# Offer #2025-08930

# PhD Position F/M Theoretically justified neural network compression

**Contract type :** Fixed-term contract

**Level of qualifications required :** Graduate degree or equivalent

**Fonction :** PhD Position

## Context

The PhD position will be funded by ERC DYNASTY (Umut Simsekli).

The position might include traveling to conferences for paper presentation. Travel expenses will be covered within the limits of the scale in force.

## Assignment

With the increasing model sizes in deep learning and with its increasing use in low-resource environments, network compression techniques are becoming ever more important. Among many network compression techniques, network pruning has been arguably the most commonly used method [1], and it is rising in popularity and success [2].

A common conclusion in pruning research is that overparametrized networks can be greatly compressed by pruning with little to no cost at the overall performance of the network, including with simple schemes such as magnitude pruning [1,2]. For example, research on *iterative magnitude pruning* [3] demonstrated the possibility of compressing trained deep learning models by iteratively eliciting a much sparser substructure.

Recently, it has been illustrated that the choice of training hyperparameters such as learning rate affects the performance of such pruning strategies [3,4,5,6] took the first step towards providing a theoretical justification for these observations. Yet, it is still highly nontrivial to understand when a pruning method will or will not be useful [1]. Hence developing compression techniques with strong theoretical guarantees is crucial.

The goal of this thesis will be to address the aforementioned issues. More precisely, it will aim at addressing the following points:

- Make a selective bibliography in the topic and identify most relevant used methods in the existing literature
- Develop new compression techniques by noise injections to optimization algorithms.
- Perform theoretical analysis on the performance of the developed algorithm.
- Support the findings by experiments conducted on several benchmarks in modern neural networks.

References

[1] Neill, James O. "An overview of neural network compression." *arXiv preprint arXiv:2006.03669* (2020).
[2] Blalock, Davis, et al. "What is the state of neural network pruning?." *Proceedings of machine learning and systems* 2 (2020): 129-146.
[3] Frankle, Jonathan, and Michael Carbin. "The lottery ticket hypothesis: Finding sparse, trainable neural networks." *arXiv preprint arXiv:1803.03635* (2018).
[4] Zhou, Hattie, et al. "Deconstructing lottery tickets: Zeros, signs, and the supermask." *Advances in neural information processing systems* 32 (2019).
[5] Renda, Alex, Jonathan Frankle, and Michael Carbin. "Comparing rewinding and fine-tuning in neural network pruning." *arXiv preprint arXiv:2003.02389* (2020).
[6] Barsbey, Melih, et al. "Heavy tails in SGD and compressibility of overparametrized neural networks." *Advances in Neural Information Processing Systems* 34 (2021): 29364-29378.

## Main activities

Main activities :

- Conduct theoretical research
- Conduct experiments for empirical verification
- Write scientific articles
- Disseminate the scientific work in appropriate venues.

## Skills

Technical skills and level required :
Languages : High-level of professional/academic English
Coding skills : Good level of coding in Python and related deep learning libraries

## 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 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
- Social security coverage

## General Information

- **Theme/Domain :** Optimization, machine learning and statistical methods Statistics (Big data) (BAP E)
- **Town/city :** Paris
- **Inria Center :** Centre Inria de Paris
- **Starting date :** 2025-10-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2025-06-22

## Contacts

- **Inria Team :** SIERRA
- **PhD Supervisor :**
  Simsekli Umut / umut.simsekli@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.

**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

**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.