**AAP Contrats doctoraux en Intelligence artificielle**

**Cofinancé par l'ANR**

# Low-complexity models and algorithms for machine learning

## 1. DESCRIPTION OF THE PHD THESIS PROJECT

### 1.1 OBJECTIVES OF THE PROJECT BASED ON THE CURRENT STATE OF THE ART

The main motivation of this PhD project is to provide **fundamental research results in machine learning** in order to deploy sustainable learning algorithms. The goal is to **revisit and lower the complexity in most of training and inference algorithms**, including traditional learning that is appropriate for many industrial needs, and deep learning that achieves successes, one after the other, with an **unbearable ecological footprint**. Indeed, all families of machine learning approaches are suffering from limitations in terms of time and space complexity: efficient GPU implementations of backpropagation in deep neural networks dramatically demand an unsustainable amount of resources [Strubell2019], especially for linear-algebra computations and for the large volumes of data required to learn today's models; other families of approaches like kernel or spectral methods cannot scale to big data settings due to their computational limitations. **In both cases, a key bottleneck is the time and space complexity of linear algebra computations**, i.e., matrix or tensor products in high dimension, while non-linearities generally applies with a O(1) cost per input coefficient.

Accelerating ML algorithms is a vast ambition addressed in many complementary perspectives, as diverse as dedicated models, parallel and distributed optimization, hardware innovations or quantum machine learning. To contribute to this challenge, **our strategy is to leverage signal processing tools like fast transforms and random projections**, expecting a high impact within a short-term period.

Matrix factorization has been widely used to model large matrices. For instance, low-rank decompositions bring regularization effects and dimensionality reduction, within convex optimization frameworks and with efficient algorithms [Williams2001, Cai2010, Halko2011]. More recently, new factorization models have accelerated computations in dense layers [Yang2015, Moczulski2016] and kernel methods [Le2013, Si2016]. They are composed of many **elementary fixed factors** with structures allowing fast multiplications -- diagonal and permutation matrices, fast Fourier, Cosine or Hadamard transforms. **Our key research hypothesis** is that we can extend such models much further by **learning operators that admit a fast-transform structure**, in order to have models with even lower complexity. Technically, recent advances in non-convex optimization and dictionary learning [LeMagoarou2016] allow to decompose an arbitrary MxM matrix as the product of sparse matrices so that the cost of a matrix-vector product drops from $O(M^2)$ to O(M log M) in time and space. While this paves the way for learning fast operators, open questions remain: what is the ability of thoses models, with O(M log M) parameters, to approximate MxM matrices and to regularize optimization problems? Can we train a machine and learn such operators at the same time with an improved time and space complexity? To which extent can we not only improve neural networks and kernel methods but also revisit many other ML algorithms?

In our **preliminary works** [Giffon2019], we have demonstrated this strategy in the context of the K-means algorithm, widely used for clustering, vector quantization, indexing, nearest-neighbor search and many other applications. Our proposed QuicK-means approach is a variant of the celebrated Lloyd algorithm and learns a fast transform that acts similarly as the centroid matrix, in a computational efficient way. This preliminary works has

been both a proof of concept for our strategy and a first experience to identify key questions to be addressed, in terms of modeling, optimization, algorithms and computational issues.

*[Cai2010] J.-F. Cai, E. J. Cand`es, and Z. Shen. A Singular Value Thresholding Algorithm for Matrix Completion.SIAM Journal on Optimization, 20(4), jan 2010.*

*[Giffon2019] L. Giffon, V. Emiya, L. Ralaivola, and H. Kadri. QuicK-means: Acceleration of K-means by learning a fast transform. preprint, https://hal.archives-ouvertes.fr/hal-02174845, 2019.*

*[Halko2011] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. SIAM Review, 53(2), 2011.*

*[Le2013] Q. Le, T. Sarlo´s, and A. Smola. Fastfood: Approximating Kernel Expansions in Loglinear Time. In Int. Conf. on Machine Learning (ICML), 2013.*

*[LeMagoarou2016] L. Le Magoarou and R. Gribonval. Flexible Multilayer Sparse Approximations of Matrices and Applications. IEEE Journal of Selected Topics in Signal Processing, 10(4), June 2016.*

*[Moczulski2016] M. Moczulski, M. Denil, J. Appleyard, and N. de Freitas. ACDC: A Structured Efficient Linear Layer. In Int. Conf. on Learning Representations (ICLR), 2016.*

*[Rahimi2007] A. Rahimi and B. Recht. Random Features for Large-Scale Kernel Machines. In Adv. in Neural Information Processing Systems, 2007.*

*[Si2016] S. Si, C.-J. Hsieh, and I. S. Dhillon. Computationally Efficient Nystro¨M Approximation Using Fast Transforms. In Int. Conf. on Machine Learning (ICML), 2016.*

*[Strubell2019] E. Strubell, A. Ganesh, and A. McCallum. Energy and Policy Considerations for Deep Learning in NLP. Proc. of ACL, 2019.*

*[Williams2001] C. K. Williams and M. Seeger. Using the Nystro¨m method to speed up kernel machines. In Adv. in Neural Information Processing Systems, pages 682–688, 2001.*

*[Yang2015] Z. Yang, M. Moczulski, M. Denil, N. de Freitas, A. Smola, L. Song, and Z. Wang. Deep Fried Convnets. In Int. Conf. on Computer Vision (ICCV), Dec. 2015.*

## 1.2 METHODOLOGY

*Several challenges will be addressed through the following research directions:*

- *improving the generic fast-transform models that can be learned: we will address a number of identified theoretical questions about the expressivity of the sparse factorization models as well as the model architecture to guarantee fast computations*
- *improving learning algorithms for such models: leveraging recent advances in non-convex optimization, the optimization methods behind our learning algorithms needs to be specialized for our low-complexity models, by taking advantage of the fast-transform structure of our models and by ensuring the convergence properties of the optimization procedure;*
- *revisiting machine learning algorithms: we will integrate our low-complexity models and algorithms in widely-used machine learning algorithms such as kernel methods and deep neural networks, by replacing*

*the high-dimensional linear operations by learned fast-transforms and by adapting the learning algorithm consequently in order to guarantee their learning properties and computational efficiency.*

## 1.3 WORK PLAN

Ideally, the work could be organized as follows:

- first year: the PhD candidate will first make a review of the state-of-the-art, will study the limit of the proposed methods and develop new models and algorithms;
- second year: we target a publication of the first year's work; the existing or new developed models will be used to revisit one or several key machine learning approaches (e.g., kernel or deep learning methods) in order to lower their complexity;
- third year: the second year's work will be published; more research directions may be investigated; a particular effort will be dedicated to release efficient code and data for reproducibility and possible transfer.

In practice, the balance between theoretical and experimental works will be adjusted according to the PhD candidate's skills and motivations.

## 1.4 SUPERVISOR AND RESEARCH GROUP DESCRIPTION

*The PhD project will be located in Marseille. It will be mainly supervised by Valentin Emiya within QARMA team at LIS lab, and cosupervised by Caroline Chaux at I2M lab.*

*As a member of QARMA team at LIS lab, Aix-Marseille University, Valentin Emiya has developed research works and collaborations in which* ***machine learning and signal processing*** *closely inweave. Low-complexity models and low time/space complexity algorithms are underlying his research: he has had many contributions in* ***sparse and low-rank*** *models and related algorithms, as diverse as* ***sparse factorizations*** *for machine learning [Giffon2019],* ***dynamic screening tests*** *for the (group-)Lasso [Bonnefoy2015], low-rank models for complex-valued time-frequency data [Emiya2018] or branch-and-bound methods for sparse models [Emiya2014]. He has also pioneered research on the reconstruction of missing audio data, named audio inpainting in a seminal paper [Adler2012] and broadened within the ANR JCJC MAD project.*

*QARMA team conducts research in machine learning, including theory, algorithms and applications. It is composed of 11 permanent and 12 non-permanent members, with complementary skills covering all fields of machine learning. QARMA is part of the Data Science department of Laboratoire d'Informatique et Systèmes (LIS), which hosts about 375 members in 21 teams.*

*As a member of the signal and image processing team at I2M, Caroline Chaux will co-supervised the PhD project with her expertise in optimization and signal processing.*

*Beyond the supervision by V. Emiya and C. Chaux, the work may be the opportunity to develop collaborations currently under discussion with other research groups (e.g., Rémi Gribonval's team at ENS Lyon) and companies (e.g., LightOn).*

[Adler2012] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley. Audio Inpainting. IEEE Trans. Audio, Speech and Language Processing, 20(3), Mar. 2012.

[Bonnefoy2015] A. Bonnefoy, V. Emiya, L. Ralaivola, and R. Gribonval. Dynamic Screening: Accelerating First-Order Algorithms for the Lasso and Group-Lasso. IEEE Trans. Signal Process., 63(19), 2015.

[Emiya2014] V. Emiya, A. Bonnefoy, L. Daudet, and R. Gribonval. Compressed sensing with unknown sensor permutation. In Proc. Int. Conf. Audio Speech and Sig. Proces. (ICASSP), Florence, Italy, May 2014.

[Emiya2018] V. Emiya, R. Hamon, and C. Chaux. Being low-rank in the time-frequency plane. In Proc. Int. Conf. Audio Speech and Sig. Proces. (ICASSP), Calgary, Canada, Apr. 2018.

[Giffon2019] L. Giffon, V. Emiya, L. Ralaivola, and H. Kadri. QuicK-means: Acceleration of K-means by learning a fast transform. preprint, https://hal.archives-ouvertes.fr/hal-02174845, 2019.

## 2. RECENT PUBLICATIONS

The following publication is the starting point of the project:

[Giffon2019] L. Giffon, V. Emiya, L. Ralaivola, and H. Kadri. QuicK-means: Acceleration of K-means by learning a fast transform. preprint, https://hal.archives-ouvertes.fr/hal-02174845, 2019.

Other publications regarding the state of the art have been given above.

## 3. EXPECTED PROFILE OF THE CANDIDATE

*The candidate should have excellent general skills in maths and computer science, ideally with some expertise in machine learning, optimization and signal processing. He or she should be motivated and rapidly efficient for:*

- *modelling optimization problems related to the project, involving both an optimization background and general knowledge on machine learning and signal processing;*
- *handling a large panel of machine learning settings and approaches that can be targeted for new contributions with low-complexity models;*
- *designing related algorithms, including their rigorous formulation, their mathematical justification, their complexity analysis, their implementation and test;*
- *conducting experiments using baseline and new algorithms and large datasets; rigorous methodology and reliable code development will be required;*
- *writing research papers in English and presenting his or her works in seminars and international conferences.*

*In addition, the candidate will be integrated within QARMA team where interactions between members are encouraged.*

## 4. SUPERVISORS' PROFILE

Main supervisor: Valentin Emiya, LIS, Aix-Marseille University.

- Valentin Emiya has developed an expertise in **machine learning and signal processing**. He has been a member of the Équipe d'Apprentissage de Marseille (QARMA) at Laboratoire d'Informatique et Systèmes (LIS) since 2011. In the last five years, Valentin Emiya has been the supervisor of **3 postdocs, 3 PhD candidates and 16 interns**. He was the **principal investigator of 4 projects including ANR JCJC MAD** and had about 25 collaborators. For three years, he has been a member of the ANR scientific evaluation comittee on Artificial Intelligence. He is currently preparing his Habilitation à Diriger les Recherches. He is also dedicating a large effort on science popularization, especially as a leader of the Treize Minutes Marseille project since 2013.

- *Valentin Emiya is currently co-supervising two PhD projects: Marina Kreme's PhD on time-frequency inpainting (2017-2020) and Raphael Sturgis' PhD on the optimization of vessel trajectories by machine learning techniques (2019-2022, in collaboration with company Searoutes).*

Co-supervisor: Caroline Chaux, I2M, Aix-Marseille University.

- Caroline Chaux has developed an expertise in **convex optimization, sparse representations and signal processing**. She is a member of the Équipe Signal et Image (SI) at Institut de Mathématiques de Marseille (I2M) since 2012. In the last five years, Caroline Chaux has been the supervisor of **1 postdoc, 3 PhD candidates and 4 interns**. She defended her Habilitation à Diriger les Recherche (HdR) in January 2019. She is the **principal investigator of the Amidex project Bifrost**. For three years, she has been a member of the ANR scientific evaluation comittee on signal processing.
- *Caroline Chaux is currently co-supervising one PhD project: Marina Kreme's PhD on time-frequency inpainting (2017-2020).*

## VISA DU RESPONSABLE DE L'INSTITUT ET DU DIRECTEUR DE LABORATOIRE CONCERNÉS

| **Visa du responsable de l'institut, NOM Prénom** | **Visa du directeur du laboratoire, NOM Prénom** |
|---|---|
| GODARD Emmanuel | |
| **Fait à Marseille, le** 15/05/2020 | **Fait à Marseille, le** |
| **Signature**  | **Signature** |