Book of abstracts.
Conference and summer school
"Structural Inference in High-Dimensional Models 2"

Christophe Andrieu (University of Bristol),
e-mail: C.Andrieu@bristol.ac.uk

Nonreversible Markov chain and process Monte Carlo methods

The aim of these lectures is twofold: (a) provide an overview of the construction of known reversible and non-reversible Markov chain Monte Carlo algorithms and their continuous time counterparts and (b) provide an overview of some theoretical results characterising their performance, for example in large dimensional setups.

Olga Klopp (ESSEC Business School),
e-mail: kloppolga@math.cnrs.fr

Matrix Completion: old and new

Low-rank matrices play a fundamental role in statistics and machine Learning. In many situations one can not observe the matrix of interest directly nor fully sample it. Then, one faces the problem of matrix completion from partial and noisy observations of a low-rank matrix or a matrix that can be well approximated by a low-rank matrix. The aim of this mini-course is to provide an overview of modern techniques for exploiting low-rank structure to perform matrix recovery in these settings. We will discuss the algorithms most commonly used in practice, the existing theoretical guarantees for these algorithms and some examples of practical applications.

Content

1. Introduction. Algorithms for Matrix Recovery.
2. Matrix LASSO. Non-commutative Bernstein inequality.
3. Confidence sets for the matrix completion problem.
4. One-bit matrix completion. Robust matrix completion.

References

[1] Afonso S. Bandeira and Ramon van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. Ann. Probab., 44(4):2479–2506, 07 2016.
[2] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. Proceedings of the IEEE, 98(6):925–936, 2010.
[3] Emmanuel J. Candes and Terence Tao. The power of convex relaxation: near-optimal matrix completion. IEEE Trans. Inform. Theory, 56(5):2053–2080, 2010.
[4] A. Carpentier, O. Klopp, M. Löffler, and R. Nickl. Adaptive confidence sets for matrix completion. Bernoulli, 24(4A): 2429 – 2460, 2018.
[5] A. Carpentier, O. Klopp and M. Löffler. Constructing confidence sets for the matrix completion problem. Nonparametric statistics, pp 103 – 118, Springer Proc. Math. Stat., 250, 2018.
[6] Sourav Chatterjee. Matrix estimation by universal singular value thresholding. The Annals of Statistics, 43(1):177–214, 2014.

[7] David Gross. Recovering low-rank matrices from few coefficients in any basis. IEEE Trans. Inform. Theory, 57(3):1548–1566, 2011.
[8] Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. J. Mach. Learn. Res., 11:2057–2078, 2010.
[9] O. Klopp. Rank penalized estimators for high-dimensional matrices. Electron. J. Statist., 5:1161–1183, 2011.
[10] O. Klopp. Noisy low-rank matrix completion with general sampling distribution. Bernoulli, 20(1):282–303, 2014.
[11] Klopp, O., Lounici, K. and Tsybakov, A. Robust Matrix Completion. Probability Theory and Related Fields Vol. 169, Issue 1 - 2, pp 523 – 564 (2017).
[12] Klopp, O., Lafond, J., Moulines, E. and Salmon J. Adaptive Multinomial Matrix Completion. Electronic Journal of Statistics Vol. 9(2), pp. 2950–2975 (2015).
[13] Klopp, O. Matrix completion by singular value thresholding: sharp bounds. Electronic Journal of Statistics Vol. 9(2) , pp. 2348–2369 (2015).
[14] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. Ann. Statist., 39(5):2302–2329, 2011.
[15] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. Journal of Machine Learning Research, 13:1665–1697, 2012.
[16] Tropp, J.A. (2011) User-friendly tail bounds for sums of random matrices. Found. Comput. Math., 11(4).

Mikhail Lifshits (St.Petersburg State University),
e-mail: mikhail@lifshits.org

How complex is a random picture

Complexity of a metric measure space may be expressed in terms of "quantization (i.e. discretization) error" telling how well it may be approximated in the average by a finite subset called dictionary. After exposing some general theory and giving examples, we dwell on a recent particular example studied by F. Aurzada (TU Darmstadt) and the lecturer.

Consider a random set (or "picture") in the unit cube of d-dimensional Euclidean space as a union of balls centered at points of a Poissonian random field and having i.i.d. radii. Let K be the minimal number of balls needed to reproduce the picture.

We study large deviation probabilities for K and prove in some cases that for large n, $\ln P(K > n) \sim -An \ln n$ where the constant A may explicitly depend on dimension, on the distribution of radii, and on the norm under consideration. In many cases the problem of finding the value of A remains open although some upper and lower bounds are available.

This asymptotics has natural corollaries in high dimensional quantization problems.

Stephane Mallat (College de France),
e-mail: stephane.mallat@ens.fr

Mathematics of Deep Convolutional Neural Networks

Deep neural networks obtain impressive results for image, sound and language recognition or to adress complex problems in physics. They are partly responsible of the renewal of artificial intelligence. Yet, we do not understand the underlying mathematics. This course will introduce some important mathematical questions, partial results and open problems.

We will review the architecture of deep convolutional neural networks, the universal approximation theorem for one-hidden layer networks and untractability of approximations in high dimension. Deep neural network have remarkable high-dimensional approximation capabilities over complex classes of functions, which seems to circumvent the curse of dimensionality. Understanding the mathematics of these networks requires to understand high-dimensional regularities. Three elements play an important role: multiscale separations, groups of symmetries and sparsity. We will show how deep neural networks can take advantage of such properties by introducing wavelet transforms, invariant representations and sparse dictionary learning. This will be related to the role of Relu non-linearities and filters in convolutional network.

We shall consider applications to unsupervised learning for the modelization of non-Gaussian ergodic stationary processes such as fluid turbulences and for the modelization of complex non-ergodic processes with autoencoders. We will also study applications to supervised learning for image classification over large data bases such as ImageNet, and in quantum chemistry for the regression of quantum molecular energies.

Denis Belomestny (Duisburg-Essen University, HSE University),
e-mail: denis.belomestny@uni-due.de

Variance-reduced Q-learning via martingale representations

In this talk we propose a novel regression-based approach for Q-learning. This approach reduces the complexity of the Monte Carlo Q-learning algorithm and has an especially simple form for Markov processes with known transition densities. We analyze the complexity of the proposed approach in the case of both fixed and increasing numbers of periods. The method is illustrated by several numerical examples.

Marco Cuturi (CREST, ENSAE),
e-mail: cuturi@google.com

Differentiable Ranks using Optimal Transport: The Sinkhorn CDF and Quantile Operator

We propose a framework to sort values that is algorithmically differentiable. We leverage the fact that sorting can be seen as a particular instance of the optimal transport (OT) problem on R, from input values to a predefined array of sorted values (e.g. $1, 2, \ldots, n$ if the input array has n elements). Building upon this link, we propose generalized ranks, CDFs and quantile operators by varying the size and weights of the target pre-sorted array. We recover differentiable algorithms by adding to the OT problem an entropic regularization, and approximate it using a few Sinkhorn iterations. We call these operators S-ranks, S-CDFs and S-quantiles, and use them in various learning settings: we benchmark them against the recently proposed neuralsort [Grover et al. 2019], propose applications to quantile regression and introduce differentiable formulations of the top-k accuracy that deliver state-of-the art performance.

Arnak Dalalyan (CREST, ENSAE),
e-mail: arnak.dalalyan@ensae.fr

On sampling from a log-concave density using kinetic Langevin diffusions

Langevin diffusion processes and their discretizations are often used for sampling from a target density. The most convenient framework for assessing the quality of such a sampling scheme corresponds to smooth and strongly log-concave densities. The present talk focuses on this framework and describes the behavior of Monte Carlo algorithms based on discretizations of the kinetic Langevin diffusion. We first present the geometric mixing property of the kinetic

Langevin diffusion with a mixing rate that is, in the overdamped regime, optimal in terms of its dependence on the condition number. We then use this result for obtaining improved guarantees of sampling using the kinetic Langevin Monte Carlo method, when the quality of sampling is measured by the Wasserstein distance. We also consider the situation where the Hessian of the log-density of the target distribution is Lipschitz-continuous. In this case, we introduce a new discretization of the kinetic Langevin diffusion and prove that this leads to a substantial improvement of the upper bound on the sampling error measured in Wasserstein distance. (joint work with L. Riou-Durand)

Alexander Gasnikov (HSE University, MIPT, IITP RAS),
e-mail: gasnikov@yandex.ru

Acceleration of Sinkhorn algorithm for optimal transportation problem and Iterative Bregman Projection algorithm for Wasserstein barycenter problem

In our recent work arXiv:1906.03622 Accelerated alternating minimization methods was proposed. This method is primal-dual and converges like fast gradient method for smooth convex problems in general case. But typically, in practice it converges much faster due to the possibility of exact auxiliary minimizations. In this talk we will demonstrate how to apply this method for two problems: dual problem for Entropy regularized Optimal Transportation problem arXiv:1802.04367 and dual problem to Entropy regularized Wasserstein barycenter problem arXiv:1901.08686.

Friedrich Götze (Bielefeld University),
e-mail: goetze@math.uni-bielefeld.de

Concentration of Measure and Entropic Convergence

We study 'higher' order concentration of measure bounds for functionals on the sphere, Euclidean and discrete spaces. These general results will be applied to the distribution of weighted sums with dependencies and to distribution questions for spin systems and unbounded functionals of polynomial type. Furthermore we discuss the entropic convergence to the Poisson law measured in relative entropy based divergences. This includes the full hierarchy of Renyi/Tsallis type divergences.

Ildar Ibragimov (PDMI RAS),
e-mail: ibr32@pdmi.ras.ru

On the estimation of intensity density functions of Poisson processes

The aim of this talk is to present some results about non-parametric estimation of the intensity density function of a Poisson process. We consider the following problem. We are observing a Poisson process $X_\varepsilon(t)$ (a Poisson random measure $X_\varepsilon(A)$) on an interval $[a, b]$ (on a region G). The non-homogeneous process $X_\varepsilon(t)$ has the intensity measure $\epsilon^{-1}\Lambda$ where $\varepsilon > 0$ is a known small parameter and Λ is an unknown measure. It is supposed that the measure Λ is absolutely continuous with respect to the Lebesgue measure and has the density (the intensity density) function $\lambda(t)$ and that the unknown density λ belongs to a known class F of functions. The basic problem is to estimate λ on the base of the observations X_ε. Denote $||.||_p$ the norm in $L_p(a, b)$. Set

$$\Delta_p(\varepsilon, F) = \Delta_p(\varepsilon) = \inf \sup \mathbf{E}_\lambda ||\hat{\lambda} - \lambda||_p$$

where sup is taken over all $\lambda \in F$ and inf is taken over all possible estimates $\hat{\lambda}$ of λ. In the talk we study the asymptotic behavior of estimates when $\varepsilon \to 0$ and in particular the rate of convergence of Δ to zero.

The rate depends on F. We study the question of dependence of the rate on characteristics of "massivity" of F, on the ε-entropy of F, its Kolmogorov diameters or some other characteristics.

Bing-Yi Jing (HKUST),
e-mail: majing@ust.hk

Recommender system incorporating social network information

We propose the so-called NetRec method in recommender system by incorporating the network information into collaborative filtering (CF). This results in a sharper error bound than previous literature under reasonable assumptions. It is also shown that the combination of the network-related penalty and the nuclear norm penalty gives better estimates than those achieved by any of them alone. The method has been shown to work well in simulations and some real data sets on Yelp.

Alexander Kolesnikov (HSE University),
e-mail: sascha77@mail.ru

Convexity and transportation: inequalities, barycenters, analysis on Wiener space

We present some new and classical results connecting several areas of research: gaussian analysis (in finite and infinite dimensions), inequalities for convex bodies, transportation inequalities. In particular, we discuss applications of optimal transportation to Minkowski-type problems and some new forms of the Blaschke-Santalo inequality. In addition, we present some results and open problems on geodesic barycenters in Wiener space.

Alexey Kroshnin (HSE University, IITP),
e-mail: akroshnin@hse.ru

Shape-based domain adaptation via optimal transportation

Domain adaptation problem aims at learning a well performing model, trained on a source data S (images, vectors, e.t.c), applied then to different (but related) target sample T. Aside from being attractive due to obvious practical utility, the setting is challenging from theoretical point of view. In this work we introduce a novel approach to supervised domain adaptation consisting in a class-dependent fitting based on ideas from optimal transportation (OT) theory which considers S and T as two mixtures of distributions. A parametrized OT distance is used as a fidelity measure between S and T, providing a toolbox for modelling of possibly independent perturbations of mixture components.

Axel Munk (Universität Göttingen),
e-mail: amunk1@gwdg.de

Empirical Optimal Transport: Inference, Algorithms, Applications

We discuss recent developments in statistical data analysis based on empirical optimal transport (EOT). Fundamental are limit laws for EOT plans and distances on finite and discrete spaces. These are characterized by dual optimal transport problems over a gaussian process. Our proofs are based on a combination of sensitivity analysis from convex optimization and discrete empirical process theory. We examine an upper bound for such limiting distributions based on a spanning tree approximation which can be computed explicitly. This can be used for

statistical inference, fast simulation, and for fast randomized computation of optimal transport in large scale data applications at pre-specified computational cost as it provides error bounds to balance computational and statistical error. Our methodology is illustrated in computer experiments and on biological data from super-resolution cell microscopy. Finally, this is contrasted and compared with recent results on regularized empirical optimal transport. This is based on joint work with M. Klatt, M. Sommerfeld, C. Tameling and Y. Zemel.

Alexey Naumov (HSE University),
e-mail: anaumov@hse.ru

Concentration inequalities for functionals of Markov chains

Empirical Variance Minimization technique to reduce variance of MCMC is naturally related to concentration and moment inequalities for U-statistics of Markov chains. In my short talk I will state the problem and give partial solution based on the transportation-information cost inequalities, W_2 – contraction and recent result by Adamzack (2015).

Salem Said (CNRS – Université de Bordeaux),
e-mail: salem.said@u-bordeaux.fr

Riemannian barycentres of Gibbs distributions : new results on concentration and convexity

Let P be a probability distribution on a Riemannian manifold M. A Riemannian barycentre of P is any point in M, which achieves the global minimum of the so-called variance function

$$\mathcal{E}(x) \,=\, \frac{1}{2}\,\int_M d^2(x,y)\,P(dy) \quad \text{ for } \; x \in M$$

where $d(x,y)$ denotes Riemannian distance. In the special case where M is a Euclidean space, P has one and only one barycentre, which is identical to the expectation of P. Making this observation, in 1948, Fréchet proposed the concept of barycentre as a generalisation of the concept of expectation, to probability distributions on Riemannian manifolds (or even on general metric spaces).

Fast-forward to our century, a bit more than fifteen years ago, this old idea was resurrected, and applied to a huge number of recent problems in data science. Today, the so-called Riemannian barycentre (or Fréchet mean) is the workhorse of data analysis, when it comes to data in Riemannian manifolds.

So much success should seem dubious, to the expert in Riemannian geometry : even in the most elementary situations, the variance function $\mathcal{E}(x)$ is non-differentiable, non-convex, and has multiple local or even global minima. Thus, in order to exploit the Riemannian barycentre, as a tool for data analysis, it is very important to understand its differentiability, convexity, uniqueness, and other properties. An important contribution, in this respect, was made by Afsari, in 2010, who proved that, as long as P is supported inside a convex geodesic ball in M, the Riemannian barycentre of P is unique and belongs to this geodesic ball.

This result is optimal, since many elementary examples show the barycentre of P can fail to be unique, if P is not supported inside a convex geodesic ball. However, it does not tell us what happens in certain important cases, where P is not supported inside, but merely concentrated on a convex geodesic ball. In particular, it does not say anything about the case where $P = P_T$ is a Gibbs distribution

$$P_T(dy) \,\propto\, \exp\left[-\frac{U(y)}{T}\right]\,v(dy) \quad (v \text{ denotes Riemannian volume})$$

My presentation will uncover some new results, which decide the properties of concentration, differentiability, convexity, and uniqueness of the Riemannian barycentre of a Gibbs distribution P_T, assuming the function U has a unique global minimum $y^* \in M$. In particular, these new results imply the following theorem: *if M is a simply-connected compact Riemannian symmetric space, with convexity radius r_{cx}, then for all $\delta < \frac{1}{2} r_{cx}$ there exists T_δ such that $T < T_\delta$ implies the Riemannian barycentre of P_T is unique and belongs to the geodesic ball $B(y^*, \delta)$. Moreover, if U is invariant by geodesic symmetry about y^*, then this Riemannian barycentre is identical to the global minimum y^*.* Remarkably, this does not require the function U to be smooth.

This theoretical result comes with an applied reward. The problem of finding the global minimum of a non-smooth function U becomes equivalent to the problem of computing the Riemannian barycentre of the Gibbs distribution P_T (provided T is chosen correctly). This gives rise to an original algorithm for black-box optimisation, based on the idea of recursive computation of the barycentre of P_T, using samples generated from a Riemannian MCMC approximation.

The performance of this algorithm, which seems quite promising, will be illustrated with two computer experiments. A theoretical understanding of this performance raises several interesting questions, which remain open in the literature.

Eugene Stepanov (PDMI RAS),
e-mail: stepanov.eugene@gmail.com

A tour of location problems: from optimal to random

The classical location (k-median) problem is that of placing k facilities modeled by points in an optimal way in the given region. We will discuss several ways of placing points, namely globally optimally, optimally one-by-one, and randomly, evaluating each time the asymptotic behavior of the average distance functional (or, equivalently, the Wasserstein distance to the reference measure).

Alexandra Suvorikova (University of Potsdam),
e-mail: a.suvorikova@gmail.com

On some geometrical intuition on multiplier bootstrap in Bures-Wasserstein space

In this talk we first briefly introduce the concept of Bures-Wasserstein (BW) barycenters of hermitian finite-dimensional matrices, and explain how they can be used for investigation of geometry of DNA molecules modelled as a union of rigid bodies. The main objective of the talk is to present an extension of multiplier bootstrapping technique to BW space, which is then used for construction of non-asymptotic confidence sets for BW barycenters, and explain the underlying geometrical intuition behind the procedure.

Lukas Szpruch (Alan Turing Institute),
e-mail: L.Szpruch@ed.ac.uk

Mean-Field Langevin Dynamics and Energy Landscape of Neural Networks

We present a probabilistic analysis of the long-time behaviour of the nonlocal, diffusive equations with a gradient flow structure in 2-Wasserstein metri. Our work is motivated by a desire to provide a theoretical underpinning for the convergence of stochastic gradient type algorithms widely used for non-convex learning tasks such as training of deep neural networks. The key insight is that the certain class of the finite dimensional non-convex problems becomes

convex when lifted to infinite dimensional space of measures. We leverage this observation and show that the corresponding energy functional defined on the space of probability measures has a unique minimiser which can be characterised by a first order condition using the notion of linear functional derivative. Next, we show that the flow of marginal laws induced by the Mean-Field Langevin Dynamics (MFLD) converges to the stationary distribution which is exactly the minimiser of the energy functional. We show that this convergence is exponential under conditions that are satisfied for highly regularised learning tasks. At the heart of our analysis is a pathwise perspective on Otto calculus used in gradient flow literature which is of independent interest. Our proof of convergence to stationary probability measure is novel and it relies on a generalisation of LaSalle's invariance principle. Importantly we do not assume that interaction potential of MFLD is of convolution type nor that has any particular symmetric structure. This is critical for applications. Finally, we show that the error between finite dimensional optimisation problem and its infinite dimensional limit is of order one over the number of parameters.

Alexander Tsybakov (CREST, ENSAE),
e-mail: Alexandre.Tsybakov@ensae.fr

Estimation of functionals in sparse vector model

Assume that we have the observations $y_i = \theta_i + \varepsilon\xi_i,\ i = 1, \dots, d$, where $\theta = (\theta_1, \dots, \theta_d) \in \mathbf{R}^d$ is a vector of unknown parameters, $\varepsilon > 0$, and ξ_i are independent identically distributed (i.i.d.) random variables. Assume also that θ belongs to the class $B_0(s)$ of all s-sparse vectors, that is, vectors in $\mathbf{R}^d$ with not more than s non-zero components, $s \in \{1, \dots, d\}$. We first consider the problem of estimation of $\|\theta\|_\gamma = \left(\sum_{i=1}^d |\theta_i|^\gamma\right)^{1/\gamma}$, $\gamma > 0$, based on observations $y = (y_1, \dots, y_d)$. We prove that, if $\varepsilon > 0$ is known and ξ_i are i.i.d. standard Gaussian variables, the minimax risk for estimation of $\|\theta\|_\gamma$ under the squared loss on the class $B_0(s)$ satisfies

$$\inf_{\hat{T}} \sup_{\theta \in B_0(s)} \mathbf{E}_\theta[(\hat{T} - \|\theta\|_\gamma)^2/\varepsilon^2] \asymp \begin{cases} s^{2/\gamma}\log(1 + d/s^2), & \text{if } s \le \sqrt{d}, \\ \dfrac{s^{2/\gamma}}{\log(1 + s^2/d)}, & \text{if } s > \sqrt{d} \text{ and } \gamma \notin E, \\ d^{1/\gamma}, & \text{if } s > \sqrt{d} \text{ and } \gamma \in E, \end{cases}$$

where E is the set of all even integers, and $\mathbf{E}_\theta$ denotes the expectation with respect to the distribution of y, and $\inf_{\hat{T}}$ is the infimum over all estimators. We also construct estimators achieving this minimax rate.

Next, for the same sparse vector model, when the noise is not necessarily Gaussian and ε is not necessarily known, we consider adaptive estimation of θ, of the norm $\|\theta\|_2$ and of the noise variance ε^2. We construct adaptive estimators and establish the optimal rates when adaptation is considered with respect to the triplet "noise level - noise distribution - sparsity". We consider classes of noise distributions with polynomially and exponentially decreasing tails as well as the case of Gaussian noise. The obtained rates turn out to be different from the minimax non-adaptive rates when the triplet is known. A crucial issue is the ignorance of the noise variance. Moreover, knowing or not knowing the noise distribution can also influence the rate. For example, the rates of estimation of the noise variance can differ depending on whether the noise is Gaussian or sub-Gaussian without a precise knowledge of the distribution. Estimation of noise variance in our setting can be viewed as an adaptive variant of robust estimation of

scale in the contamination model, where instead of fixing the nominal distribution in advance, we assume that it belongs to some class of distributions.

Dmitry Zaporozhets (PDMI RAS),
e-mail: zap1979@gmail.com

Generalized Busemann inequality

Based on a joint work with Alexander Litvak.

We will discuss a result that generalizes both the Busemann intersection inequality and the Busemann random simplex inequality.

Nikita Zhivotovskiy (Google, HSE University),
e-mail: nzhivotovskiy@hse.ru

Noise sensitivity of the top eigenvector of random matrices

In this talk we discuss the noise sensitivity of the top eigenvector of a Wigner matrix in the following sense. Let v be the top eigenvector of an $N \times N$ Wigner matrix. Suppose that k randomly chosen entries of the matrix are resampled, resulting in another realization of the Wigner matrix with top eigenvector u. We prove that when k is much greater than $N^{5/3}$, then u is almost orthogonal to v, and this threshold is sharp. This result is closely related to a particular case of superconcentration phenomenon.

Based on a joint work Charles Bordenave and Gabor Lugosi