research and higher education institutions (University Grenoble Alpes, CNRS, CEA, INRAE, …), but also with key economic players in the area.

Inria Grenoble is active in the fields of high-performance computing, verification and embedded systems, modeling of the environment at multiple levels, and data science and artificial intelligence. The center is a top-level scientific institute with an extensive network of international collaborations in Europe and the rest of the world.

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 **is 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 in general since meaningful information for the global simulation depends on the other nodes' output. The goal of this PhD is to **leverage machine learning techniques to bypass IO bottlenecks in the context of physics simulation on high-performance computing (HPC) clusters.** This work is thus placed in a broader ``Machine Learning for Science'' context, which aims at using ML to solve key problems arising in traditional sciences.

More specifically, we will focus on distributed dimensionality reduction techniques, which critically reduce the communication and storage needed to retain most of the information.

Environment. The PhD will take place at Inria Grenoble, in the Thoth team. 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 PhD will include frequent visits to the MIND team, at Inria Saclay. 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. The results from the PhD will thus be integrated into the software stack for these applications. This PhD thus provides the unique opportunity to discuss with scientists from other fields and to improve their workflows through IA 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.

Assignment

The project will first focus on dimensionality reduction techniques, and in particular the standard PCA method. To fit the requirements imposed by the HPC setting, we will consider distributed incremental PCA methods [8], that work with 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 physics of the problem. 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 [3] 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 investi-
   gate how to efficiently leverage the inter-node communication to improve existing distributed PCA methods [5, 4]. The convergence of the proposed methods will be analyzed and we will provide tight convergence bounds.

While the initial focus will be on PCA, more advanced compression methods will be considered throughout the project, in particular with spatial compression [1, 2], mesh-based wavelets [7], or auto-encoders [6].

References
[1] Andrés Hoyos-Idrobo, Gaël Varoquaux, Jonas Kahn, and Bertrand Thirion. Recursive nearest agglomeration (ReNA): Fast clustering for approximation of structured signals. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(3):669–681, 2019.

[2] 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.

[3] Thomas Moreau, Mathurin Massias, Alexandre Gramfort, Pierre Ablin, Pierre-Antoine Bannier, Benjamin Charlier, Mathieu Dagréou, Tom Dupré la Tour, Ghislain Durif, Cassio F. Dantas, Quentin Klopfenstein, Johan Larsson, En Lai, Tanguy Lefort, Benoit Malézieux, Badr Moufad, Binh T. Nguyen, Alain Rakotomamonjy, Zaccharie Ramzi, Joseph Salmon, and Samuel Vaiter. Benchopt: Reproducible, efficient and collaborative optimization benchmarks. In Advances in Neural Information Processing Systems (NeurIPS), volume 36, New-Orleans, LA, USA, November 2022.

[4] 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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

Main activities

Main activities :

– Read papers and state of the art

- Benchmark existing algorithms

– Write problem formulation, proofs of convergence.

– Adapt the formulation to the target scenario.

– Propose a new dedicated algorithm.
– Program, run and analyse simulation results.

Complementary activities

– Participate to the teams activities : scientific meetings, seminars, scientific presentations.

Skills

- Strong mathematical background. Knowledge in numerical optimization is a plus.

- Good programming skills in Python. Knowledge of a distributed computation framework is a plus.

- The candidate should be proficient in English. Knowing French is not necessary, as daily communication in the team is mostly in English due to the strong international environment.

Benefits package

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking (90 days / year) and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Complementary health insurance under conditions

Remuneration

1st and 2nd year: 2 100 euros gross salary /month

3rd year: 2 190 euros gross salary / month

General Information

- **Theme/Domain :** Optimization, machine learning and statistical methods
  Statistics (Big data) (BAP E)
- **Town/city :** Montbonnot
- **Inria Center :** Centre Inria de l'Université Grenoble Alpes
- **Starting date :** 2024-09-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2024-06-30

Contacts

- **Inria Team :** THOTH
- **PhD Supervisor :**
  Hendrikx Hadrien / hadrien.hendrikx@inria.fr

About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

The keys to success

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.

> **Warning** : you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

Instruction to apply

Applications must be submitted online on the Inria website.

Processing of applications sent by other channels is not guaranteed.

Defence Security :

This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST).Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy :
As part of its diversity policy, all Inria positions are accessible to people with disabilities.