LOL 2022 – Learning and Optimization in Luminy
Workshop CIRM

Organizers

- Aymeric Dieuleveut
- Claire Boyer
- Thomas Moreau
- Soledad Villar
- Alexandre d'Aspremont

Schedule

Monday

09:00-09:45	Anders Hansen
09:45-10:30	Jérôme Bolte
10:30-11:00	**Coffee break**
11:00-11:45	Raphaël Berthier
11:45-12:30	Francis Bach
12:30-14:00	**Lunch**
16:00-16:50	Jelani Nelson
16:55-17:45	Irène Waldspurger
17:50-18:40	Pierre Laforgue
18:45-19:25	Lightning Talks
19:30	**Diner**

Tuesday

09:00-09:45	Edouard Pauwels
09:45-10:30	Gabriele Steidl
10:30-11:00	**Coffee break**
11:00-11:45	Kevin Scaman
11:45-12:30	Stéphane Chrétien
12:30-14:00	**Lunch**
16:00-16:25	Hugo Cui
16:25-16:50	Adeline Fermanian
16:50-17:15	Baptiste Goujaud
17:15-17:45	**Break**
17:45-18:10	Scott Pesme
18:10-18:35	Etienne Boursier
18:35-19:00	Mathurin Massias
19:30	**Diner**

Wednesday

09:00-09:45	Robert Gower
09:45-10:30	Joseph Salmon
10:30-11:00	**Coffee break**
11:00-11:45	Cristóbal Guzmán
11:45-12:30	Gersende Fort
12:30-14:00	**Lunch**
14:00-19:00	**Free Afternoon**
19:30	**Diner**

Thursday

09:00-09:45	Andrea Simonetto
09:45-10:30	Aurélien Bellet
10:30-11:00	**Coffee break**
11:00-11:45	Mikael Johansson
11:45-12:30	Hadrien Hendrikx
12:30-14:00	**Lunch**
16:00-16:50	Pierre Ablin
16:50-17:40	Michael Arbel
17:45-19:15	**Poster Session**
19:30	**Gala Diner**

Friday

09:00-09:45	Florentin Goyens
09:45-10:30	Mikhail Belkin
10:30-11:00	**Coffee break**
11:00-11:45	Arthur Mensch
11:45-12:30	Vianney Perchet
12:30-14:00	**Lunch**
14:00-15:00	**Goodbye :)**

Monday

09:00 - 09:45 – Anders Hansen

Title: Generalised hardness of approximation: On the extended Smale's 9th problem and the maths of "why things don't work"

Abstract: Alchemists wanted to create gold, Hilbert wanted an algorithm to solve Diophantine equations, researchers want to make deep learning robust in AI, MATLAB wants (but fails) to detect when it provides wrong solutions to linear programs, etc. Why do we fail in so many of these fundamental cases? The reason is typically methodological barriers. The history of science is full of methodological barriers — reasons for why we never succeed in reaching certain goals. In many cases, this is due to barriers in the foundations of mathematics. This talk introduces new such barriers from foundations: the phenomenon of generalised hardness of approximation (GHA). GHA grows out of our solution to the extended 9th problem from Smale's list of mathematical problems for the 21st century. This phenomenon is not a rare issue — but happens on a daily basis in computational mathematics — and causes modern software such as MATLAB to fail on basic problems, and even certify nonsensical solutions as correct.

GHA is close in spirit to hardness of approximation (HA) in computer science. Assuming P $\neq$ NP, HA is the phenomenon that one can easily compute an epsilon-approximation to the solution of a discrete computational problem for $\epsilon > \epsilon_0 > 0$, but for $\epsilon < \epsilon_0$ it suddenly becomes intractable. HA was discovered decades ago and has been transformative being the subject of several Goedel, Nevanlinna and ACM prizes. GHA is a similar but distinct mathematical phenomenon that requires a new set of mathematical tools in computational mathematics rather than computer science (NB: GHA is independent of P vs. NP). The GHA phenomenon has so far been detected in optimisation, inverse problems, deep learning and AI, as well as computer-assisted proofs. It is essential in the Solvability Complexity Index (SCI) hierarchy and for understanding "why things don't work", as well as a key tool to understand "why things sometimes work".

09:45 - 10:30 – Jérôme Bolte

Title: A Glance at Nonsmooth Automatic differentiation

Abstract: Nonsmooth automatic differentiation is one of the core learning mechanism in modern AI. I will show how a recent theory we developed — ≪Conservative gradients≫— helps to understand this process and related fundamental phenomena, such as the convergence of learning phases in deep learning, the optimization of learning parameters, the nonsmooth cheap gradient principle, or the differentiation of algorithms. Joint work with E. Pauwels

10:30 - 11:00 – Coffee break

11:00 - 11:45 – Raphaël Berthier

Title: Incremental Learning in Diagonal Linear Networks

Abstract: Diagonal linear networks (DLNs) are a toy simplification of artificial neural networks; they consist in a quadratic reparametrization of linear regression inducing a sparse implicit regularization. In this talk, I will describe the trajectory of the gradient flow of DLNs in the limit of small initialization. I will show that incremental learning is effectively performed in the limit: coordinates are successively activated, while the iterate is the minimizer of the loss constrained to have support on the active coordinates only. This shows that the sparse implicit regularization of DLNs decreases with time.

11:45 - 12:30 – Francis Bach

Title: Sum-of-Squares Relaxations for Information Theory and Variational Inference

Abstract: We consider extensions of the Shannon relative entropy, referred to as f-divergences. Three classical related computational problems are typically associated with these divergences: (a) estimation from moments, (b) computing normalizing integrals, and (c) variational inference in probabilistic models. These problems are related to one another through convex duality, and for all them, there are many applications throughout data science, and we aim for computationally tractable approximation algorithms that preserve properties of the original problem such as potential convexity or monotonicity. In order to achieve this, we derive a sequence of convex relaxations for computing these divergences from non-centered covariance matrices associated with a given feature vector: starting from the typically non-tractable optimal lower-bound, we consider an additional relaxation based on "sums-of-squares", which is is now computable in polynomial time as a semidefinite program, as well as further computationally more efficient relaxations based on spectral information divergences from quantum information theory (see https://arxiv.org/abs/2206.13285 for details).

12:30 - 14:00 – Lunch

16:00 - 16:50 – Jelani Nelson

Title: Private Frequency Estimation via Projective Geometry

Abstract: Many of us use smartphones and rely on tools like auto-complete and spelling auto-correct to make using these devices more pleasant, but building these tools presents a challenge. On the one hand, the machine-learning algorithms used to provide these features require data to learn from, but on the other hand, who among us is willing to send a carbon copy of all our text messages to device manufacturers to provide that data? "Local differential privacy" and related concepts have emerged as the gold standard model in which to analyze tradeoffs between losses in utility and privacy for solutions to such problems. In this talk, we give a new state-of-the-art algorithm for estimating histograms of user data, making use of projective geometry over finite fields coupled with a reconstruction algorithm based on dynamic programming.

This talk is based on joint work with Vitaly Feldman (Apple), Huy Le Nguyen (Northeastern), and Kunal Talwar (Apple).

16:55 - 17:45 – Irène Waldspurger

Title: Sketching semidefinite programs for super-resolution problems

Abstract: In this talk, we will consider the canonical example of a super-resolution problem: the recovery of a measure on [0;1] from its first Fourier coefficients, assuming that the measure is the sum of a few spikes. Under weak assumptions, it is known that the measure to be recovered is the solution of a convex infinite-dimensional problem, which is in turn equivalent to a semidefinite program. This property yields a polynomial-time reconstruction algorithm with strong correctness guarantees.

Unfortunately, the size of the semidefinite program can be extremely large, even when the measure contains a very small number of spikes. I will present a sketching approach to reduce this size. Proving that this approach retains the correctness guarantees is still an ongoing work. I will present a byproduct of our efforts to find a proof, namely an algorithm to automatically find (simple) upper bounds on some integrals with parameters.

This work is a collaboration with Augustin Cosse and Gabriel Peyré.

17:50 - 18:40 – Pierre Laforgue

Title: Multitask Online Mirror Descent

Abstract: We introduce and analyze MT-OMD, a multitask generalization of Online Mirror Descent (OMD) which operates by sharing updates between tasks. We prove that the regret of MT-OMD is of order $\sqrt{1+\sigma^2(N-1)}\sqrt{T}$, where σ^2 is the task variance according to the geometry induced by the regularizer, N is the number of tasks, and T is the time horizon. Whenever tasks are similar, that is $\sigma^2 \leq 1$, our method improves upon the $\sqrt{NT}$ bound obtained by running independent OMDs on each task. We further provide a matching lower bound, and show that our multitask extensions of Online Gradient Descent and Exponentiated Gradient, two major instances of OMD, enjoy closed-form updates, making them easy to use in practice. Finally, we present experiments which support our theoretical findings. Paper: https://arxiv.org/abs/2106.02393

18:45 - 19:25 – Lightning Talks

LT1 – Le bars Baptiste *Refined convergence and topology learning for Decentralized SGD with heterogeneous data*

Abstract: One of the key challenges in decentralized and federated learning is to design algorithms that efficiently deal with highly heterogeneous data distributions across agents. In this work, we revisit the analysis of Decentralized Stochastic Gradient Descent algorithm (D-SGD) under data heterogeneity. We first exhibit the key role played by a new quantity, called neighborhood heterogeneity, on the convergence rate of D-SGD. We then argue that neighborhood heterogeneity provides a natural criterion to learn data-dependent topologies that reduce the detrimental effect of data heterogeneity on the convergence of D-SGD.

LT2 – Mishenko Konstantin *Super-Universal Regularized Newton Method*

Abstract: We analyze the performance of a variant of Newton method with quadratic regularization for solving composite convex minimization problems. At each step of our method, we choose regularization parameter proportional to a certain power of the gradient norm at the current point. We introduce a family of problem classes characterized by Hölder continuity of either the second or third derivative. Then we present the method with a simple adaptive search procedure allowing an automatic adjustment to the problem class with the best global complexity bounds, without knowing specific parameters of the problem. In particular, for the class of functions with Lipschitz continuous third derivative, we get the global $O(1/k^3)$ rate, which was previously attributed to third-order tensor methods. When the objective function is uniformly convex, we justify an automatic acceleration of our scheme, resulting in a faster global rate and local superlinear convergence. The switching between the different rates (sublinear, linear, and superlinear) is automatic. Again, for that, no a priori knowledge of parameters is needed.

LT3 – Lefort Tanguy *Improve learning combining crowsourced labels by weighting area under the margin*

Abstract: In supervised learning – for instance in image classification – modern massive datasets are commonly labelled by a crowd of workers. The obtained labels in this crowdsourcing setting are then aggregated for training. The aggregation step generally leverages a per worker trust score. Yet, such worker-centric approaches discard each task ambiguity. Some intrinsically ambiguous tasks might even fool expert workers, which could eventually be harmful for the learning step. In a stan- dard supervised learning setting – with one label per task and balanced classes – the Area Under the Margin (AUM) statistic is tailored to identify mislabeled data. We adapt the AUM to identify ambiguous tasks in crowdsourced learning scenar- ios, introducing the Weighted AUM (WAUM). The WAUM is an average of AUMs weighted by worker and task dependent scores. We show that the WAUM can help discarding ambiguous tasks from the training set.

LT4 – Varre Aditya *Accelerated SGD for Non-Strongly-Convex Least Squares*
Abstract: We consider stochastic approximation for the least squares regression problem in the non-strongly convex setting. We present the first practical algorithm that achieves the optimal prediction error rates in terms of dependence on the noise of the problem, as $O(d/t)$ while accelerating the forgetting of the initial conditions to $O(d/t^2)$. Our new algorithm is based on a simple modification of the accelerated gradient descent. We provide convergence results for both the averaged and the last iterate of the algorithm. In order to describe the tightness of these new bounds, we present a matching lower bound in the noiseless setting and thus show the optimality of our algorithm.

LT5 – Lezane Clement *Optimal Algorithms for Composite Stochastic Mirror Descent*
Abstract: Inspired by regularization techniques in statistics and machine learning, we study complementary composite minimization in the stochastic setting. This problem corresponds to the minimization of the sum of a (weakly) smooth function endowed with a stochastic first-order oracle, and a structured uniformly convex (possibly nonsmooth and non-Lipschitz) regularization term. Despite intensive work on closely related settings, prior to our work no complexity bounds for this problem were known. We close this gap by providing novel excess risk bounds, both in expectation and with high probability. Our algorithms are nearly optimal, which we prove via novel lower complexity bounds for this class of problems. We conclude by providing numerical results comparing our methods to the state of the art.

LT6 – Rebjock Quentin *On the relationship between the Polyak-Lojasiewicz inequality and the Morse-Bott property*
Abstract: Lojasiewicz inequalities are ubiquitous in optimization; they are a powerful tool to analyze the iterations produced by an algorithm in the vicinity of critical points. A C2 function is Morse-Bott at a critical point if, locally, the critical set is a submanifold and the Hessian is non-degenerate in the normal spaces. It is easy to see that if a C2 function is Morse-Bott at a critical point then it satisfies Polyak-Lojasiewicz (PL) around that point. In this talk we show that the converse is also true: if a C2 function satisfies the PL inequality around a critical point then the function is Morse-Bott at that point. This offers a dual perspective to prove local convergence guarantees for second-order algorithms.

LT7 – Philippenko Constantin *Preserved central model for faster bidirectional compression in distributed settings*
Abstract: We develop a new approach to tackle communication constraints in a distributed learning problem with a central server. We propose and analyze a new algorithm that performs bidirectional compression and achieves the same convergence rate as algorithms using only uplink (from the local workers to the central server) compression. To obtain this improvement, we design MCM, an algorithm such that the downlink compression only impacts local models, while the global model is preserved. As a result, and contrary to previous works, the gradients on local servers are computed on perturbed models. Consequently, convergence proofs are more challenging and require a precise control of this perturbation. To ensure it, MCM additionally combines model compression with a memory mechanism. This analysis opens new doors, e.g. incorporating worker dependent randomized-models and partial participation.

LT8 – Paul-Dubois-Taine Benjamin *Fast Stochastic Composite Minimization and an Accelerated Frank Wolfe Algorithm under Parallelizationf*

Abstract: We consider the problem of minimizing the sum of two convex functions. One of those functions has Lipschitz-continuous gradients, and can be accessed via stochastic oracles, whereas the other is "simple". We provide a Bregman-type algorithm with accelerated convergence in function values to a ball containing the minimum. The radius of this ball depends on problem-dependent constants, including the variance of the stochastic oracle. We further show that this algorithmic setup naturally leads to a variant of Frank-Wolfe achieving acceleration under parallelization. More precisely, when minimizing a smooth convex function on a bounded domain, we show that one can achieve an O primal-dual gap (in expectation) in $O(1/\sqrt{\epsilon})$ iterations, by only accessing gradients of the original function and a linear maximization oracle with $O(1/\sqrt{\epsilon})$ computing units in parallel. We illustrate this fast convergence on synthetic numerical experiments.

LT9 – Marion Pierre *Framing RNN as a kernel method: a neural ODE approach*
Abstract: Building on the interpretation of a recurrent neural network (RNN) as a continuous-time neural differential equation, we show, under appropriate conditions, that the solution of a RNN can be viewed as a linear function of a specific feature set of the input sequence, known as the signature. This connection allows us to frame a RNN as a kernel method in a suitable reproducing kernel Hilbert space. As a consequence, we obtain theoretical guarantees on generalization and stability for a large class of recurrent networks.

LT10 – Richard Hugo *Scheduling with predictions*
Abstract: A popular approach to go beyond the worst-case analysis of online algorithms is to assume the existence of predictions that can be leveraged to improve performances. Those predictions are usually given by some external sources that cannot be fully trusted. Instead, we argue that trustful predictions can be built by algorithms, while they run. We investigate this idea in the illustrative context of static scheduling with exponential job sizes. Indeed, we prove that algorithms agnostic to this structure do not perform better than in the worst case. In contrast, when the expected job sizes are known, we show that the best algorithm using this information, called Follow-The-Perfect-Prediction (FTPP), exhibits much better performances. Then, we introduce two adaptive explore-then-commit types of algorithms: they both first (partially) learn expected job sizes and then follow FTPP once their self-predictions are confident enough. On the one hand, ETCU explores in ”series”, by completing jobs sequentially to acquire information. On the other hand, ETCRR, inspired by the optimal worst-case algorithm Round-Robin (RR), explores efficiently in ”parallel”. We prove that both of them asymptotically reach the performances of FTPP, with a faster rate for ETCRR. Those findings are empirically evaluated on synthetic data.

LT11 – Mourcer Céline *A systematic approach to Lyapunov analyses of continuous-time models in convex optimization*
Abstract: First-order methods are often analyzed via their continuous-time models, where their worst-case convergence properties are usually approached via Lyapunov functions. In this work, we provide a systematic and principled approach to find and verify Lyapunov functions for classes of ordinary and stochastic differential equations. More precisely, we extend the performance estimation framework, originally proposed by Drori and Teboulle [14], to continuous-time models. We retrieve convergence results comparable to those of discrete-time methods using fewer assumptions and convexity inequalities, and provide new results for a family of stochastic accelerated gradient flows.

LT12 – Arnould Ludovic *Interpolation and Random Forests*
Abstract: Statistical wisdom suggests that very complex models, interpolating training data, will be poor at predicting unseen examples. Yet, this aphorism has been recently challenged by the identification of benign overfitting regimes, specially studied in the case of parametric models: generalization capabilities may be preserved despite model high complexity. While it is widely known that fully-grown

decision trees interpolate and, in turn, have bad predictive performances, the same behavior is yet to be analyzed for random forests. In this paper, we study the trade-off between interpolation and consistency for several types of random forest algorithms. Theoretically, we prove that interpolation regimes and consistency cannot be achieved simultaneously for non-adaptive random forests. Since adaptivity seems to be the cornerstone to bring together interpolation and consistency, we study interpolating Median RF which are proved to be consistent in a noiseless scenario. Numerical experiments show that Breiman's random forests are consistent while exactly interpolating, when no bootstrap step is involved. We theoretically control the size of the interpolation area, which converges fast enough to zero, so that exact interpolation and consistency occur in conjunction.

LT13 – Mangold Paul *High-Dimensional Private Empirical Risk Minimization by Greedy Coordinate Descent*
Abstract: In differentially private empirical risk minimization (DP-ERM), worst-case utility decreases as the dimension increases. This is a major obstacle to privately learning large machine learning models. In high dimension, some model's parameters typically carry more information than others. We propose a differentially private greedy coordinate descent (DP-GCD) algorithm that can exploit this property to reduce the dependence on the dimension.

LT14 – Ayme Alexis *Near-optimal rate of consistency for linear models with missing values*
Abstract: Missing values arise in most real-world data sets due to the aggregation of multiple sources and intrinsically missing information (sensor failure, unanswered questions in surveys...). In fact, the very nature of missing values usually prevents us from running standard learning algorithms. In this talk, we focus on the extensively-studied linear models, but in presence of missing values, which turns out to be quite a challenging task. Indeed, the Bayes rule can be decomposed as a sum of predictors corresponding to each missing pattern. This eventually requires to solve a number of learning tasks, exponential in the number of input features, which makes predictions impossible for current real-world datasets. First, we propose a rigorous setting to analyze a least-square type estimator and establish a bound on the excess risk which increases exponentially in the dimension. Consequently, we leverage the missing data distribution to propose a new algorithm, and derive associated adaptive risk bounds that turn out to be minimax optimal.

19:30 – Diner

Tuesday

09:00 - 09:45 – Edouard Pauwels

Title: Curiosities and counterexamples in smooth convex optimization

Abstract: We present a list of counterexamples to conjectures in smooth convex coercive optimization. We will detail two extensions of the gradient descent method, of interest in machine learning: gradient descent with exact line search, and Bregman descent (also known as mirror descent). We show that both are non convergent in general. These examples are based on general smooth convex interpolation results. Given a decreasing sequence of convex compact sets in the plane, whose boundaries are Ck curves (k ¿ 1, arbitrary) with positive curvature, there exists a Ck convex function for which each set of the sequence is a sublevel set. The talk will provide proof arguments for this results and detail how it can be used to construct the anounced counterexamples.

09:45 - 10:30 – Gabriele Steidl

Title: Stochastic Normalizing Flows and the Power of Patches in Inverse Problems

Abstract: Learning neural networks using only a small amount of data is an important ongoing research topic with tremendous potential for applications. We introduce a regularizer for the variational modeling of inverse problems in imaging based on normalizing flows, called patchNR. It involves a normalizing flow learned on patches of very few images. The subsequent reconstruction method is completely unsupervised and the same regularizer can be used for different forward operators acting on the same class of images. By investigating the distribution of patches versus those of the whole image class, we prove that our variational model is indeed a MAP approach. Numerical examples for low-dose CT, limited-angle CT and superresolution of material images demonstrate that our method provides high quality results among unsupervised methods, but requires only very few data. Further, the appoach also works if only the low resolution image is available.

In the second part of the talk I will generalize normalizing flows to stochastic normalizing flows to improve their expressivity.Normalizing flows, diffusion normalizing flows and variational autoencoders are powerful generative models. A unified framework to handle these approaches appear to be Markov chains. We consider stochastic normalizing flows as a pair of Markov chains fulfilling some properties and show how many state-of-the-art models for data generation fit into this framework. Indeed including stochastic layers improves the expressivity of the network and allows for generating multimodal distributions from unimodal ones. The Markov chains point of view enables us to couple both deterministic layers as invertible neural networks and stochastic layers as Metropolis-Hasting layers, Langevin layers, variational autoencoders and diffusion normalizing flows in a mathematically sound way. Our framework establishes a useful mathematical tool to combine the various approaches.

Joint work with F. Altekrüger, A. Denker, P. Hagemann, J. Hertrich, P. Maass

10:30 - 11:00 – Coffee break

11:00 - 11:45 – Kevin Scaman

Title: Non-convex SGD and Lojasiewicz-type conditions for deep learning

Abstract: First-order non-convex optimization is at the heart of neural networks training. Recent analyses showed that the Polyak-Lojasiewicz condition is particularly well-suited to analyze the convergence of the training error for these architectures. In this short presentation, I will propose extensions of this condition that allows for more flexibility and application scenarios, and show how stochastic gradient descent converges under these conditions. Then, I will show how to use these conditions to prove the convergence of the test error for simple deep learning architectures in an online setting.

11:45 - 12:30 – Stéphane Chrétien

Title: Relationship between sample size and architecture for the estimation of Sobolev functions using deep neural networks

Abstract: The statistical problem of estimating a Sobolev function using a deep network is studied using the Neuberger theorem and recent approximation results by Gurhing Kutyniok and Petersen. The problem is addressed by decoupling the statistical and the approximation problems and is shown to boil down to the computation of the Sobolev norm of bump functions

12:30 - 14:00 – Lunch

16:00 - 16:25 – Hugo Cui

Title: Error rates for kernel methods under source and capacity conditions

Abstract: We investigate the rates of decay of the prediction error for kernel methods under the Gaussian design and source/capacity assumptions. For kernel ridge regression, we derive all the observable rates, and characterize the regimes in which each hold. In particular, we show that the decay rate may transition from a fast, noiseless rate to a slow, noisy rate as the sample complexity is increased. For noiseless kernel classification, we derive the rates for two standard classifiers, margin-maximizing SVMs and ridge classifiers, and contrast the two methods. In both cases, the derived rates also describe to a good degree the learning curves of a number of real datasets. This is joint work with Bruno Loureiro, Florent Krzakala and Lenka Zeborová.

16:25 - 16:50 – Adeline Fermanian

Title: Scaling ResNets in the Large-depth Regime

Abstract: Deep ResNets are recognized for achieving state-of-the-art results in complex machine learning tasks. However, the remarkable performance of these architectures relies on a training procedure that needs to be carefully crafted to avoid vanishing or exploding gradients, particularly as the depth L increases. No consensus has been reached on how to mitigate this issue, although a widely discussed strategy consists in scaling the output of each layer by a factor α_L. We show in a probabilistic setting that with standard i.i.d. initializations, the only non-trivial dynamics is for $\alpha_L = 1/\sqrt{L}$ (other choices lead either to explosion or to identity mapping). This scaling factor corresponds in the continuous-time limit to a neural stochastic differential equation, contrarily to a widespread interpretation that deep ResNets are discretizations of neural ordinary differential equations. By contrast, in the latter regime, stability is obtained with specific correlated initializations and $\alpha_L = 1/L$. Our analysis suggests a strong interplay between scaling and regularity of the weights as a function of the layer index. Finally, in a series of experiments, we exhibit a continuous range of regimes driven by these two parameters, which jointly impact performance before and after training.

16:50 - 17:15 – Baptiste Goujaud

Title: PEPit: a computer assistant to study first order optimization

Abstract: In recent years, a general performance estimation framework has been developed to study more easily the optimization algorithms' performance. This framework, called PEP, has been implemented in Matlab under the name Pesto. In a matter of performance and mostly to help more researchers, we are proposing a version of this framework in the most widely used language, namely Python. In this talk I will mainly introduce the principles of PEP, the theoretical framework. Then, I will quickly show one recent result obtained through the use of PEPit.

17:15 - 17:45 – Break

17:45 - 18:10 – Scott Pesme

Title: Implicit Bias of SGD for Diagonal Linear Networks: a Provable Benefit of Stochasticity

Abstract: Understanding the implicit bias of training algorithms is of crucial importance in order to explain the success of overparametrised neural networks. We study the dynamics of stochastic gradient descent over diagonal linear networks through its continuous time version, namely stochastic gradient flow. We explicitly characterise the solution chosen by the stochastic flow and prove that it always enjoys better generalisation properties than that of gradient flow. Quite surprisingly, we show that the convergence speed of the training loss controls the magnitude of the biasing effect: the slower the convergence, the better the bias. To fully complete our analysis, we provide convergence guarantees for the dynamics. We also give experimental results which support our theoretical claims. Our findings highlight the fact that structured noise can induce better generalisation and they help explain the greater performances of stochastic gradient descent over gradient descent observed in practice.

18:10 - 18:35 – Etienne Boursier

Title: Gradient flow dynamics of shallow ReLU networks for square loss and orthogonal inputs

Abstract: The training of neural networks by gradient descent methods is a cornerstone of the deep learning revolution. Yet, despite some recent progress, a complete theory explaining its success is still missing. This article presents, for orthogonal input vectors, a precise description of the gradient flow dynamics of training one-hidden layer ReLU neural networks for the mean squared error at small initialisation. In this setting, despite non-convexity, we show that the gradient flow converges to zero loss and characterise its implicit bias towards minimum variation norm. Furthermore, some interesting phenomena are highlighted: a quantitative description of the initial alignment phenomenon and a proof that the process follows a specific saddle to saddle dynamics.

18:35 - 19:00 – Mathurin Massias

Title: Iterative regularization for low complexity regularizers

Abstract: Iterative regularization exploits the implicit bias of an optimization algorithm to regularize ill-posed problems. Constructing algorithms with such built-in regularization mechanisms is a classic challenge in inverse problems but also in modern machine learning, where it provides both a new perspective on algorithms analysis, and significant speed-ups compared to explicit regularization. In this work, we propose and study the first iterative regularization procedure able to handle biases described by non smooth and non strongly convex functionals, prominent in low-complexity regularization. Our approach is based on a primal-dual algorithm of which we analyze convergence and stability properties, even in the case where the original problem is unfeasible. The general results are illustrated considering the special case of sparse recovery with the ℓ_1 penalty. Our theoretical results are complemented by experiments showing the computational benefits of our approach.

19:30 – Diner

Wednesday

09:00 - 09:45 – Robert Gower

Title: Cutting Some Slack for SGD with Adaptive Polyak Stepsizes

Abstract: Tuning the step size of stochastic gradient descent is tedious and error prone. This has motivated the development of methods that automatically adapt the step size using readily available information. We investigate the family of SPS (Stochastic gradient with a Polyak Stepsize) adaptive methods for setting the step size. These are methods that make use of gradient and loss value at the sampled points to adaptively adjust the step size. We first show that SPS and its recent variants can all be seen as extensions of the Passive-Aggressive methods applied to nonlinear problems. We use this insight to develop new variants of the SPS method that are better suited to nonlinear models. Our new variants are based on introducing a slack variable into the interpolation equations. This single slack variable tracks the loss function across iterations and is used in setting a stable step size. We provide extensive numerical results supporting our new methods and a convergence theory.

09:45 - 10:30 – Joseph Salmon

Title: Stochastic smoothing of the top-K calibrated hinge loss for deep imbalanced classification

Abstract: Modern classification tasks can include several thousand of possibly very similar classes. One such example is the Pl@ntNet application, which aiStochastic smoothing of the top-K calibrated hinge loss for deep imbalanced classificationms at providing users with the correct plant species given an input image. In this context, the high ambiguity results in low top-1 accuracy. This motivates top-K classification, in which K possible classes are returned. Yet, proposing top-K losses (to minimize the top-K error) tailored for deep learning remains a challenge, both theoretically and practically. We will present a stochastic top-K hinge loss for deep learning inspired by recent developments on top-K calibrated losses. The proposal is based on the smoothing of the top-K operator building on the flexible ”perturbed optimizer” framework. We show that our loss function performs well for balanced datasets. In addition, we propose a simple variant of our loss to handle imbalanced cases and that significantly outperforms other baseline loss functions on Pl@ntNet-300K. The latter is an open dataset of plant images obtained from the Pl@ntNet application, characterized by high ambiguity and a long-tailed distribution, that we recently released.

This is joint work with Camille Garcin, Alexis Joly and Maximilien Servajean.

10:30 - 11:00 – Coffee break

11:00 - 11:45 – Cristóbal Guzmán

Title: Non-Euclidean Differentially Private Stochastic Convex Optimization

Abstract: In this talk, I will present upper and lower bounds on the excess risk for differentially private stochastic convex optimization in non-Euclidean settings, in particular for ℓ_p spaces. Interestingly, our upper bounds for the ell_1 case are near dimension independent, and for any $1 < p \leq 2$ we obtain a sharp transition of the excess risk, showing a polynomial dependence on dimension is necessary. Our algorithms are guaranteed to succeed both in expectation and with high probability. If time allows, I will discuss more recent work involving nonconvex losses and other related problems.

11:45 - 12:30 – Gersende Fort

Title: Stochastic Variable Metric Proximal Gradient with variance reduction for non-convex composite optimization

Abstract: This talk introduces a novel algorithm, the Perturbed Proximal Preconditioned SPIDER algorithm (3P-SPIDER) , designed to solve finite sum non-convex composite optimization. It is a stochastic Variable Metric Forward-Backward algorithm, which allows approximate preconditioned forward operator and uses a variable metric proximity operator as the backward operator; it also proposes a mini-batch strategy with variance reduction to address the finite sum setting. We show that 3P-SPIDER extends some Stochastic preconditioned Gradient Descent-based algorithms and some Incremental Expectation Maximization algorithms to composite optimization and to the case the forward operator can not be computed in closed form. We also provide an explicit control of convergence in expectation of 3P-SPIDER , and study its complexity in order to satisfy the approximate epsilon-stationary condition. Our results are the first to combine the composite optimization setting, the non-convex one, a variance reduction technique to tackle the finite sum setting by using a minibatch strategy and, to allow deterministic or random approximations of the preconditioned forward operator. Finally, through an application to inference in a logistic regression model with random effects, we numerically compare 3P-SPIDER to other stochastic forward-backward algorithms and discuss the role of some design parameters of 3P-SPIDER. This is a joint work with Eric Moulines (Ecole Polytechnique, CMAP, France). Talk based on the paper HAL-03781216.

12:30 - 14:00 – Lunch

14:00 - 19:00 – Free Afternoon

19:30 – Diner

Thursday

09:00 - 09:45 – Andrea Simonetto

Title: Personalized Time-Varying Optimization

Abstract: We look at optimization problems that change over time and have a component that capture user's satisfaction to the decisions. The talk is going to be a blend of optimization algorithms and regression methods to learn user's satisfaction.

09:45 - 10:30 – Aurélien Bellet

Title: Better Privacy Guarantees for Decentralized Optimization

Abstract: Decentralized optimization is increasingly popular in machine learning for its scalability and efficiency. Intuitively, it should also provide better privacy guarantees, as nodes only observe the messages sent by their neighbors in the network graph. But formalizing and quantifying this gain is challenging: existing results are limited to Local Differential Privacy (LDP) guarantees that overlook the advantages of decentralization. In this talk, I will introduce appropriate relaxations of differential privacy and show how they can be used to show stronger privacy guarantees for gossip-based and random walk-based SGD, matching the privacy-utility trade-off of centralized SGD in some settings. Interestingly, some of these algorithms amplify privacy guarantees as a function of the distance between nodes in the graph, which aligns well with the privacy expectations of users in some use-cases. I will conclude with a discussion of open questions.

10:30 - 11:00 – Coffee break

11:00 - 11:45 – Mikael Johansson

Title: A Fast and Accurate Splitting Method for Optimal Transport: Analysis and Implementation

Abstract: We develop a fast and reliable method for solving large-scale optimal transport (OT) problems at an unprecedented combination of speed and accuracy.

Built on the celebrated Douglas-Rachford splitting technique, our method tackles the original OT problem directly instead of solving an approximate regularized problem, as many state-of-the-art techniques do. This allows us to provide sparse transport plans and avoid numerical issues of methods that use entropic regularization.

The algorithm has the same cost per iteration as the popular Sinkhorn method, and each iteration can be executed efficiently, in parallel. The proposed method enjoys an iteration complexity $O(1/\epsilon)$ compared to the best-known $O(1/\epsilon^2)$ of the Sinkhorn method. In addition, we establish a linear convergence rate for our formulation of the OT problem.

Finally, we detail an efficient GPU implementation of the proposed method that maintains a primal-dual stopping criterion at no extra cost. Substantial experiments demonstrate the effectiveness of our method, both in terms of computation times and robustness.

The talk is based on joint work with Vien V. Mai and Jacob Lindbäck.

11:45 - 12:30 – Hadrien Hendrikx

Title: Beyond spectral gap: The role of the topology in decentralized learning

Abstract: In data-parallel optimization of machine learning models, workers collaborate to improve their estimates of the model: more accurate gradients allow them to use larger learning rates and optimize faster. We consider the decentralized setting, in which all workers communicate over a sparse graph, and in which current theory fails to capture important aspects of real-world behavior. First, the 'spectral gap' of the communication graph is not predictive of its empirical performance in (deep) learning. Second, current theory does not explain that collaboration enables larger learning rates than training alone. In fact, it prescribes smaller learning rates, which further decrease as graphs become larger, failing to explain convergence in infinite graphs. This paper aims to paint an accurate picture of sparsely-connected distributed optimization. We quantify how the graph topology influences convergence in a quadratic toy problem and provide theoretical results for general smooth and (strongly) convex objectives. Our theory matches empirical observations in deep learning, and accurately describes the relative merits of different graph topologies.

12:30 - 14:00 – Lunch

16:00 - 16:50 – Pierre Ablin

Title: A framework for bilevel optimization that enables stochastic and global variance reduction algorithms

Abstract: Bilevel optimization, the problem of minimizing a value function which involves the arg-minimum of another function, appears across many areas of machine learning. The goal of this talk is to give an overview of some applications of bilevel optimization in machine learning, a description of the difficulties to develop large scale methods and of recent progress to overcome them.

16:50 - 17:40 – Michael Arbel

Title: Bilevel Optimization in Machine Learning: Beyond Strong Convexity

Abstract: Bilevel optimization problems involve two nested objectives, where an upper-level objective depends on a solution to a lower-level problem. In the first part of this talk, we consider the case when the lower-level problem is strongly convex and show that simple algorithms based on inexact implicit differentiation and a warm-start strategy can match the computational complexity of oracle methods that have access to an unbiased estimate of the gradient, thus outperforming many existing results for bilevel optimization. In the second part of the talk, we consider non-convex lower-level problems, with possibly multiple critical points, thus leading to an ambiguous definition of the bilevel problem. We introduce a key ingredient for resolving this ambiguity through the concept of a selection map which allows one to choose a particular solution to the lower-level problem. Using such maps, we define a class of hierarchical games between two agents that resolve the ambiguity in bilevel problems. We show that many existing algorithms for bilevel optimization, such as unrolled optimization, solve these games up to approximation errors due to finite computational power. This new class of games requires introducing new analytical tools in Morse theory to characterize their evolution. In particular, these tools allow us to study the differentiability of the selection, an essential property when analyzing gradient-based algorithms for solving these games.

17:45 - 19:15 – Poster Session

19:30 – Gala Diner

Friday

09:00 - 09:45 – Florentin Goyens

Title: Nonlinear matrix completion, denoising and registration

Abstract: Nonlinear matrix completion is the task of completing a partially observed matrix whose columns obey a nonlinear structure. Such matrices are in general full rank, but it is often possible to exhibit a low rank structure when the data is embedded in a higher dimensional space of features. We try to find the best formulation as a nonconvex optimisation problem and leverage existing results from matrix completion theory. Optimization methods on manifolds and alternative minimization algorithms are explored. We give convergence guarantees and provide numerical results for several test cases. With similar tools, we also address the problems of denoising and point cloud registration for a data set that is represented by an algebraic variety.

09:45 - 10:30 – Mikhail Belkin

Title: Neural networks, wide and deep, singular kernels and Bayes optimality.

Abstract: Wide and deep neural networks are used in many important practical setting. In this talk I will discuss some aspects of width and depth related to optimization and generalization. I will first discuss what happens when neural networks become infinitely wide, giving a general result for the transition to linearity (i.e., showing that neural networks become linear functions of parameters) for a broad class of wide neural networks corresponding to directed graphs.

I will then proceed to the question of depth, showing equivalence between infinitely wide and deep fully connected networks trained with gradient descent and Nadaraya-Watson predictors based on certain singular kernels. Using this connection we show that for certain activation functions these wide and deep networks are (asymptotically) optimal for classification but, interestingly, never for regression.

Based on joint work with Chaoyue Liu, Adit Radhakrishnan, Caroline Uhler and Libin Zhu.

10:30 - 11:00 – Coffee break

11:00 - 11:45 – Arthur Mensch

Title: Improving the efficiency of large language model training

Abstract: In recent years, natural language processing has owed most of its advances to the increase in model sizes. Today's best language models have more than 100B parameters; they have been trained on so much data that they can be conditioned at inference time into solving many different tasks. Such a trend in "scaling models" asks different questions.

First, we consider how to best choose the model size given a computational budget. By training over 500 language models ranging from 70M to over 16B parameters on 5 to 500 billion tokens, we find that for compute-optimal training, the model size and the number of training tokens should be scaled equally: for every doubling of model size the number of training tokens should also be doubled. We argue how this finding can be grounded in learning theory and use our optimality predictions to train a 70B model outperforming most of its much larger competitors.

Second, we show that fully parametric models may be inefficient at solving NLP tasks, and combine them with a non-parametric training mechanism. We enhance auto-regressive language models by conditioning on document chunks retrieved from a large corpus, based on local similarity with preceding tokens. With a 2 trillion token database, our Retrieval-Enhanced Transformer (RETRO) obtains comparable performance to GPT-3 and Jurassic-1 on the Pile, despite using 25x fewer parameters. We show that RETRO can be finetuned efficiently on NLP tasks, and discuss how it outlines train/test leakage issues in current language model training.

11:45 - 12:30 – Vianney Perchet

Title: An algorithmic solution to the Blotto game using multi-marginal couplings

Abstract: We describe an efficient algorithm to compute solutions for the general two-player Blotto game on n battlefields with heterogeneous values. While explicit constructions for such solutions have been limited to specific, largely symmetric or homogeneous, setups, this algorithmic resolution covers the most general situation to date: value-asymmetric game with asymmetric budget. The proposed algorithm rests on recent theoretical advances regarding Sinkhorn iterations for matrix and tensor scaling. An important case which had been out of reach of previous attempts is that of heterogeneous but symmetric battlefield values with asymmetric budget. In this case, the Blotto game is constant-sum so optimal solutions exist, and our algorithm samples from an ϵ-optimal solution in time $O(n^2+\epsilon^{-4})$, independently of budgets and battlefield values. In the case of asymmetric values where optimal solutions need not exist but Nash equilibria do, our algorithm samples from an ϵ-Nash equilibrium with similar complexity but where implicit constants depend on various parameters of the game such as battlefield values.

12:30 - 14:00 – Lunch

14:00 - 15:00 – Goodbye :)