RESEARCH CENTRE

Paris

IN PARTNERSHIP WITH:

Ecole normale supérieure de Paris, CNRS

2021 ACTIVITY REPORT

Project-Team SIERRA

Statistical Machine Learning and Parsimony

IN COLLABORATION WITH: Département d'Informatique de l'Ecole Normale Supérieure

DOMAIN

Applied Mathematics, Computation and Simulation

THEME

Optimization, machine learning and statistical methods

Contents

Project-Team SIERRA

Creation of the Project-Team: 2012 January 01

Keywords

Computer sciences and digital sciences

A3.4. – Machine learning and statistics

A5.4. – Computer vision

A6.2. – Scientific computing, Numerical Analysis & Optimization

A7.1. – Algorithms

A8.2. – Optimization

A9.2. – Machine learning

Other research topics and application domains

B9.5.6. – Data science

1 Team members, visitors, external collaborators

Research Scientists

- Francis Bach [Team leader, Inria, Senior Researcher, HDR]
- Laurent El Ghaoui [Inria, Advanced Research Position, from May 2021 until Jun 2021, HDR]
- Alessandro Rudi [Inria, Researcher]
- Umut Simsekli [Inria, Researcher]
- Adrien Taylor [Inria, Starting Research Position]
- Alexandre d'Aspremont [CNRS, Senior Researcher]

Post-Doctoral Fellows

- Martin Arjovsky [Inria]
- Pierre Cyril Aubin [Inria, from Sep 2021]
- Seyed Daneshmand [Inria]
- Hans Kersting [Inria]
- Ziad Kobeissi [Institut Louis Bachelier]
- Boris Muzellec [Inria]
- Anant Raj [Inria, from Nov 2021]
- Blake Woodworth [Inria, from Oct 2021]

PhD Students

- Mathieu Barre [École Normale Supérieure de Paris, until Oct 2021]
- Eloise Berthier [DGA]
- Raphael Berthier [Inria, until Sep 2021]
- Gaspard Beugnot [Inria, from Apr 2021]
- Vivien Cabannes [Inria]
- Radu Alexandru Dragomir [École polytechnique, until Oct 2021]
- Bertille Follain [École Normale Supérieure de Paris, from Sep 2021]
- Gautier Izacard [CNRS]
- Remi Jezequel [École Normale Supérieure de Paris]
- Marc Lambert [DGA]
- Clement Lezane [Université de Twente - Pays-Bas, from Sep 2021]
- Ulysse Marteau-Ferey [Inria]
- Gregoire Mialon [Inria]
- Celine Moucer [Ecole normale supérieure Paris-Saclay, from Sep 2021]
- Alex Nowak Vila [Inria, until Sep 2021]

- Benjamin Paul-Dubois-Taine [Université Paris-Saclay, from Sep 2021]
- Manon Romain [CNRS]
- Lawrence Stewart [Inria, from Oct 2021]

Technical Staff

- Loïc Estève [Inria, Engineer]
- Anant Raj [Inria, Engineer, from Sep 2021 until Oct 2021]

Interns and Apprentices

- Theophile Cantelobre [Inria, from Apr 2021 until Sep 2021]
- Megi Dervishi [Inria, from Mar 2021 until May 2021]
- Clement Lezane [Inria, from May 2021 until Sep 2021]
- Tomas Rigaux [Inria, until Feb 2021]
- Louis Romain Roux [Inria, from Apr 2021 until Oct 2021]
- Lawrence Stewart [Inria, from Apr 2021 until Sep 2021]
- Badr Youbi Idrissi [Inria, from Sep 2021 until Oct 2021]

Administrative Assistants

- Helene Bessin Rousseau [Inria, from Feb 2021]
- Helene Milome [Inria]
- Scheherazade Rouag [Inria, until Apr 2021]

Visiting Scientist

- Benjamin Paul-Dubois-Taine [Université Paris-Saclay, from Apr 2021 until Aug 2021]

External Collaborator

- Laurent El Ghaoui [University of California-Berkeley, from Jul 2021, HDR]

2 Overall objectives

2.1 Statement

Machine learning is a recent scientific domain, positioned between applied mathematics, statistics and computer science. Its goals are the optimization, control, and modelisation of complex systems from examples. It applies to data from numerous engineering and scientific fields (e.g., vision, bioinformatics, neuroscience, audio processing, text processing, economy, finance, etc.), the ultimate goal being to derive general theories and algorithms allowing advances in each of these domains. Machine learning is characterized by the high quality and quantity of the exchanges between theory, algorithms and applications: interesting theoretical problems almost always emerge from applications, while theoretical analysis allows the understanding of why and when popular or successful algorithms do or do not work, and leads to proposing significant improvements.

Our academic positioning is exactly at the intersection between these three aspects—algorithms, theory and applications—and our main research goal is to make the link between theory and algorithms,

and between algorithms and high-impact applications in various engineering and scientific fields, in particular computer vision, bioinformatics, audio processing, text processing and neuro-imaging.

Machine learning is now a vast field of research and the team focuses on the following aspects: supervised learning (kernel methods, calibration), unsupervised learning (matrix factorization, statistical tests), parsimony (structured sparsity, theory and algorithms), and optimization (convex optimization, bandit learning). These four research axes are strongly interdependent, and the interplay between them is key to successful practical applications.

3 Research program

3.1 Supervised Learning

This part of our research focuses on methods where, given a set of examples of input/output pairs, the goal is to predict the output for a new input, with research on kernel methods, calibration methods, and multi-task learning.

3.2 Unsupervised Learning

We focus here on methods where no output is given and the goal is to find structure of certain known types (e.g., discrete or low-dimensional) in the data, with a focus on matrix factorization, statistical tests, dimension reduction, and semi-supervised learning.

3.3 Parsimony

The concept of parsimony is central to many areas of science. In the context of statistical machine learning, this takes the form of variable or feature selection. The team focuses primarily on structured sparsity, with theoretical and algorithmic contributions.

3.4 Optimization

Optimization in all its forms is central to machine learning, as many of its theoretical frameworks are based at least in part on empirical risk minimization. The team focuses primarily on convex and bandit optimization, with a particular focus on large-scale optimization.

4 Application domains

4.1 Applications for Machine Learning

Machine learning research can be conducted from two main perspectives: the first one, which has been dominant in the last 30 years, is to design learning algorithms and theories which are as generic as possible, the goal being to make as few assumptions as possible regarding the problems to be solved and to let data speak for themselves. This has led to many interesting methodological developments and successful applications. However, we believe that this strategy has reached its limit for many application domains, such as computer vision, bioinformatics, neuro-imaging, text and audio processing, which leads to the second perspective our team is built on: Research in machine learning theory and algorithms should be driven by interdisciplinary collaborations, so that specific prior knowledge may be properly introduced into the learning process, in particular with the following fields:

- Computer vision: object recognition, object detection, image segmentation, image/video processing, computational photography. In collaboration with the Willow project-team.
- Bioinformatics: cancer diagnosis, protein function prediction, virtual screening. In collaboration with Institut Curie.
- Text processing: document collection modeling, language models.

- Audio processing: source separation, speech/music processing.
- Neuro-imaging: brain-computer interface (fMRI, EEG, MEG).

5 Highlights of the year

5.1 Awards

- Alessandro Rudi is recipient of the ERC Starting Grant. The ERC project is named REAL (947908) and corresponds to a grant of 1.5 millions € for the period 2021-2026.
- Outstanding paper award at NeurIPS 2021 for the paper *Continuized Accelerations of Deterministic and Stochastic Gradient Descents, and of Gossip Algorithms* (Mathieu Even, Raphaël Berthier, Francis Bach, Nicolas Flammarion, Hadrien Hendrikx, Pierre Gaillard, Laurent Massoulié, Adrien Taylor).
- Test-of-time award at NeurIPS 2021 for the paper *Online Learning for Latent Dirichlet Allocation* by Matthew Hoffman, David Blei, and Francis Bach.

5.2 Books

- Alexandre d'Aspremont, Damien Scieur and Adrien Taylor (2021), "Acceleration Methods", Foundations and Trends® in Optimization: Vol. 5: No. 1-2, pp 1-245.

6 New software and platforms

6.1 New software

6.1.1 ProxASAGA

Keyword: Optimization

Functional Description: A C++/Python code implementing the methods in the paper "Breaking the Nonsmooth Barrier: A Scalable Parallel Method for Composite Optimization", F. Pedregosa, R. Leblond and S. Lacoste-Julien, Advances in Neural Information Processing Systems (NIPS) 2017. Due to their simplicity and excellent performance, parallel asynchronous variants of stochastic gradient descent have become popular methods to solve a wide range of large-scale optimization problems on multi-core architectures. Yet, despite their practical success, support for nonsmooth objectives is still lacking, making them unsuitable for many problems of interest in machine learning, such as the Lasso, group Lasso or empirical risk minimization with convex constraints. In this work, we propose and analyze ProxASAGA, a fully asynchronous sparse method inspired by SAGA, a variance reduced incremental gradient algorithm. The proposed method is easy to implement and significantly outperforms the state of the art on several nonsmooth, large-scale problems. We prove that our method achieves a theoretical linear speedup with respect to the sequential version under assumptions on the sparsity of gradients and block-separability of the proximal term. Empirical benchmarks on a multi-core architecture illustrate practical speedups of up to 12x on a 20-core machine.

URL: https://github.com/fabianp/ProxASAGA

Contact: Fabian Pedregosa

6.1.2 object-states-action

Keyword: Computer vision

Functional Description: Code for the paper Joint Discovery of Object States and Manipulation Actions, ICCV 2017: Many human activities involve object manipulations aiming to modify the object state. Examples of common state changes include full/empty bottle, open/closed door, and attached/detached car wheel. In this work, we seek to automatically discover the states of objects and the associated manipulation actions. Given a set of videos for a particular task, we propose a joint model that learns to identify object states and to localize state-modifying actions. Our model is formulated as a discriminative clustering cost with constraints. We assume a consistent temporal order for the changes in object states and manipulation actions, and introduce new optimization techniques to learn model parameters without additional supervision. We demonstrate successful discovery of seven manipulation actions and corresponding object states on a new dataset of videos depicting real-life object manipulations. We show that our joint formulation results in an improvement of object state discovery by action recognition and vice versa.

URL: https://github.com/jalayrac/object-states-action

Publication: hal-01676084

Contact: Jean-Baptiste Alayrac

Participants: Jean-Baptiste Alayrac, Josef Sivic, Ivan Laptev, Simon Lacoste-Julien

7 New results

7.1 On the Effectiveness of Richardson Extrapolation in Data Science

Richardson extrapolation is a classical technique from numerical analysis that can improve the approximation error of an estimation method by combining linearly several estimates obtained from different values of one of its hyperparameters without the need to know in details the inner structure of the original estimation method. The main goal of this paper is to study when Richardson extrapolation can be used within data science beyond the existing applications to step-size adaptations in stochastic gradient descent. We identify two situations where Richardson interpolation can be useful: (1) when the hyperparameter is the number of iterations of an existing iterative optimization algorithm with applications to averaged gradient descent and Frank–Wolfe algorithms (where we obtain asymptotically rates of $O(1/k^2)$ on polytopes, where k is the number of iterations) and (2) when it is a regularization parameter with applications to Nesterov smoothing techniques for minimizing nonsmooth functions (where we obtain asymptotically rates close to $O(1/k^2)$ for nonsmooth functions) and kernel ridge regression. In all these cases, we show that extrapolation techniques come with no significant loss in performance but with sometimes strong gains, and we provide theoretical justifications based on asymptotic developments for such gains, as well as empirical illustrations on classical problems from machine learning.

7.2 Batch Normalization Orthogonalizes Representations in Deep Random Networks

This paper underlines a subtle property of batch-normalization (BN): Successive batch normalizations with random linear transformations make hidden representations increasingly orthogonal across layers of a deep neural network. We establish a non-asymptotic characterization of the interplay between depth, width, and the orthogonality of deep representations. More precisely, under a mild assumption, we prove that the deviation of the representations from orthogonality rapidly decays with depth up to a term inversely proportional to the network width. This result has two main implications: 1) Theoretically, as the depth grows, the distribution of the representation –after the linear layers– contracts to a Wasserstein-2 ball around an isotropic Gaussian distribution. Furthermore, the radius of this Wasserstein ball shrinks with the width of the network. 2) In practice, the orthogonality of the representations directly influences the performance of stochastic gradient descent (SGD). When representations are initially aligned, we observe SGD wastes many iterations to orthogonalize representations before the classification. Nevertheless, we experimentally show that starting optimization from orthogonal representations is sufficient to accelerate SGD, with no need for BN.

7.3 A Continuized View on Nesterov Acceleration for Stochastic Gradient Descent and Randomized Gossip

We introduce the "continuized" Nesterov acceleration, a close variant of Nesterov acceleration whose variables are indexed by a continuous time parameter. The two variables continuously mix following a linear ordinary differential equation and take gradient steps at random times. This continuized variant benefits from the best of the continuous and the discrete frameworks: as a continuous process, one can use differential calculus to analyze convergence and obtain analytical expressions for the parameters; and a discretization of the continuized process can be computed exactly with convergence rates similar to those of Nesterov original acceleration. We show that the discretization has the same structure as Nesterov acceleration, but with random parameters. We provide continuized Nesterov acceleration under deterministic as well as stochastic gradients, with either additive or multiplicative noise. Finally, using our continuized framework and expressing the gossip averaging problem as the stochastic minimization of a certain energy function, we provide the first rigorous acceleration of asynchronous gossip algorithms

7.4 A Dimension-free Computational Upper-bound for Smooth Optimal Transport Estimation

It is well-known that plug-in statistical estimation of optimal transport suffers from the curse of dimensionality. Despite recent efforts to improve the rate of estimation with the smoothness of the problem, the computational complexity of these recently proposed methods still degrade exponentially with the dimension. In this paper, thanks to an infinite-dimensional sum-of-squares representation, we derive a statistical estimator of smooth optimal transport which achieves a precision ε from $\tilde{O}(\varepsilon^{-2})$, independent and identically distributed samples from the distributions, for a computational cost of $\tilde{O}(\varepsilon^{-4})$ when the smoothness increases, hence yielding dimension-free statistical and computational rates, with potentially exponentially dimension-dependent constants.

7.5 Fast rates in structured prediction

Discrete supervised learning problems such as classification are often tackled by introducing a continuous surrogate problem akin to regression. Bounding the original error, between estimate and solution, by the surrogate error endows discrete problems with convergence rates already shown for continuous instances. Yet, current approaches do not leverage the fact that discrete problems are essentially predicting a discrete output when continuous problems are predicting a continuous value. In this paper, we tackle this issue for general structured prediction problems, opening the way to "super fast" rates, that is, convergence rates for the excess risk faster than n^{-1}, where n is the number of observations, with even exponential rates with the strongest assumptions. We first illustrate it for predictors based on nearest neighbors, generalizing rates known for binary classification to any discrete problem within the framework of structured prediction. We then consider kernel ridge regression where we improve known rates in $n^{-1/4}$ to arbitrarily fast rates, depending on a parameter characterizing the hardness of the problem, thus allowing, under smoothness assumptions, to bypass the curse of dimensionality.

7.6 Disambiguation of weak supervision with exponential convergence rates

Machine learning approached through supervised learning requires expensive annotation of data. This motivates weakly supervised learning, where data are annotated with incomplete yet discriminative information. In this project, we focus on partial labelling, an instance of weak supervision where, from a given input, we are given a set of potential targets. We review a disambiguation principle to recover full supervision from weak supervision, and propose an empirical disambiguation algorithm. We prove exponential convergence rates of our algorithm under classical learnability assumptions, and we illustrate the usefulness of our method on practical examples

7.7 Deep Equals Shallow for ReLU Networks in Kernel Regimes

Deep networks are often considered to be more expressive than shallow ones in terms of approximation. Indeed, certain functions can be approximated by deep networks provably more efficiently than by shallow ones, however, no tractable algorithms are known for learning such deep models. Separately, a recent line of work has shown that deep networks trained with gradient descent may behave like (tractable) kernel methods in a certain over-parameterized regime, where the kernel is determined by the architecture and initialization, and this paper focuses on approximation for such kernels. We show that for ReLU activations, the kernels derived from deep fully-connected networks have essentially the same approximation properties as their "shallow" two-layer counterpart, namely the same eigenvalue decay for the corresponding integral operator. This highlights the limitations of the kernel framework for understanding the benefits of such deep architectures. Our main theoretical result relies on characterizing such eigenvalue decays through differentiability properties of the kernel function, which also easily applies to the study of other kernels defined on the sphere.

7.8 Explicit Regularization of Stochastic Gradient Methods through Duality

We consider stochastic gradient methods under the interpolation regime where a perfect fit can be obtained (minimum loss at each observation). While previous work highlighted the implicit regularization of such algorithms, we consider an explicit regularization framework as a minimum Bregman divergence convex feasibility problem. Using convex duality, we propose randomized Dykstra-style algorithms based on randomized dual coordinate ascent. For non-accelerated coordinate descent, we obtain an algorithm which bears strong similarities with (non-averaged) stochastic mirror descent on specific functions, as it is equivalent for quadratic objectives, and equivalent in the early iterations for more general objectives. It comes with the benefit of an explicit convergence theorem to a minimum norm solution. For accelerated coordinate descent, we obtain a new algorithm that has better convergence properties than existing stochastic gradient methods in the interpolating regime. This leads to accelerated versions of the perceptron for generic ℓ_p-norm regularizers, which we illustrate in experiments.

7.9 The recursive variational Gaussian approximation (R-VGA)

We consider the problem of computing a Gaussian approximation to the posterior distribution of a parameter given N observations and a Gaussian prior. Owing to the need of processing large sample sizes N, a variety of approximate tractable methods revolving around online learning have flourished over the past decades. In the present work, we propose to use variational inference (VI) to compute a Gaussian approximation to the posterior through a single pass over the data. Our algorithm is a recursive version of the variational Gaussian approximation we have called recursive variational Gaussian approximation (RVGA). We start from the prior, and for each observation we compute the nearest Gaussian approximation in the sense of Kullback-Leibler divergence to the posterior given this observation. In turn, this approximation is considered as the new prior when incorporating the next observation. This recursive version based on a sequence of optimal Gaussian approximations leads to a novel implicit update scheme which resembles the online Newton algorithm, and which is shown to boil down to the Kalman filter for Bayesian linear regression. In the context of Bayesian logistic regression the implicit scheme may be solved, and the algorithm is shown to perform better than the extended Kalman filter, while being far less computationally demanding than its sampling counterparts.

7.10 Restarting Frank-Wolfe

Conditional Gradients (aka Frank-Wolfe algorithms) form a classical set of methods for constrained smooth convex minimization due to their simplicity, the absence of projection step, and competitive numerical performance. While the vanilla Frank-Wolfe algorithm only ensures a worst-case rate of O(1/epsilon), various recent results have shown that for strongly convex functions, the method can be slightly modified to achieve linear convergence. However, this still leaves a huge gap between sublinear O(1/epsilon) convergence and linear O(log 1/epsilon) convergence to reach an eps-approximate solution. Here, we present a new variant of Conditional Gradients, that can dynamically adapt to the function's

geometric properties using restarts and thus smoothly interpolates between the sublinear and linear regimes.

7.11 Approximation Bounds for Sparse Programs

We show that sparsity-constrained optimization problems over low dimensional spaces tend to have a small duality gap. We use the Shapley-Folkman theorem to derive both data-driven bounds on the duality gap, and an efficient primalization procedure to recover feasible points satisfying these bounds. These error bounds are proportional to the rate of growth of the objective with the target cardinality k, which means in particular that the relaxation is nearly tight as soon as k is large enough so that only uninformative features are added.

7.12 Linear Bandits on Uniformly Convex Sets

Linear bandit algorithms yield two types of structural assumptions lead to better pseudo-regret bounds. When K is the simplex or an ℓ_p ball with p in $]1,2]$, there exist bandits algorithms with $O(\sqrt{nT})$ pseudo-regret bounds. Here, we derive bandit algorithms for some strongly convex sets beyond l_p balls that enjoy pseudo-regret bounds of $O(\sqrt{nT})$, which answers an open question from (BCB12, S 5.5). Interestingly, when the action set is uniformly convex but not necessarily strongly convex, we obtain pseudo-regret bounds with a dimension dependency smaller than $O(\sqrt{n})$. However, this comes at the expense of asymptotic rates in T varying between $O(\sqrt{T})$ and $O(T)$.

7.13 Local and Global Uniform Convexity Conditions

We review various characterizations of uniform convexity and smoothness on norm balls in finite-dimensional spaces and connect results stemming from the geometry of Banach spaces with scaling inequalities used in analyzing the convergence of optimization methods. In particular, we establish local versions of these conditions to provide sharper insights on a recent body of complexity results in learning theory, online learning, or offline optimization, which rely on the strong convexity of the feasible set. While they have a significant impact on complexity, these strong convexity or uniform convexity properties of feasible sets are not exploited as thoroughly as their functional counterparts, and this work is an effort to correct this imbalance. We conclude with some practical examples in optimization and machine learning where leveraging these conditions and localized assumptions lead to new complexity results.

7.14 Fractal Structure and Generalization Properties of Stochastic Optimization Algorithms

Understanding generalization in deep learning has been one of the major challenges in statistical learning theory over the last decade. While recent work has illustrated that the dataset and the training algorithm must be taken into account in order to obtain meaningful generalization bounds, it is still theoretically not clear which properties of the data and the algorithm determine the generalization performance. In this study, we approach this problem from a dynamical systems theory perspective and represent stochastic optimization algorithms as *random iterated function systems* (IFS). Well studied in the dynamical systems literature, under mild assumptions, such IFSs can be shown to be ergodic with an invariant measure that is often supported on sets with a *fractal structure*. As our main contribution, we prove that the generalization error of a stochastic optimization algorithm can be bounded based on the 'complexity' of the fractal structure that underlies its invariant measure. Then, by leveraging results from dynamical systems theory, we show that the generalization error can be explicitly linked to the choice of the algorithm (e.g., stochastic gradient descent – SGD), algorithm hyperparameters (e.g., step-size, batch-size), and the geometry of the problem (e.g., Hessian of the loss). We further specialize our results to specific problems (e.g., linear/logistic regression, one hidden-layered neural networks) and algorithms (e.g., SGD and preconditioned variants), and obtain analytical estimates for our bound. For modern neural networks, we develop an efficient algorithm to compute the developed bound and support our theory with various experiments on neural networks.

7.15 PSD Representations for Effective Probability Models

Finding a good way to model probability densities is key to probabilistic inference. An ideal model should be able to concisely approximate any probability while being also compatible with two main operations: multiplications of two models (product rule) and marginalization with respect to a subset of the random variables (sum rule). In this work, we show that a recently proposed class of positive semi-definite (PSD) models for non-negative functions is particularly suited to this end. In particular, we characterize both approximation and generalization capabilities of PSD models, showing that they enjoy strong theoretical guarantees. Moreover, we show that we can perform efficiently both sum and product rule in closed form via matrix operations, enjoying the same versatility of mixture models. Our results open the way to applications of PSD models to density estimation, decision theory and inference.

7.16 Sampling from Arbitrary Functions via PSD Models

In many areas of applied statistics and machine learning, generating an arbitrary number of independent and identically distributed (i.i.d.) samples from a given distribution is a key task. When the distribution is known only through evaluations of the density, current methods either scale badly with the dimension or require very involved implementations. Instead, we take a two-step approach by first modeling the probability distribution and then sampling from that model. We use the recently introduced class of positive semi-definite (PSD) models, which have been shown to be efficient for approximating probability densities. We show that these models can approximate a large class of densities concisely using few evaluations, and present a simple algorithm to effectively sample from these models. We also present preliminary empirical results to illustrate our assertions.

7.17 Mixability made efficient: Fast online multiclass logistic regression

Mixability has been shown to be a powerful tool to obtain algorithms with optimal regret. However, the resulting methods often suffer from high computational complexity which has reduced their practical applicability. For example, in the case of multiclass logistic regression, the aggregating forecaster (Foster et al. (2018)) achieves a regret of $O(\log(Bn))$ whereas Online Newton Step achieves $O(e^B \log(n))$ obtaining a double exponential gain in B (a bound on the norm of comparative functions). However, this high statistical performance is at the price of a prohibitive computational complexity $O(n^{37})$.

7.18 Beyond Tikhonov: Faster Learning with Self-Concordant Losses via Iterative Regularization

The theory of spectral filtering is a remarkable tool to understand the statistical properties of learning with kernels. For least squares, it allows to derive various regularization schemes that yield faster convergence rates of the excess risk than with Tikhonov regularization. This is typically achieved by leveraging classical assumptions called source and capacity conditions, which characterize the difficulty of the learning task. In order to understand estimators derived from other loss functions, Marteau-Ferey et al. have extended the theory of Tikhonov regularization to generalized self concordant loss functions (GSC), which contain, e.g., the logistic loss. In this paper, we go a step further and show that fast and optimal rates can be achieved for GSC by using the iterated Tikhonov regularization scheme, which is intrinsically related to the proximal point method in optimization, and overcomes the limitation of the classical Tikhonov regularization.

8 Bilateral contracts and grants with industry

8.1 Bilateral grants with industry

- Alexandre d'Aspremont, Francis Bach, Martin Jaggi (EPFL): Google Focused award.
- Francis Bach: Gift from Facebook AI Research.
- Alexandre d'Aspremont: fondation AXA, "Mécénat scientifique", optimisation & machine learning.

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 Associate Teams in the framework of an Inria International Lab or in the framework of an Inria International Program

4TUNE

Title: Adaptive, Efficient, Provable and Flexible Tuning for Machine Learning

Duration: 2020 ->

Coordinator: Peter Grünwald (pdg@cwi.nl)

Partners:

- CWI

Inria contact: Francis Bach

Summary:

FOAM

Title: First-Order Accelerated Methods for Machine Learning.

Duration: 2020 ->

Coordinator: Cristobal Guzman (crguzmanp@mat.uc.cl)

Partners:

- Pontificia Universidad Católica de Chile

Inria contact: Alexandre D'Aspremont

Summary:

9.2 European initiatives

9.2.1 Horizon Europe

- Alessandro Rudi is recipient of the ERC Starting Grant. The ERC project is named REAL (947908) and corresponds to a grant of 1.5 millions € for the period 2021-2026.

9.3 National initiatives

- Alexandre d'Aspremont: IRIS, PSL "Science des données, données de la science".

10 Dissemination

10.0.1 Journal

Member of the editorial boards

- Alexandre d'Aspremont, associate editor for SIAM Journal on the Mathematics of Data Science.
- Alexandre d'Aspremont, associate editor for SIAM Journal on Optimization.
- Alexandre d'Aspremont, associate editor for Mathematics of Operations Research.
- Francis Bach, co-editor-in-chief, Journal of Machine Learning Research
- Francis Bach, associate Editor, Mathematical Programming
- Francis Bach, associate editor, Foundations of Computational Mathematics (FoCM)

Reviewer - reviewing activities

- Alessandro Rudi: area chair for “International Conference on Machine Learning 2021”
- Alessandro Rudi: area chair for “Advances in Neural Information Processing Systems 2021”
- Adrien Taylor: reviewer for “Conference on Learning Theory 2021”.
- Adrien Taylor: reviewer for “Computational Optimization And Applications”.
- Adrien Taylor: reviewer for “IMA Journal on Numerical Analysis”.
- Adrien Taylor: reviewer for “Journal of Optimization Theory and Applications”.
- Adrien Taylor: reviewer for “Mathematical Programming”.
- Adrien Taylor: reviewer for “SIAM Journal on Optimization”.
- Umut Simsekli: area chair for “International Conference on Machine Learning 2021”
- Umut Simsekli: area chair for “Advances in Neural Information Processing Systems 2021”

10.0.2 Invited talks

- Adrien Taylor: invited talk at Europt (July 2021, online).
- Adrien Taylor: invited talk at the “All-Russian optimization seminar” (May 2021, online).
- Adrien Taylor: invited talk at the “Optimization without Borders” workshop (July 2021, Sochi/online).
- Adrien Taylor: invited talk at EPFL (October 2021, Lausanne).
- Adrien Taylor: invited talk at SUTD (December 2021, Singapore/Online).
- Alexandre d’Aspremont: invited talk at Cornell ORIE.
- Alexandre d’Aspremont: invited talk at MIT OR.
- Alexandre d’Aspremont: invited talk at "Optimization without Borders" workshop, Sochi.
- Francis Bach: invited talk at Caltech University (January 2021, online)
- Francis Bach: invited talk at Journées Math / IA (March 2021, online)
- Francis Bach: invited talk at Stanford University (April 2021, online)
- Francis Bach: invited talk at Georgia Tech University (September 2021, online)
- Francis Bach: invited talk at IMSI, Chicago (October 2021, online)
- Francis Bach: invited talk at MIT (November 2021, online)
- Francis Bach: invited talk at University of Michigan (November 2021, online)
- Umut Simsekli: invited talk at University of California, Los Angeles - Max Planck Institute (February 2021, online)
- Umut Simsekli: invited talk at the Mathematical Statistics and Learning Workshop (June 2021, Barcelona)
- Umut Simsekli: invited talk at ESSEC Business School (November 2021, Cergy)
- Umut Simsekli: invited talk at Current Developments in MCMC Methods Workshop (December 2021, Warsaw)

- Umut Simsekli: invited talk at University of Bristol (November 2021, online)
- Alessandro Rudi: invited talk "Finding global minima via kernel approximations", RWTH Chair for Mathematics of Information Processing, RWTH Aachen University, 14 June 2021.
- Alessandro Rudi: invited talk "Towards energy-aware ML From first principles", ECO-INFO CNRS meeting, 21 June 2021.
- Alessandro Rudi: invited talk "Effective models for non-negative functions", Mathematical Statistics and Learning 2021, Barcelona, 29 June 2021.
- Alessandro Rudi: invited talk "PSD models for Non-convex optimization and beyond", Statistics Seminars, Sorbonne Université, 9 Nov 2021.

10.0.3 Leadership within the scientific community

- Francis Bach: president of the board of ICML

10.0.4 Research administration

- Francis Bach: Deputy Scientific director, Inria Paris

10.1 Teaching - Supervision - Juries

10.1.1 Teaching

- Master: Alexandre d'Aspremont, Optimisation Combinatoire et Convexe, avec Zhentao Li, (2015-Present) cours magistraux 30h, Master M1, ENS Paris.
- Master: Alexandre d'Aspremont, Optimisation convexe: modélisation, algorithmes et applications cours magistraux 21h (2011-Present), Master M2 MVA, ENS PS.
- Master : Francis Bach, Optimisation et apprentissage statistique, 20h, Master M2 (Mathématiques de l'aléatoire), Université Paris-Sud, France.
- Master : Francis Bach, Learning theory from first principles, 27h, Master M2 MASH, Université Paris Dauphine PSL, France.
- Master : Francis Bach, Machine Learning, 20h, Master ICFP (Physique), Université PSL.
- Master: Alessandro Rudi, Umut Simsekli. Introduction to Machine Learning, 52h, L3, ENS, Paris.

10.1.2 Supervision

- PhD in progress: Grégoire Mialon, Sample Selection Methods, 2018, Alexandre d'Aspremont (joint with Julien Mairal)
- PhD in progress: Manon Romain, Causal Inference Algorithms, 2020, Alexandre d'Aspremont
- PhD in progress: Theophile Cantelobre, supervised by Alessandro Rudi, Benjamin Guedj, Carlo Ciliberto (UCL).
- PhD in progress: Gaspard Beugnot, supervised by Alessandro Rudi, Julien Mairal.
- PhD in progress: Ulysse Marteau Ferey, supervised by Alessandro Rudi and Francis Bach.
- PhD in progress: Vivien Cabannes, supervised by Francis Bach and Alessandro Rudi.
- PhD in progress: Eloise Berthier, supervised by Francis Bach.
- PhD in progress: Theo Ryffel, supervised by Francis Bach and David Pointcheval.

- PhD in progress: Rémi Jezequel, supervised by Pierre Gaillard and Alessandro Rudi.
- PhD in progress: Antoine Bambade, supervised by Jean-Ponce (Willow), Justin Carpentier (Willow), and Adrien Taylor.
- PhD in progress: Marc Lambert, supervised by Francis Bach and Silvère Bonnabel.
- PhD in progress: Ivan Lerner, co-advised with Anita Burgun et Antoine Neuraz.
- PhD in progress: Lawrence Stewart, co-advised by Francis Bach and Jean-Philippe Vert.
- PhD in progress: Céline Moucer, supervised by Adrien Taylor and Francis Bach
- PhD in progress: Bertille Follain, supervised by Umut Simsekli and Francis Bach
- PhD defended: Raphaël Berthier, supervised by Francis Bach and Pierre Gaillard.
- PhD defended: Radu - Dragomir Alexandru, Bregman Gradient Methods, Alexandre d'Aspremont (joint with Jérôme Bolte)
- PhD defended: Mathieu Barré, Accelerated Polyak Methods, Alexandre d'Aspremont
- PhD defended: Alex Nowak-Vila, supervised by Francis Bach and Alessandro Rudi.
- PhD defended: Hadrien Hendrikx, supervised by Francis Bach and Laurent Massoulié.

10.1.3 Juries

- Francis Bach: HDR committee of Emilie Kaufmann
- Francis Bach: HDR committee of Aurélien Bellet
- Francis Bach: HDR committee of Samuel Vaiter
- Umut Simsekli: PhD committee of François-Pierre Paty

10.2 Popularization

10.2.1 Interventions

- Francis Bach: Keynote talk at GPAI summit (November 2021)
- Francis Bach: Presentation on scientific challenges of AI, Ecole de Guerre (September 2021)

11 Scientific production

11.1 Major publications

[1] U. Marteau-Ferey, F. Bach and A. Rudi. 'Non-parametric Models for Non-negative Functions'. working paper or preprint. July 2020. URL: https://hal.inria.fr/hal-02891640.

[2] V. Roulet and A. D'Aspremont. 'Sharpness, Restart and Acceleration'. In: *SIAM Journal on Optimization* 30.1 (Oct. 2020), pp. 262–289. DOI: 10.1137/18M1224568. URL: https://hal.archives-ouvertes.fr/hal-02983236.

11.2 Publications of the year

International journals

[3] F. Bach. 'On the Effectiveness of Richardson Extrapolation in Data Science'. In: *SIAM Journal on Mathematics of Data Science* 3.4 (2021), pp. 1251–1277. URL: https://hal.archives-ouvertes.fr/hal-02470950.

[4] A. D'Aspremont, M. Cucuringu and H. Tyagi. 'Ranking and synchronization from pairwise measurements via SVD'. In: *Journal of Machine Learning Research* 22.19 (11th Feb. 2021), pp. 1–63. URL: https://hal.archives-ouvertes.fr/hal-02340372.

[5] R.-A. Dragomir, A. Taylor, A. D'Aspremont, J. Bolte and A. d'Aspremont. 'Optimal Complexity and Certification of Bregman First-Order Methods'. In: *Mathematical Programming* (21st Apr. 2021). DOI: 10.1007/s10107-021-01618-1. URL: https://hal.inria.fr/hal-02384167.

[6] M. Lambert, S. Bonnabel and F. Bach. 'The recursive variational Gaussian approximation (R-VGA)'. In: *Statistics and Computing* (2021). URL: https://hal.inria.fr/hal-03086627.

International peer-reviewed conferences

[7] G. Beugnot, J. Mairal and A. Rudi. 'Beyond Tikhonov: Faster Learning with Self-Concordant Losses via Iterative Regularization'. In: NeurIPS 2021 – 35th Annual Conference on Neural Information Processing Systems. Advances in Neural Information Processing Systems 34. Virtual, France, 6th Dec. 2021, pp. 1–37. URL: https://hal.inria.fr/hal-03406072.

[8] A. Bietti and F. Bach. 'Deep Equals Shallow for ReLU Networks in Kernel Regimes'. In: ICLR 2021 - International Conference on Learning Representations. Virtual, Austria, 3rd May 2021, pp. 1–22. URL: https://hal.inria.fr/hal-02963250.

[9] T. Birdal, A. Lou, L. Guibas and U. Şimşekli. 'Intrinsic Dimension, Persistent Homology and Generalization in Neural Networks'. In: NeurIPS 2021 - Thirty-fifth Conference on Neural Information Processing Systems. Virtual, France, 3rd Dec. 2021. URL: https://hal.inria.fr/hal-03530322.

[10] V. Cabannes, F. Bach and A. Rudi. 'Fast Rates for Structured Prediction'. In: COLT 2021 - 34th Annual Conference on Learning Theory. Vol. 134. Proceedings of Machine Learning Research. Boulder, Colorado, United States, 15th July 2021. URL: https://hal.archives-ouvertes.fr/hal-03384304.

[11] V. Cabannes, L. Pillaud-Vivien, F. Bach and A. Rudi. 'Overcoming the curse of dimensionality with Laplacian regularization in semi-supervised learning'. In: NeurIPS 2021 - Thirty-fifth conference on Neural Information Processing Systems (NeurIPS). Online, Unknown Region, 6th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03454809.

[12] V. Cabannes, A. Rudi and F. Bach. 'Disambiguation of Weak Supervision leading to Exponential Convergence rates'. In: *Proceedings of the 38th International Conference on Machine Learning*. ICML 2021 - 38th International Conference on Machine Learning. Vol. 139. Virtual, France, 18th July 2021, pp. 1147–1157. URL: https://hal.archives-ouvertes.fr/hal-03383710.

[13] A. Camuto, G. Deligiannidis, M. A. Erdogdu, M. Gürbüzbalaban, U. Şimşekli and L. Zhu. 'Fractal Structure and Generalization Properties of Stochastic Optimization Algorithms'. In: NeurIPS 2021 - Thirty-fifth Conference on Neural Information Processing Systems. Virtual, France, 3rd Dec. 2021. URL: https://hal.inria.fr/hal-03530568.

[14] A. Camuto, X. WANG, L. Zhu, C. Holmes, M. Gürbüzbalaban and U. Şimşekli. 'Asymmetric Heavy Tails and Implicit Bias in Gaussian Noise Injections'. In: ICML 2021 - Thirty-eighth annual conference International Conference on Machine Learning. Virtual, France, 18th July 2021. URL: https://hal.inria.fr/hal-03530315.

[15] O. Cífka, A. Ozerov, U. Şimşekli and G. Richard. 'Self-Supervised VQ-VAE for One-Shot Music Style Transfer'. In: *ICASSP 2021 - IEEE International Conference on Acoustics, Speech and Signal Processing*. ICASSP 2021 - IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto / Virtual, Canada, 6th June 2021. DOI: 10.1109/ICASSP39728.2021.9414235. URL: https://hal.telecom-paris.fr/hal-03132940.

[16] R.-A. Dragomir, H. Hendrikx and M. Even. 'Fast Stochastic Bregman Gradient Methods: Sharp Analysis and Variance Reduction'. In: ICML 2021- 38th International Conference on Machine Learning. Vol. 139. Proceedings of the 38th International Conference on Machine Learning. virtual, United States, 1st July 2021, pp. 2815–2825. URL: https://hal.archives-ouvertes.fr/hal-03383164.

[17] M. Even, R. Berthier, F. Bach, N. Flammarion, P. Gaillard, H. Hendrikx, L. Massoulié and A. Taylor. 'A Continuized View on Nesterov Acceleration for Stochastic Gradient Descent and Randomized Gossip'. In: *Advances in Neural Information Processing Systems 34*. NeurIPS 2021 - 35th Conference on Neural Information Processing Systems. Sydney (virtual), Australia: Morgan Kaufmann Publishers, 1st Dec. 2021, pp. 1–32. URL: https://hal.archives-ouvertes.fr/hal-03405165.

[18] M. Gurbuzbalaban, U. Şimşekli and L. Zhu. 'The Heavy-Tail Phenomenon in SGD'. In: ICML2021 - Thirty-eighth annual conference on International Conference on Machine Learning. Virtual, France, 18th July 2021. URL: https://hal.inria.fr/hal-03530319.

[19] h. daneshmand hadi, A. Joudaki and F. Bach. 'Batch Normalization Orthogonalizes Representations in Deep Random Networks'. In: NeurIPS 2021 - 35th Conference on Neural Information Processing Systems. Virtual, France, 6th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03454243.

[20] R. Jézéquel, P. Gaillard and A. Rudi. 'Mixability made efficient: Fast online multiclass logistic regression'. In: NeurIPS 2021. Thirty-fifth Conference on Neural Information Processing Systems. Online, France, 6th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03370530.

[21] A. Liutkus, O. Cífka, S.-L. Wu, U. Şimşekli, Y.-H. Yang and G. Richard. 'Relative Positional Encoding for Transformers with Linear Complexity'. In: ICML 2021 - 38th International Conference on Machine Learning. Proceedings of the 38th International Conference on Machine Learning. Virtual Only, United States, 18th July 2021. URL: https://hal.telecom-paris.fr/hal-03256451.

[22] G. Mialon, D. Chen, A. D'Aspremont and J. Mairal. 'A Trainable Optimal Transport Embedding for Feature Aggregation and its Relationship to Attention'. In: ICLR 2021 - The Ninth International Conference on Learning Representations. Virtual, France, 4th May 2021. URL: https://hal.archives-ouvertes.fr/hal-02883436.

[23] K. Nadjahi, A. Durmus, P. E. Jacob, R. Badeau and U. Şimşekli. 'Fast Approximation of the Sliced-Wasserstein Distance Using Concentration of Random Projections'. In: NeurIPS 2021 - Thirty-fifth Conference on Neural Information Processing Systems. Virtual, France, 6th Dec. 2021. URL: https://hal.telecom-paris.fr/hal-03494781.

[24] H. Wang, M. Gürbüzbalaban, L. Zhu, U. Şimşekli and M. A. Erdogdu. 'Convergence Rates of Stochastic Gradient Descent under Infinite Noise Variance'. In: NeurIPS 2021 - Thirty-fifth Conference on Neural Information Processing Systems. Virtual, France, 3rd Dec. 2021. URL: https://hal.inria.fr/hal-03530384.

[25] P. Zhang, A. Orvieto and h. daneshmand hadi. 'Rethinking the Variational Interpretation of Accelerated Optimization Methods'. In: NeurrIPS 2021. Virtual, Unknown Region, 1st Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03454342.

[26] P. Zhang, A. Orvieto, h. daneshmand hadi, T. Hofmann and R. Smith. 'Revisiting the Role of Euler Numerical Integration on Acceleration and Stability in Convex Optimization'. In: AISTATS 2021 - 24th International Conference on Artifi-cial Intelligence and Statistics. Virtual, Unknown Region, 13th Apr. 2021. URL: https://hal.archives-ouvertes.fr/hal-03454377.

Conferences without proceedings

[27] E. Berthier, J. Carpentier and F. Bach. 'Fast and Robust Stability Region Estimation for Nonlinear Dynamical Systems'. In: European Control Conference (ECC) 2021. Rotterdam, Netherlands, 29th June 2021. URL: https://hal.archives-ouvertes.fr/hal-02984348.

[28] G. Izacard and E. Grave. 'Distilling Knowledge from Reader to Retriever for Question Answering'. In: ICLR 2021 - 9th International Conference on Learning Representations. Vienna, Austria, 4th May 2021. URL: https://hal.archives-ouvertes.fr/hal-03463398.

[29] G. Izacard and E. Grave. 'Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering'. In: EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics. Kiev, Ukraine: Association for Computational Linguistics, 19th Apr. 2021, pp. 874–880. DOI: 10.18653/v1/2021.eacl-main.74. URL: https://hal.archives-ouvertes.fr/hal-03463108.

[30] A. Vacher, B. Muzellec, A. Rudi, F. Bach and F.-X. Vialard. 'A Dimension-free Computational Upper-bound for Smooth Optimal Transport Estimation'. In: COLT 2021 - 34th Annual Conference on Learning Theory. Boulder, United States, 15th Aug. 2021. URL: https://hal.archives-ouvertes.fr/hal-03454237.

Doctoral dissertations and habilitation theses

[31] R.-A. Dragomir. 'Bregman Gradient Methods for Relatively-Smooth Optimization'. UT1 Capitole, 14th Sept. 2021. URL: https://hal.inria.fr/tel-03389344.

[32] H. Hendrikx. 'Accelerated Methods for Distributed Optimization'. PSL, 20th Sept. 2021. URL: https://hal.archives-ouvertes.fr/tel-03475383.

Reports & preprints

[33] A. Askari, A. D'Aspremont and L. E. Ghaoui. *Approximation Bounds for Sparse Programs*. 10th Mar. 2021. URL: https://hal.archives-ouvertes.fr/hal-03165622.

[34] F. Bach and L. Chizat. *Gradient Descent on Infinitely Wide Neural Networks: Global Convergence and Generalization*. 15th Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03379011.

[35] M. Barré, C. Giron, M. Mazzolini and A. D'Aspremont. *Averaging Atmospheric Gas Concentration Data using Wasserstein Barycenters*. 10th Mar. 2021. URL: https://hal.archives-ouvertes.fr/hal-03165617.

[36] M. Barré, A. Taylor and F. Bach. *A note on approximate accelerated forward-backward methods with absolute and relative errors, and possibly strongly convex objectives*. 14th Oct. 2021. URL: https://hal.inria.fr/hal-03377374.

[37] E. Berthier, J. Carpentier, A. Rudi and F. Bach. *Infinite-Dimensional Sums-of-Squares for Optimal Control*. 14th Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03377120.

[38] R. Berthier, F. Bach, N. Flammarion, P. Gaillard and A. Taylor. *A Continuized View on Nesterov Acceleration*. 11th Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03138823.

[39] L. Brogat-Motte, A. Rudi, C. Brouard, J. Rousu and F. d'Alché-Buc. *Learning Output Embeddings in Structured Prediction*. 9th July 2021. URL: https://hal.inrae.fr/hal-03282445.

[40] A. D'Aspremont, D. Scieur and A. Taylor. *Acceleration Methods*. 1st Mar. 2021. URL: https://hal.inria.fr/hal-03154589.

[41] A. Défossez, N. Usunier, L. Bottou and F. Bach. *Music Source Separation in the Waveform Domain*. 28th Apr. 2021. URL: https://hal.archives-ouvertes.fr/hal-02379796.

[42] Y. Drori and A. Taylor. *On the oracle complexity of smooth strongly convex minimization*. 1st Mar. 2021. URL: https://hal.inria.fr/hal-03154582.

[43] C. Gerbelot and R. Berthier. *Graph-based Approximate Message Passing Iterations*. 22nd Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03397846.

[44] B. Goujaud, D. Scieur, A. Dieuleveut, A. Taylor and F. Pedregosa. *Super-Acceleration with Cyclical Step-sizes*. 14th Oct. 2021. URL: https://hal.inria.fr/hal-03377367.

[45] T. Kerdreux, A. D'Aspremont and S. Pokutta. *Local and Global Uniform Convexity Conditions*. 10th Mar. 2021. URL: https://hal.archives-ouvertes.fr/hal-03165620.

[46] T. Kerdreux, C. Roux, A. D'Aspremont and S. Pokutta. *Linear Bandits on Uniformly Convex Sets*. 15th Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03379835.

[47] M. Lambert, S. Bonnabel and F. Bach. *The limited-memory recursive variational Gaussian approximation (L-RVGA)*. 23rd Dec. 2021. URL: https://hal.inria.fr/hal-03501920.

[48] T. Lauvaux, C. Giron, M. Mazzolini, A. D'Aspremont, R. Duren, D. Cusworth, D. Shindell and P. Ciais. *Global assessment of oil and gas methane ultra-emitters*. 15th Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03379815.

[49] U. Marteau-Ferey, F. Bach and A. Rudi. *Sampling from Arbitrary Functions via PSD Models*. 27th Oct. 2021. URL: https://hal.inria.fr/hal-03386544.

[50] G. Mialon, D. Chen, M. Selosse and J. Mairal. *GraphiT: Encoding Graph Structure in Transformers*. 10th June 2021. URL: https://hal.archives-ouvertes.fr/hal-03256708.

[51] B. Muzellec, F. Bach and A. Rudi. *A Note on Optimizing Distributions using Kernel Mean Embeddings*. 18th June 2021. URL: https://hal.archives-ouvertes.fr/hal-03454259.

[52] B. Muzellec, F. Bach and A. Rudi. *Learning PSD-valued functions using kernel sums-of-squares*. 22nd Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03454277.

[53] B. Muzellec, A. Vacher, F. Bach, F.-X. Vialard and A. Rudi. *Near-optimal estimation of smooth transport maps with kernel sums-of-squares*. 3rd Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03466696.

[54] M. Romain and A. D'Aspremont. *A Bregman Method for Structure Learning on Sparse Directed Acyclic Graphs*. 10th Mar. 2021. URL: https://hal.archives-ouvertes.fr/hal-03165618.

[55] A. Taylor and Y. Drori. *An optimal gradient method for smooth (possibly strongly) convex minimization*. 1st Mar. 2021. URL: https://hal.inria.fr/hal-03154583.

[56] S. Vaswani, B. Dubois-Taine and R. Babanezhad. *Towards Noise-adaptive, Problem-adaptive Stochastic Gradient Descent*. 30th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03456663.

[57] H. Zenati, A. Bietti, M. Martin, E. Diemert and J. Mairal. *Counterfactual Learning of Stochastic Policies with Continuous Actions: from Models to Offline Evaluation*. 19th Aug. 2021. URL: https://hal.archives-ouvertes.fr/hal-02883423.