Mathematisches Forschungsinstitut Oberwolfach

Report No. 25/2022

DOI: 10.4171/OWR/2022/25

Re-thinking High-dimensional Mathematical Statistics

Organized by
Florentina Bunea, Ithaca
Robert Nowak, Madison
Alexandre Tsybakov, Palaiseau

15 May – 21 May 2022

Abstract. The workshop highlighted recent theoretical advances on inference in high-dimensional statistical models based on the interplay of techniques from mathematical statistics, machine learning, theoretical computer science and related areas. The workshop brought together about 50 researchers in order to present new results, exchange ideas and explore open problems.

Mathematics Subject Classification (2020): 62-xx.

Introduction by the Organizers

The mathematical treatment of high dimensional statistical problems has been at the core of recent research in Statistics, Machine Learning and Artificial Intelligence. The last several decades have seen both positive and negative results that made it evident that statistical inference in very high dimensions, possibly larger than the size of the observed samples, is reliable and practical only for data that implicitly or explicitly are generated from probability models that exhibit low dimensional structure.

Understanding the nature of this structure has been and continues to be an area of active investigation. The last two decades showed the crucial role played by sparsity in generative models, and lead to the growth of a sub-area devoted to the understanding of restricted (regularized) statistical estimators. The last few years have seen a more intense focus on other structures, chiefly low dimensional latent generative structures, broadly defined. The interplay between the ways in which various simpler structures can explain phenomena in high dimensions lead to

new impactful mathematical developments that require re-visiting and re-thinking existing paradigms in mathematical statistics.

The workshop talks and discussions tackled several new challenges and open problems. The statistical aspects of neural networks and deep learning are not well understood, and this has led to the study of over-parameterized models in high dimensional statistical problems. This includes principal component analysis, classification and regression problems under the (sometimes implicit) assumption on the existence of a lower dimensional latent generative model. While latent models have been studied and applied for years, their behavior and effect on analyses is only well understood in low dimensions. Connecting them with modern Machine Learning and Artificial Intelligence challenges, as well as studying them in high dimensions requires a re-thinking of existing proof techniques, while also opening new avenues for research. Other important trends include non-asymptotic approaches to robustness in high dimensions and methods of online learning, in particular, via developing novel bandit algorithms as well as sampling techniques. Contributions to all these directions were provided at the workshop. The talks given at the workshop can be, at a high level, subdivided into the following groups.

- Structured high-dimensional inference and learning: The talks by Boaz Nadler, Jaouad Mourtada, Sara van de Geer, Nicolas Verzelen, Mathias Drton, Marten Wegkamp, Olga Klopp, Corinne Emmenegger, Martin Wahl. A particular focus on tensor models was provided in the talks by Ming Yuan and Cun-Hui Zhang.
- Deep learning: The talks by Peter Bartlett, Andrea Montanari, Helmut Bölcskei.
- Bandit algorithms: The talks by Solenne Gaucher, Karim Lounici, Alexandra Carpentier.
- Robustness: The talks by Arnak Dalalyan and Mohamed Ndaoud.
- Sampling techniques and stochastic optimization: The talks by Richard Nickl, Arya Akhavan, Vladimir Spokoiny, Kevin Jamieson, Evgenii Chzhen.
- Other topics: The talks by Philippe Rigollet (statistical optimal transport), Angelika Rohde (testing in diffusion models), Johannes Schmidt-Hieber (foundations of inference), László Györfi (classification in metric spaces).

Workshop: Re-thinking High-dimensional Mathematical Statistics

Table of Contents

Abstracts

On sampling high-dimensional non-log concave posterior measures

Richard Nickl

The problem of efficiently generating random samples from high-dimensional and non-log-concave posterior measures arising from nonlinear regression problems is considered. Extending investigations from [2], local and global stability properties of the model are identified under which such posterior distributions can be approximated in Wasserstein distance by suitable log-concave measures. This allows the use of fast gradient based sampling algorithms, for which convergence guarantees are established that scale polynomially in all relevant quantities (assuming 'warm' initialisation). Applications to a variety of PDE models are discussed, such as non-Abelian X-ray transforms [1] or elliptic diffusion or Schrödinger equations.

References

[1] F. Monard, R. Nickl, G.P. Paternain, Consistent inversion of non-Abelian X-ray transforms. *Comm. Pure Appl. Math.* (2021).

[2] R. Nickl, S. Wang, On polynomial-time computation of high-dimensional posterior measures by Langevin-type algorithms, *J.Eur.Math.Soc.* (2022), to appear.

[3] J. Bohr, R. Nickl, On log-concave approximations of high-dimensional posterior measures and stability properties in non-linear inverse problems, *arXiv* (2021).

Completing large matrices with only few observed entries: A one-line algorithm with provable guarantees

Boaz Nadler

(joint work with Pini Zilber)

Suppose you observe very few entries from a large matrix. Can we predict the missing entries, say assuming the matrix is (approximately) low rank ? In [1] we describe a very simple and computationally efficient method to solve this non-convex matrix completion problem. Our approach, denoted GNMR, is factorization based and relies on a Gauss-Newton linearization of the quadratic objective. In details, given a current guess of the two factor matrices U_0 and V_0, at each iteration GNMR searches for an update $U_1 = U_0 + \Delta U, V_1 = V_0 + \Delta V$ by minimizing the following objective

$$\min_{\Delta U, \Delta V} \|\mathcal{P}_\Omega(U_0 V_0^\top + U_0 \Delta V^\top + \Delta U V_0^T) - \mathcal{P}_\Omega(X^*)\|_F.$$

Here X^* is the true underlying matrix, Ω is the subset of observed entries and $\mathcal{P}_\Omega$ is the projection of a matrix onto the set of indices Ω. In the above objective, the non-convex quadratic term $\Delta U \Delta V^\top$ has been neglected. Hence, at each iteration GNMR solves a simple least squares problem. On the empirical front, we show that GNMR is able to recover matrices from very few entries and/or with ill conditioned matrices, where many other popular methods fail. Furthermore, due to

its simplicity, it is easy to extend our method to incorporate additional knowledge on the underlying matrix, for example to solve the inductive matrix completion problem [2]. On the theoretical front, we prove that GNMR enjoys some of the strongest available theoretical recovery guarantees. Specifically, near the global optimum its convergence rate is quadratic. To derive these theoretical results we develop a sharp RIP-like guarantee for matrix completion. This theorem, in particular, implies a uniform RIP for the difference between two incoherent matrices, provided they are sufficiently close to each other. This settles an open question posed in [3]. Finally, for inductive matrix completion, we prove that under suitable conditions the recovery problem itself has a benign optimization landscape with no bad local minima.

Our work also raises several open questions and challenges. For example, comparing the empirical results of our method with the theoretical recovery guarantees highlights that there exist a significant gap. Specifically, nearly all available guarantees, including ours, scale at least quadratically with the condition number of the underlying matrix. In contrast, empirically our method shows little sensitivity to the condition number and successfully recovers highly ill conditioned matrices from few observations. Deriving guarantees for factorization based methods that close this gap is an interesting question for future research.

References

[1] P. Zilber and B. Nadler, *GNMR: A Provable One-Line Algorithm for Low Rank Matrix Recovery*, SIAM Journal on the Mathematics of Data (2022), to appear.
[2] P. Zilber and B. Nadler, *Inductive Matrix Completion: No Bad Local Minima and a Fast Algorithm*, International Conference on Machine Learning (ICML), 2022.
[3] M.A. Davenport and J. Romberg, *An overview of low-rank matrix recovery from incomplete observations*, IEEE Journal on Selected Topics in Signal Processing **10** (2016), 608–622.

Coding convex bodies under Gaussian noise, and the Wills functional

Jaouad Mourtada

In sequential probability assignment, one aims to assign a large probability to a sequence of observations (unknown a priori), close to that of the best a posteriori distribution within a prescribed model. This prediction problem is intimately connected to that of lossless coding in information theory.

In this work, we study the case of a sequence of real-valued observations, modeled by a subset of the Gaussian sequence model with mean constrained to a general convex body. This can be thought of as an information-theoretic analogue of fixed-design regression. We show that the minimax-optimal error is exactly given by a certain functional of the constraint set from convex geometry called the Wills functional. As a consequence, we express the optimal error in terms of basic geometric quantities associated to the convex body, namely its intrinsic volumes. After comparing the optimal error to the Gaussian width of the constraint set, and

to fixed points of local Gaussian widths, we state a fundamental concavity property of the error, and deduce some strong monotonicity properties with respect to noise and sample size.

References

[1] J. Mourtada, *Coding convex bodies under Gaussian noise, and the Wills functional*, In preparation, 2022.

Sample complexity of entropic optimal transport

Philippe Rigollet
(joint work with Austin J. Stromme)

Fueled by recent computational advances, optimal transport (OT) techniques have become preponderant in a variety of statistical applications. Given two measures μ and ν on $\mathbb{R}^d$ and a cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$, the OT problem of interest here is the infinite dimensional linear optimization problem given by

$$\inf_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|^2 \mathrm{d}\pi(x, y)\,, \tag{1}$$

where the infimum is taken over the set $\Pi(\mu, \nu)$ of couplings between μ and ν and $\|\cdot\|$ denotes the Euclidean norm. Recall that $\pi \in \Pi(\mu, \nu)$ is a valid coupling between μ and ν if π is a probability measure on $\mathbb{R}^d \times \mathbb{R}^d$, such that for any measurable $A \subset \mathbb{R}^d$, it holds that $\pi(A \times \mathbb{R}^d) = \mu(A)$ and $\pi(\mathbb{R}^d \times A) = \nu(A)$. We assume further that μ and ν have bounded support. Under these conditions, (1) admits a unique minimizer π_0, called the *OT coupling*. Furthermore, Brenier's theorem states that under mild regularity conditions on μ, the OT coupling π_0 is supported on the graph of a deterministic map T called the *Brenier map*. In other words, $(X, Y) \sim \pi_0$ if and only if $X \sim \mu$ and $Y = T(X) \sim \nu$.

The OT coupling and the Brenier map have a dynamical interpretation in terms of energy minimization that has fueled a conceptual shift from the traditional statistical toolbox in many areas including statistics, economics, computer graphics, computational biology, and machine learning. Indeed, a central application of optimal transport is transfer learning, where the goal is to transfer information from one dataset to another using the Brenier map which, in turn needs to be estimated from data.

Unfortunately, a long line of work has provided strong evidence that the OT coupling suffers from a statistical curse of dimensionality. Indeed, without further assumption, the minimax rate for estimating the OT cost is at least $n^{-2/d}$ and a similar rate is conjectured to hold for the problem of estimating the Brenier map T. While recent theoretical effort has been devoted to showing that this inefficiency can be alleviated by making structural assumptions—chiefly smoothness—on the transport map, finding computationally efficient and smoothness-adaptive estimators is a challenging and ongoing research topic.

In this talk, we presented an alternative to the OT coupling that we call the *Schrödinger coupling*. It arises as the solution to the entropically regularized OT problem given by, for $\eta > 0$,

$$S(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \Big\{ \int \|x-y\|^2 \mathrm{d}\pi(x,y) + \frac{1}{\eta}\mathsf{KL}(\pi\|\mu\otimes\nu)\Big\}, \tag{2}$$

where KL denotes the Kullback-Leibler divergence. Under our conditions, the Schrödinger coupling exists and is unique and denoted $\pi_\star$. This regularized problem dates back to early work of Schrödinger and recently has largely eclipsed the OT coupling in applications because it offers significant computational advantages. Like the OT coupling, the Schrödinger coupling also arises from a minimum energy paradigm in the context of stochastic control and is also justified in the aforementioned applications even for a fixed η, say $\eta = 1$. While it does not give the Brenier map, it produces a map, called the *entropic regression function*, and defined by $x \mapsto \mathbb{E}[Y|X = x]$ where $(X, Y) \sim \pi_\star$. This map is just as useful as the Brenier map in applications. However, unlike the latter, a key finding of this work is that estimation of the entropic regression function does not suffer from the curse of dimensionality.

More specifically, we focused on the estimation of the Schrödinger coupling and quantities that are derived from it. To that end, assume that we observe $X_1, \ldots, X_n \sim \mu$ i.i.d., and $Y_1, \ldots, Y_n \sim \nu$ i.i.d. The corresponding empirical measures are denoted μ_n and ν_n respectively and we consider the optimization problem associated to computing $S(\mu_n, \nu_n)$. In particular, this *plug-in* version of the problem outputs a coupling of μ_n and ν_n from which estimators of various quantities of interest, including the entropic regression function, can be computed. Crucially these estimators escape the curse of dimensionality and, instead, converge at a fast, parametric rate. Our proofs use an elementary approach to bypass the control of suprema of empirical processes, and may be of independent interest.

Re-thinking logistic regression

Sara van de Geer

Theoretical results for the logistic regression model typically assume that the observed binary label has probabilities staying away from zero. We study a situation where this assumption is violated, which is the case where the the noise level is low. There may even be no noise at all, which is often assumed in the literature on 1-bit compressed sensing.

Consider a feature vector $\mathbf{x} \in \mathbb{R}^s$ and a vector of regression coefficients $\beta^* \in \mathbf{S}_2^{s-1} := \{b \in \mathbb{R}^s : \|b\|_2 = 1\}$, with $s \log n \ll n$ and let $\mathbf{y} \in \{\pm 1\}$ be the sign of $\mathbf{x}\beta^* + \sigma\zeta$ where the noise $\zeta \sim \mathbf{N}(0, 1)$ is independent of $\mathbf{x}$ and $0 < \sigma \leq 1$. Then $1/\sigma$ is the signal-to-noise level, since only the ratio of signal strength and noise level is identified.

We observe n i.i.d. copies $\{(X_i, Y_i)\}_{i=1}^n$ of $(\mathbf{x}, \mathbf{y})$. With feature vector $x \in \mathbb{R}^s$ and label $y \in \{\pm 1\}$ the logistic loss function is $l_c(x, y) := \log(1 + \mathrm{e}^{-yxc})$. We examine the estimator $\hat{\gamma} := \arg\min_{c\in\mathbb{R}^s} \sum_{i=1}^n l_c(X_i, Y_i)$.

Let $\hat{\beta} := \hat{\gamma}/\|\hat{\gamma}\|_2$. We show that for the case of Gaussian design, the rate of convergence for $\|\hat{\beta} - \beta^*\|_2$ is of order $\sqrt{\sigma s \log n/n}$ when σ is of larger order $s \log n/n$. For the case where σ is of order $s \log n/n$ we obtain a rate of order $s \log n/n$ for $\|\hat{\beta} - \beta^*\|_2$ where $\hat{\beta} = \hat{\gamma}/\|\hat{\gamma}\|_2$ is now the logistic regression estimator with the constraint that $\|\hat{\gamma}\|_2$ is bounded by a large constant. Empirical risk minimization with $\{0,1\}$ loss $1\{\text{sign}(xc) \neq y\}$ has the rate $(\sigma^2 s \log n/n)^{1/3} \vee s \log n/n$. Thus, with Gaussian design, logistic regression appears to have a faster rate.

Estimator	loss function	$\|\hat{\beta} - \beta^*\|_2$
ERM	$1\{\text{sign}(xc) \neq y\}$	$\left(\frac{\sigma^2 s \log n}{n}\right)^{\frac{1}{3}} \vee \frac{s \log n}{n}$
logistic regression	$\log(1 + \mathrm{e}^{-xc})$	$\left(\frac{\sigma s \log n}{n}\right)^{\frac{1}{2}} \vee \frac{s \log n}{n}$
linear	$-yxc + \|c\|_2^2$	$\left(\frac{s}{n}\right)^{\frac{1}{2}}$
hybrid	$-yxb + \|xb\|$	$\left(\frac{\sigma s \log n}{n}\right)^{\frac{1}{2}} \vee \frac{s \log n}{n}$

TABLE 1. Rates of convergence for various estimators $\hat{\beta}$ of β^*. The noise is assumed to be $\mathbf{N}(0, \sigma^2)$-distributed and independent of the features. The result for ERM follows from [1]. The hybrid estimator faces the norm constraint $\|b\|_2 = 1$.

After the reparamatrization $c \mapsto (\tau, b)$, where $\tau = \|c\|_2 \in \mathbb{R}_+$ is the "signal strength" and $b := c/\|c\|_2 \in \mathbf{S}_2^{s-1}$ is the "direction", one may rewrite the logistic loss function as

$$l_{\tau,b}(x,y) = \log(1 + \mathrm{e}^{-\tau|xb|}) + \tau \underbrace{|xb| 1\{\text{sign}(xb) \neq y\}}_{\text{hybrid loss}}$$

so that the logistic risk is

$$\sum_{i=1}^n l_{\tau,b}(X_i, Y_i) = \sum_{i=1}^n \log(1 + \mathrm{e}^{-\tau|X_i b|}) + \sum_{i=1}^n \tau |X_i b| 1\{\text{sign}(X_i b) \neq Y_i\}.$$

The first term induces a normalization, whereas the second term promotes a small number of misclassifications. Note that in the noiseless case, one can interpolate the signs by taking $b = \beta^*$: $\text{sign}(X_i \beta^*) = Y_i$ for all $i \in \{1, \ldots, n\}$. Then the second term is zero and then the first term is minimized by taking $\tau \to \infty$. In other words, then the logistic regression estimator does not exist. For very small noise, one can interpolate with a probability staying away from zero. Therefore, when there is no noise or very small noise, one needs to restrict the estimator or add a penalty term.

We show that for $\sigma > 0$, the score for estimating $\tau^* := \|\gamma^*\|_2$ is orthogonal to the score for estimating β^*. Moreover, we explain that the rate of convergence for

estimating τ^* is much slower than the rate for estimating β^*. This leads to using a weighted Euclidean distance between (τ^*, β^*) and $(\hat{\tau}, \hat{\beta})$ where $\hat{\tau} = \|\hat{\gamma}\|_2$. We show a type of sandwich formula result when σ is of larger order $s \log n/n$: lower bounds for the excess risk and upper bounds for the empirical process, both in terms of the weighted Euclidean distance. For σ or order $s \log n/n$ we bound $1/\hat{\tau}$ first and then derive the rate for $\hat{\beta}$.

References

[1] P. Massart, and É. Nédélec, *Risk bounds for statistical learning*, The Annals of Statistics, **34** (2006) 2326–2366.

Optimal ranking in crowd-sourcing problems

Nicolas Verzelen

(joint work with Alexandra Carpentier, Emmanuel Pilliat)

We consider a crowd-sourcing problem where we have n experts and d tasks. The average ability of each expert for each task is stored in an unknown matrix M, from which we have incomplete and noisy observations. We make no semi-parametric assumptions, but assume that both experts and tasks can be perfectly ordered: so that if an expert A is better than an expert B, the ability of A is higher than that of B for all tasks - and that the same holds for the tasks. This implies that if the matrix M, up to permutations of its rows and columns, is bi-isotonic. We focus on the problem of recovering the optimal ranking of the experts when the ordering of the tasks is known to the statistician. In other words, we aim at estimating a permutation π^* of the rows of M such that the corresponding permuted matrix is bi-isotonic.

This problem has attracted a lot of attention in the last years [2, 3]. Unfortunately, there remains a large gap between the minimax estimation rate achieved by exponential-time methods and the much more worse performances achieves by known polynomial time methods. Recently, Liu and Moitra [1] have introduced a minimax polynomial time procedure in the specific square case ($n = d$), where one has polylogarithmic noisy observations of the matrix M. The purpose of this presentation is to go beyond this specific situation.

Let us describe more formally the setting. Let $M \in [0,1]^{n \times d}$ and π^* be a permutation of $[n]$ such that the corresponding permuted matrix $M_{\pi^{*-1}}$ is bi-isotonic. For an estimator $\widehat{\pi}$ of π^*, we quantify its error by the loss

$$l(\pi^*, \widehat{\pi}) = \|M_{\pi^{*-1}} - M_{\widehat{\pi}^{-1}}\|_F^2 \ .$$

We consider a partial observation scheme defined as follows. Given $\lambda > 0$, we have $\mathcal{P}(\lambda nd)$ observations of the form (y_s, x_s) where the position $x_s \in [n] \times [d]$ is sampled uniformly and $y_s = M_{x_s} + E_{x_s}$ is a noisy observation of M_{x_s}. Here, E_{x_s} is distributed as a standard Gaussian variable.

Our main contribution is the construction of a quadratic-time estimator $\widehat{\pi}$ achieving the risk bound

$$\mathbb{E}[l(\pi^*, \widehat{\pi})] \leq c \log^{c'}(nd) \left[\frac{nd^{1/6}}{\lambda^{5/6}} \wedge \frac{n^{3/4} d^{1/4}}{\lambda^{3/4}} + \frac{n}{\lambda} \right] ,$$

where c and c' are numerical constants. Up to polylogarithmic multiplicative terms, this risk bound is minimax optimal for all n, d, and all $\lambda \in [1/d, 1]$, which encompasses all non-trivial regimes of partial observations.

Among others, the construction of $\widehat{\pi}$ combines hierarchical clustering ideas with spectral clustering and change-point detection methods.

References

[1] A. Liu and A. Moitra, *Better algorithms for estimating non-parametric models in crowdsourcing and rank aggregation*, in Conference on Learning Theory (2020), 2780–282.

[2] M. Cheng, A. Pananjady, and M. Wainwright, *Towards optimal estimation of bivariate isotonic matrices with unknown permutations*, The Annals of Statistics **48** (2020), 3183–3205.

[3] N. Shah, S. Balakrishnan, A. Guntuboyina, and M. Wainwright, *Stochastically Transitive Models for Pairwise Comparisons: Statistical and Computational Issues*, IEEE Transactions on Information Theory **63** (2016), 934–959.

Robust estimation of the Gaussian mean by spectral dimension reduction

Arnak Dalalyan

(joint work with Amir-Hossein Bateni, Arshak Minasyan)

The broad goal of robust estimation is to design statistical procedures that are not very sensitive to small changes in data or to small departures from the modeling assumptions. A typical example, extensively studied in the literature, and considered in the present work, is when the data set contains outliers. In their vast majority, the well-established approaches of robust estimation treated the dimension of the parameter as a fixed and small constant. This simple setting was convenient for mathematical analysis and for computational purposes, but somewhat disconnected from many practical situations. Furthermore, it was hiding some fascinating phenomena that emerge only when the dimension is considered as a parameter that might be large, in the same way as the sample size.

In particular, it turns out that under the Huber contamination, the componentwise median is not minimax-rate optimal whereas the Tukey median is. More precisely, if a p-dimensional mean vector is to be estimated from n independent vectors drawn from the mixture distribution $(1-\varepsilon)\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \varepsilon \mathbf{Q}$ (where $\varepsilon \in (0, 1/2)$ is the rate of contamination and $\mathbf{Q}$ is the unknown distribution of outliers), then the mean squared error of the componentwise median is of order $p/n + p\varepsilon^2$ while that of Tukey's median is of order $p/n + \varepsilon^2$. Thus, as long as only statistical properties of the estimators are considered, Tukey's median is superior to its competitors, the componentwise and the geometric medians. However, the componentwise and the

geometric medians are better than the Tukey's median in terms of the breakdown point and in terms of computational complexity.

This observation led to the development of a number of computationally tractable estimators having an error with nearly the same dependence on dimension as that of Tukey's median. The goal of this talk is to make a step forward by designing an estimator which is not only nearly rate optimal and computationally tractable, but also has a breakdown point equal to $1/2$, which is the highest possible value of the breakdown point. To construct the estimator, termed iterative spectral dimension reduction or SDR, we combine and suitably adapt ideas from [1], [2] and [3]. The main underlying observation is that if we remove some clear outliers and restrict our attention to the subspace spanned by the eigenvectors of the sample covariance matrix corresponding to small eigenvalues, then the sample mean of the projected data points is a rate-optimal estimator. This allows us to iteratively reduce the dimension and eventually to estimate the remaining low-dimensional component of the mean by a standard robust estimator such as the componentwise median or the trimmed mean.

Importantly, the SDR estimator does not require as input the rate of contamination ε but only an upper bound on ε. We establish an upper bound on the error of the SDR estimator, showing that it is nearly minimax-rate optimal and has a breakdown point equal to $1/2$. This is done in the general case of a sub-Gaussian distribution with heterogeneous covariance matrix contaminated by adversarial noise. We then extend this result to the case where only an approximation to the covariance matrix is available.

References

[1] I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart, *Being robust (in high dimensions) can be practical.* In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Proceedings of Machine Learning Research **70** (2017), 999–1008.

[2] K. Lai, A. Rao and S. Vempala, *Agnostic estimation of mean and covariance.* In IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 665–674.

[3] A. Dalalyan and A. Minasyan, *All-in-one robust estimator of the Gaussian mean.* The Annals of Statistics **50** (2), 1193–1219.

Adaptive robustness and sub-Gaussian deviations in Sparse Linear Regression through Pivotal Double SLOPE

Mohamed Ndaoud

(joint work with Stas Minsker)

Consider the sparse linear model where some of the entries can be corrupted and the noise heavy-tailed, i.e.

$$Y = X\beta^* + \sqrt{n}\theta^* + \sigma\xi,$$

where β^* and θ^* are both sparse vectors of sparsity s and o respectively, and ξ the noise vector.

After deriving the minimax quadratic risk for estimation of β^*, we propose a practical and fully adaptive procedure that is optimal. Our procedure corresponds to solving the following pivotal estimation problem

$$\min_{\beta\in\mathbf{R}^p,\theta\in\mathbf{R}^n} \frac{1}{\sqrt{n}}\|Y - X\beta - \sqrt{n}\theta\| + \|\theta\|_\mu + \|\beta\|_\lambda,$$

where $\|\cdot\|_\eta$ corresponds to a sorted ℓ_1 norm that depends on some sequence $(\eta)_m$. Our procedure is not only minimax optimal and robust but also enjoys sub-Gaussian deviations even in the presence of heavy-tailed noise.

Identification and model selection for graphical continuous Lyapunov models

Mathias Drton

(joint work with Philipp Dettling, Carlos Améndola, Niels Richard Hansen, Mladen Kolar, Roser Homs Pons)

Graphical continuous Lyapunov models offer a new perspective on modeling causally interpretable dependence structure in multivariate data by treating each independent observation as a one-time cross-sectional snapshot of a temporal process. Specifically, the models consider multivariate Ornstein-Uhlenbeck processes in equilibrium [1, 2]. This leads to Gaussian models in which the covariance matrix is determined by the continuous Lyapunov equation. In this setting, each graphical model assumes a sparse drift matrix with support determined by a directed graph. In the research presented in Oberwolfach we discussed identifiability of such sparse drift matrices as well as their ℓ_1-regularized estimation.

Identifiability in graphical continuous Lyapunov models. We study the crucial problem of parameter identifiability in the class of graphical continuous Lyapunov models. Indeed, given a statistical model induced by a graph, it is essential for statistical analysis to clarify if it is possible to uniquely recover the parameters from the joint distribution of the observed variables. We show that this question can be reduced to analyzing the rank of certain sparse matrices with covariances as entries. Depending on the graph under consideration, the structure of these matrices changes in subtle ways. We study identifiability for different classes of graphs. For directed acyclic graphs, the matrices to be studied are block upper triangular and global identifiability is easily derived. However, cyclic graphs may also yield a globally identifiable parametrization and we prove, in our main result, that global identifiability holds if and only if the graph is simple (i.e., contains at most one edge between any two nodes). We computationally classify all graphs with up to 5 nodes, and present intriguing examples of non-simple graphs for which the associated model has generically identifiable parameters.

In future work it will be interesting to pursue a characterization of which non-simple graphs yield generically identifiable models. Moreover, the work we presented focused on the case where the volatility matrix of the Ornstein-Uhlenbeck

process is known up to a multiplicative constant, and it will be interesting to consider models in which the volatility matrix is unknown (but structured so).

On the Lasso for graphical continuous Lyapunov models. A natural approach to model selection for the considered graphical models is to use an ℓ_1-regularization approach that seeks to find a sparse approximate solution to the Lyapunov equation when given a sample covariance matrix. We study the model selection properties of the resulting lasso technique by applying the primal-dual witness technique for support recovery. Our analysis uses special spectral properties of the Hessian of the considered loss function to arrive at a consistency result. While the lasso technique is able to recover useful structure, our results also demonstrate that the relevant irrepresentability condition may be violated in subtle ways, preventing perfect recovery even in seemingly favorable settings.

Given the obstacles encountered with the direct application of ℓ_1 regularization, it would be interesting to consider alternative forms of regularized estimation and theoretically explore other possible loss functions. At a more fundamental level, it remains to develop a better understanding of whether/to which extent different graphs may induce identical statistical models.

References

[1] K. Fitch, *Learning directed graphical models from Gaussian data*, arXiv (2019).
[2] G. Varando and N.R. Hansen, *Graphical continuous Lyapunov models*, Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (2020), PMLR 124:989–998.

Optimal discriminant analysis in high-dimensional latent factor models

Marten Wegkamp
(joint work with Xin Bing)

In high-dimensional classification problems, a commonly used approach is to first project the high-dimensional features into a lower dimensional space, and base the classification on the resulting lower dimensional projections. In this talk, we formulate a latent-variable model with a hidden low-dimensional structure to justify this two-step procedure. We derive minimax lower bounds for the misclassification regret under these latent-variable models. We propose a computationally efficient classifier that takes certain principal components (PCs) of the observed features as projections, with the number of retained PCs selected in a data-driven way. Next, we show that our proposed method also performs favorably relative to other existing discriminant methods on three real data examples. A general theory is established for analyzing such two-step classifiers based on any low-dimensional projections. We derive explicit rates of convergence of the excess risk of the proposed PC-based classifier. The obtained rates are further shown to be optimal up to logarithmic factors in the minimax sense, only provided we modify the classifier slightly that involves data-splitting to handle the bias due to the low-dimensional projection. Interestingly, simulations show that there are scenarios when the rates are suboptimal if this data-splitting device isn't implemented. Our theory allows,

but does not require, the lower dimension to grow with the sample size and is also valid even when the feature dimension exceeds the sample size. We will show simulations that corroborate our theoretical findings.

Assigning topics to documents by successive projections

Olga Klopp

(joint work with Maxim Panov, Suzanne Sigalla, Alexandre B. Tsybakov)

Assigning topics to documents is an important task in several applications. For example, press agencies need to identify articles of interest to readers based on the topics of articles that they have read in the past. Analogous goals are pursued by many other text-mining applications such as, for example, recommending blogs from among the millions of blogs available. A popular approach to the problem of estimating hidden thematic structures in a corpus of documents is based on topic modeling.

In this paper, we adopt the *probabilistic Latent Semantic Indexing* (pLSI) model introduced in [5]. The pLSI model deals with three types of variables, namely, documents, topics and words. Topics are latent variables, while the observed variables are words and documents. Assume that we have a dictionary of p words and a collection of n documents. Documents are sequences of words from the dictionary. The number of topics is denoted by K. Usually, $K \ll \min(p, n)$. Throughout this paper, we assume that $2 \leq K \leq \min(p, n)$. The pLSI model assumes that the probability of occurrence of word j in a document discussing topic k is independent of the document. Therefore, by the total probability formula,

$$\mathbb{P}(\text{word } j|\text{document } i) = \sum_{k=1}^{K} \mathbb{P}(\text{topic } k|\text{document } i)\mathbb{P}(\text{word } j|\text{topic } k).$$

Introducing the notation $\Pi_{ij} := \mathbb{P}(\text{word } j|\text{document } i), W_{ik} := \mathbb{P}(\text{topic } k|\text{document } i)$ and $A_{kj} := \mathbb{P}(\text{word } j|\text{topic } k)$ we may write $\Pi_{ij} = W_i^{\mathrm{T}} A_j$, where $W_i = (W_{i1}, \ldots, W_{iK})^{\mathrm{T}} \in [0,1]^K$ is the topic probability vector for document i and $A_j = (A_{1j}, \ldots, A_{Kj})^{\mathrm{T}} \in [0,1]^K$ is the vector of word j probabilities under topics $k = 1, \ldots, K$:

$$\boldsymbol{\Pi} = \boldsymbol{W}\boldsymbol{A}, \tag{1}$$

where $\boldsymbol{\Pi}$ is the document-word matrix of size $n \times p$ with entries Π_{ij}, $\boldsymbol{W} := (W_1, \ldots, W_n)^{\mathrm{T}}$ is the document-topic matrix of size $n \times K$ and $\boldsymbol{A} := (A_1, \ldots, A_p)$ is the topic-word matrix of size $K \times p$. The rows of these matrices are probability vectors,

$$\sum_{m=1}^{K} W_{im} = 1, \ \sum_{j=1}^{p} A_{kj} = 1, \ \sum_{j=1}^{p} \Pi_{ij} = 1 \text{ for any } i = 1, \ldots, n, \ k = 1, \ldots K. \tag{2}$$

The value Π_{ij} is the probability of occurrence of word j in document i. It is not available but we have access to the corresponding empirical frequency X_{ij}. Thus, we have a document-word matrix $\boldsymbol{X} = (X_{ij})$ of size $n \times p$ such that for each

document i in $1,\dots,n$, and each word j in $1,\dots,p$, the entry X_{ij} is the observed frequency of word j in document i. Let N_i denote the (non-random) number of sampled words in document i. We can write the observation model in a "signal + noise" form:

$$\boldsymbol{X} = \boldsymbol{\Pi} + \boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A} + \boldsymbol{Z}, \tag{3}$$

where $\boldsymbol{Z} := \boldsymbol{X} - \boldsymbol{\Pi}$ is a zero mean noise. The objective in topic modeling is to estimate matrices $\boldsymbol{A}$ and $\boldsymbol{W}$ based on the observed frequency matrix $\boldsymbol{X}$ and on the known $N_1,\dots,N_n$. The recovery of $\boldsymbol{A}$ and the recovery of $\boldsymbol{W}$ address different purposes. An estimator of matrix $\boldsymbol{A}$ identifies the topic distribution on the dictionary. An estimator of $\boldsymbol{W}$ indicates the topics associated to each document. Estimation of $\boldsymbol{W}$ has multiple applications and has been extensively discussed in the literature, mainly in the Bayesian perspective. The focus was on Latent Dirichlet Allocation (LDA) and related techniques. These methods are computationally slow and, to the best of our knowledge, no theoretical guarantees on their performance are available.

On the other hand, estimation of matrix $\boldsymbol{A}$ is well-studied in the theory. Several papers provide bounds on the performance of different estimators of $\boldsymbol{A}$. Most of the results use the *anchor word assumption* postulating that for every topic there is at least one word, which occurs only in this topic, see [2, 4, 6]. At first sight, it seems that results on estimation of matrix $\boldsymbol{A}$ can be applied to estimation of $\boldsymbol{W}$ by simply taking the transpose of (2) and inverting the roles of these two matrices. However, such an argument is not valid since the resulting models are different. Indeed, the rows of matrix $\boldsymbol{X}^{\mathrm{T}}$ are not independent and the rows of matrices $\boldsymbol{\Pi}^{\mathrm{T}}, \boldsymbol{A}^{\mathrm{T}}, \boldsymbol{W}^{\mathrm{T}}$ do not sum up to 1, which leads to a different statistical analysis.

In the present paper, we change the framework by focusing on estimation of matrix $\boldsymbol{W}$ rather than $\boldsymbol{A}$. We introduce the following assumption, the *Anchor document assumption*: for each topic $k = 1,\dots,K$, there exists at least one document i (called an anchor document) such that $W_{ik} = 1$ and $W_{il} = 0$ for all $l \neq k$. Both anchor word and anchor document assumptions are very relevant in real word applications. Since each document is identified with a mixture of K topics, anchor document assumption means that, for each topic, there is a document devoted solely to this topic.

Our approach to estimation of matrix $\boldsymbol{W}$ that we call Successive Projection Overlapping Clustering (SPOC) is inspired by the Successive Projection Algorithm (SPA) initially proposed for non-negative matrix factorization [1]. The idea of such methods is to start with the singular value decomposition (SVD) of matrix $\boldsymbol{X}$, and launch an iterative procedure that, at each step, chooses the maximum norm row of the matrix composed of singular vectors and then projects on the linear subspace orthogonal to the selected row. From a geometric perspective, the rows of the matrix composed of singular vectors of $\boldsymbol{\Pi}$ belong to a simplex in $\mathbb{R}^K$. The documents can be identified with some points in this simplex and the anchor documents with its vertices. Our algorithm iteratively finds estimators of the vertices, based on which we finally estimate $\boldsymbol{W}$.

Note that the idea of exploiting simplex structures for estimation of matrix $\boldsymbol{A}$ rather than $\boldsymbol{W}$ was previously developed in, for example, [2, 6], among others. For example, the method to estimate $\boldsymbol{A}$ suggested in [6] is based on an exhaustive search over all size K subsets of $\{1, \ldots, p\}$. Its goal is to select K vertices of a p-dimensional simplex and its computational cost is at least of the order p^K. Our algorithm for estimating $\boldsymbol{W}$ recovers the vertices of much less complex object, which is a K-dimensional simplex (recall that $K \ll p$), and has much lower computational cost. Another important point is that existing simplex-based methods for estimation of matrix $\boldsymbol{A}$ require the number K of topics to be known. In the present paper, we propose a procedure that is adaptive to unknown K.

Our theoretical results deal only with the problem of estimating the topic-document matrix $\boldsymbol{W}$, for which the theory was not developed in prior work. But in practice, our method can be used for estimation of matrix $\boldsymbol{A}$ as well. Based on the SPOC estimator of $\boldsymbol{W}$, we immediately obtain an estimator of matrix $\boldsymbol{A}$ by a computationally fast procedure. Our simulation studies indicate that this estimator exhibits a behavior similar to LDA on average while being more stable.

We prove that the SPOC estimator of $\boldsymbol{W}$ converges in the Frobenius norm and in the ℓ_1-norm with the rates $\sqrt{n/N}$ and $n/\sqrt{N}$ (up to a weak factor[1]), respectively, assuming that $N_i = N$ for $i = 1, \ldots, n$. We also prove lower bounds of the order $\sqrt{n/N}$ and $n/\sqrt{N}$, respectively, implying near optimality of the proposed method. One of the conclusions, both from the theory and the numerical experiments, is that the error of the SPOC algorithm does not grow significantly with the size of the dictionary p, in contrast to what one observes for Latent Dirichlet Allocation. We also introduce an estimator for the number K of topics, which is usually unknown in practice. We show that SPOC algorithm using the estimator of K preserves its optimal properties in this more challenging setting.

References

[1] Mario Cesar Ugulino Araujo, Teresa Cristina Bezerra Saldanha, Roberto Kawakami Harrop Galvao, Takashi Yoneyama, Henrique Caldas Chame, and Valeria Visani. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems*, 57(2):65 – 73, 2001.

[2] Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*, pages 280–288, 2013.

[3] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 1–10. IEEE, 2012.

[4] Xin Bing, Florentina Bunea, and Marten Wegkamp. A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics. *Bernoulli 26(3): 1765-1796*, 2020.

[5] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

[1] In what follows, we mean by *weak factor* a small power of K multiplied by a term logarithmic in the parameters of the problem. We will ignore weak factors when discussing the convergence rates.

[6] Zheng Tracy Ke and Minzhe Wang. A new svd approach to optimal topic estimation. *arXiv preprint arXiv:1704.07016*, 2017.

Optimization and generalization in high dimensions: wide global minima of empirical risk

Peter L. Bartlett

(joint work with Olivier Bousquet, Philip M. Long, Gabor Lugosi, Alexander Tsigler)

We consider the impact of optimization methodology on statistical performance in high-dimensional prediction problems, motivated by the success of deep learning. We review empirical evidence that optimization algorithms that favor wide minima give better performance, and describe some recent analysis of one such algorithm, sharpness-aware minimization, highlighting questions about how wide global minima of empirical risk behave in these high-dimensional settings.

Zero order optimization of highly smooth functions

Arya Akhavan

(joint work with Evgenii Chzhen, Massimiliano Pontil, Alexandre Tsybakov)

This work studies minimization problems with zero-order noisy oracle information under the assumption that the objective function is highly smooth and possibly satisfies additional properties. The studied algorithm uses a gradient estimator based on randomization on the ℓ_2 sphere. The precise form that we consider is due to [2] and it has been used for zero order optimization of strongly convex functions. We present an improved analysis of this algorithm for the same class of functions and we derive rates of convergence for more general function classes. In particular, we consider functions which satisfies the Polyak-Łojasiewicz condition instead of strong convexity, and the larger class of highly smooth non-convex functions. We also analyse the case of quadratic, but not necessary strongly convex, functions, establishing improved rates of convergence. The improvements are achieved by new bounds on bias and variance for this algorithm, which is obtained via a Poincaré type inequality for uniform distribution on ℓ_2 sphere. The optimality of the upper bounds is discussed and a slightly more general lower bound than the state-of-the art bound in [1] is presented. These results imply that the proposed algorithm is nearly minimax optimal.

References

[1] Arya Akhavan, Massimiliano Pontil, and Alexandre B. Tsybakov. *Exploiting higher order smoothness in derivative-free optimization and continuous bandits*, Advances in Neural Information Processing Systems **33** (2020), 9017–9027.

[2] Francis Bach and Vianney Perchet. *Highly-smooth zero-th order online optimization*, Conference on Learning Theory (2016), 257–283.

Double machine learning methods: Beyond independence

Corinne Emmenegger
(joint work with Peter Bühlmann, Meta-Lina Spohn)

Double machine learning is a tool to combine machine learning algorithms and statistical models to estimate and make inference on low-dimensional parameters. These parameter estimators converge at the parametric rate and follow a Gaussian distribution asymptotically. We present three use cases: estimating the linear part in a partially linear endogenous model and in a partially linear mixed-effects model and treatment effect estimation for observational network data. For the latter, we provide more in-depth discussions.

Double machine learning and use cases. Semiparametric methods combine the flexibility of nonparametric approaches with ease of interpretation of parametric approaches. Double machine learning [1] is a tool to estimate and make inference on a low-dimensional parameter θ^0 in the presence of high- or infinite-dimensional nuisance components η^0 that satisfy some moment conditions

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\psi(S_i,\theta^0,\eta^0)]=0,$$

where N denotes the number of experimental units and ψ is a suitable function on the data S_i of the experimental units. The method uses sample splitting and cross-fitting. The data is partitioned into K many sets $I_1, I_2, \ldots, I_K$ of approximately equal size. For each $k \in \{1, 2, \ldots, K\}$, the nuisance components η^0 are estimated on the complement of I_k using an arbitrary machine learning algorithm and plugged into the estimating equation for θ^0. The data from I_k is then used to build an estimator $\hat{\theta}^{I_k}$ of θ^0 using this estimating equation. The final estimator of θ^0 averages over the $\hat{\theta}^{I_k}$, and it converges at the parametric rate, $N^{-1/2}$, and follows a Gaussian distribution asymptotically, provided ψ is Neyman orthogonal and the machine learning errors decay fast enough. Typically, these errors decay at the rate $o_P(N^{-1/4})$ if the problem is smooth and sufficiently sparse. Neyman orthogonality requires that the Gateau derivative of ψ vanishes at the true θ^0 and η^0, which makes ψ insensitive to inserting biased machine learning estimators of η^0. The algorithm is called "double" machine learning because η^0 consists of at least two objects, which means that machine learning algorithms are applied at least twice. Nonparametric components can also be estimated without sample splitting [6], but complex machine learners do not satisfy the entropy conditions these results require [1]. Consequently, sample splitting is essential.

This double machine learning base method can be extended to estimate the linear parameter in a partially linear endogenous model [2] or in a partially linear mixed-effects model [3] or to estimate the treatment effect from interacting units [4]. First, if endogeneity is present, so-called two-stage least squares is frequently applied, but it often outputs a large standard error. The regularization scheme proposed in [2] reduces the standard error and thus the confidence interval length, leading to preciser results. Second, mixed-effects models are frequently

used in clinical trials because they account for the correlation from observing the same patients repeatedly. Third, causal inference methods for treatment effect estimation usually assume independent experimental units. However, this assumption is often questionable because experimental units may interact. The vaccination (treatment) of a person not only influences this person's health status (outcome), but can also protect the health status of other people the person is interacting with. Ignoring interactions may yield biased estimators and invalid inference and contributes to the replication crisis. Subsequently, we describe our solution presented in [4] in more detail.

Treatment effect estimation from observational network data. For a dichotomous Bernoulli treatment $W_i \in \{0,1\}$ and a continuous outcome Y_i for units $i = 1, 2, \ldots, N$, our goal is to estimate and make inference for the expected average treatment effect (EATE)

$$\theta_N^0 = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\Big[Y_i^{do(W_i=1)} - Y_i^{do(W_i=0)}\Big],$$

where we use the do-notation of [8]. The EATE measures how, on average, the outcome Y_i of unit i is causally affected by its own treatment W_i in the presence of so-called spillover effects from other units. We consider a structural equation model as a data generating mechanism. The data on the unit level comes from sequentially evaluating the structured equations

$$\begin{aligned} C_i &\leftarrow \varepsilon_{C_i} \\ W_i &\leftarrow h^0(C_i, Z_i) + \varepsilon_{W_i} \\ Y_i &\leftarrow W_i g_1^0(C_i, X_i) + (1 - W_i) g_0^0(C_i, X_i) + \varepsilon_{Y_i}, \end{aligned}$$

where Z_i and X_i are user-specified features that capture spillover effects. The variables C_i are observed confounders. The ε_{C_i} as well as the ε_{Y_i} are i.i.d., and the endogenous error terms ε_{W_i} are independent across units and satisfy $\mathbb{E}[\varepsilon_{W_i} | C_i, Z_i] = 0$ within units. This approach can be extended to discrete responses [4]. Interactions among the units are encoded by the edges of a network, which is an undirected graph on the units. Spillover effects are along network paths. The Z_i features account for spillover from other units' confounders, and the X_i features account for spillover from other units' confounders and treatment assignments. For example, the user may choose as X_i the average number of treated neighbors and/or treated neighbors of neighbors of a unit.

We use the estimating equation

$$\begin{aligned} \theta_N^0 = \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \mathbb{E}\big[& g_1^0(C_i, X_i) - g_0^0(C_i, X_i) \\ & + \tfrac{W_i}{h^0(C_i, Z_i)} \big(Y_i - g_1^0(C_i, X_i)\big) - \tfrac{1 - W_i}{1 - h^0(C_i, Z_i)} \big(Y_i - g_0^0(C_i, X_i)\big) \big] \end{aligned}$$

for the EATE, which reminds us of augmented inverse probability weighting. The two summands in the second line above serve as a "bias correction" for biased machine learning estimators of $\eta^0 = (g_1^0, g_0^0, h^0)$ because they make the underlying ψ-function Neyman orthogonal. Sample splitting and cross-fitting are used to estimate the EATE. The propensity function h^0 and the two functions g_1^0 and

g_0^0 characterizing the outcome model are estimated using machine learning algorithms. In each of the K steps, the data used to estimate these functions needs to be independent from the data used to compute θ_N^0 from its estimating equation. Consequently, unit-level data points $S_i = (W_i, C_i, X_i, Z_i, Y_i)$ that are not independent from the data in I_k are removed to estimate η^0. The number of such removed data points depends on how far-reaching the Z- and X-features are. The resulting estimator of θ_N^0 converges at the parametric rate, $N^{-1/2}$, and follows a Gaussian distribution asymptotically. The underlying network cannot be arbitrarily complex, but it can become more dense as the number of units increases. The simulation study in [4] demonstrates the effectiveness of our method. Approaches that cannot correctly account for the correlation structure induced by the network yield biased results and invalid confidence intervals, whereas our method does not.

Other authors either uniformly limit the number of edges in the network, consider spillover from direct neighbors only, estimate densities, and assume a semiparametric model [7]; they do not incorporate observed confounding variables and consider spillover from the average of treated neighbors [5]; or they assume the network consists of multiple independent groups. We leverage all of these shortcomings at once: we perform entirely nonparametric regressions with arbitrary machine learning algorithms, include observed confounding variables, and allow the units to interact beyond direct neighborhoods in a general and increasingly complex network.

References

[1] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, *Double/debiased machine learning for treatment and structural parameters*, The Econometrics Journal **21:1** (2018), C1–C68.

[2] C. Emmenegger and P. Bühlmann, *Regularizing double machine learning in partially linear endogenous models*, Electronic Journal of Statistics **15:2** (2021), 6461–6543.

[3] C. Emmenegger and P. Bühlmann, *Plug-in machine learning for partially linear mixed-effects models with repeated measurements*, preprint arXiv:2108.13657 (2021).

[4] C. Emmenegger, M. Spohn, and P. Bühlmann, *Treatment effect estimation from observational network data using augmented inverse probability weighting and machine learning*, preprint arXiv:2206.14591 (2022).

[5] S. Li and S. Wager, *Random graph asymptotics for treatment effect estimation under network interference*, preprint arXiv:2007.13302 (2022).

[6] E. Mammen and S. van de Geer, *Penalized quasi-likelihood estimation in partial linear models*, The Annals of Statistics **25:3** (1997), 1014–1035.

[7] E. Ogburn, O. Sofrygin, I. Diaz, and M. Laan, *Causal inference for social network data*, preprint arXiv:1705.08527 (2017).

[8] J. Pearl, *Causal diagrams for empirical research*, Biometrika **82:4** (1995), 669–688.

The price of unfairness in linear bandits with biased feedback

Solenne Gaucher
(joint work with Alexandra Carpentier and Christophe Giraud)

Artificial intelligence is increasingly used in a wide range of decision making scenarii with higher and higher stakes. Recent works have shown that the decisions made by algorithms can be dangerously biased against certain categories of people, and have endeavored to mitigate this behavior. This work addresses the problem of online decision making under biased feedback. In the present talk, we consider a variant of the linear bandit problem, where the agent only has access to an unfair assessment of the action taken, that is systematically biased against a group of actions. For example, examiners may be prejudiced against people from a minority group, and give them lower grades; similarly, algorithms trained on biased data may produce unfair assessments of the credit risk of individuals belonging to a minority group. The problem of sequential decision making under biased feedback can be formalized as follows.

Biased linear bandit problem. A player is presented with a set of k distinct actions characterized by covariates $x \in \mathcal{X} \subset \mathbb{R}^d$, and by known sensitive attributes $z_x \in \{-1, 1\}$ indicating the group of the action. At each round $t \leq T$, the player chooses the action x_t and receives an unobserved reward $x_t^\top \gamma^*$, where $\gamma^* \in \mathbb{R}^d$ is the regression parameter specifying the true value of the action. The regret of the player is given by

$$R_T = \mathbb{E}\Big[\sum_{t \leq T} (x^* - x_t)^\top \gamma^*\Big], \quad \text{where} \quad x^* \in \arg\max_{x \in \mathcal{X}} x^\top \gamma^*.$$

By contrast to the classical linear bandit, the player does not observe a noisy version of the unbiased reward $x_t^\top \gamma^*$. Instead, she observes an unfair evaluation y_t of the value of the action $x_t^\top \gamma^*$, given by the following biased linear model:

$$y_t = x_t^\top \gamma^* + z_{x_t} \omega^* + \xi_t$$

where $\xi_t \overset{i.i.d}{\sim} \mathcal{N}(0, 1)$ is a noise term. The evaluation are systematically biased against a certain group: this unequal treatment of the groups is captured by the bias parameter $\omega^* \in \mathbb{R}$.

In the following, we assume that all covariates $x \in \mathcal{X}$ are distinct, which implies that the group z_x of action x is well defined. We also assume that no group is empty, that the set $\{\begin{pmatrix} x \\ z_x \end{pmatrix} : x \in \mathcal{X}\}$ spans $\mathbb{R}^{d+1}$ (which guarantees identifiability of the parameters), and that the rewards are bounded: $\max_{x \in \mathcal{X}} |x^\top \gamma^*| \leq 1$.

Preliminary discussion. The biased linear bandit is a variant of the classical linear bandit, where the agent observes a noisy version of the reward. Obviously, applying directly an algorithm designed for linear bandit to biased linear bandits without correcting the evaluations could lead to a linear regret if the evaluation mechanism is prejudiced against the group of the best action in terms of reward. To avoid this pitfall, one must estimate the bias in order to correct the evaluations.

This implies a change in the exploration-exploitation trade-off, as exploration becomes more expensive. Indeed, in classical bandit problems, one can compare the rewards of two actions by repeatedly sampling them - or, to put it differently, one can find the best action by sampling only those actions that seem optimal. This does not hold in the biased linear bandit: if, at some point, the set of potentially optimal actions contains representatives from both groups, and does not span $\mathbb{R}^d$, one is forced to sample sub-optimal actions to estimate the bias and improve the estimation of the unbiased rewards. This underlines the necessity to ensure sufficient estimation of the bias parameter, even when it implies sampling sub-optimal actions.

Worst case regret of the biased linear bandit. The following theorem characterizes the worst-case regret of the biased linear bandit. Before stating it, we introduce further notations : we write $\kappa_* = \min_{\pi \in \mathcal{P}^{\mathcal{X}}_{e_{d+1}}} e_{d+1}^\top \left(\sum_x \pi(x) \begin{pmatrix} x \\ z_x \end{pmatrix} \begin{pmatrix} x \\ z_x \end{pmatrix}^\top \right)^+ e_{d+1}$, where e_{d+1} denotes the $d+1$-th vector of the canonical basis in $\mathbb{R}^{d+1}$, M^+ denotes any generalized inverse of a matrix M, and $\mathcal{P}^{\mathcal{X}}_{e_{d+1}}$ denotes the set of probability measures π on $\mathcal{X}$ such that $e_{d+1} \in \text{Span}\left\{ \begin{pmatrix} x \\ z_x \end{pmatrix} : \pi(x) > 0 \right\}$.

Theorem 1 ([1])**.** *There exists a numerical constant $C > 0$ such that the following bound on the regret of the* Fair Phased Elimination *algorithm* [1] *holds*

$$R_T \leq C \kappa_*^{1/3} T^{2/3} \log(T)^{1/3}$$

for $T \geq T_{\kappa_*,k,d}$, where $T_{\kappa_*,k,d}$ is a constant depending on k, d and κ_*.

Lower bounds established in [1] show that the upper bound obtained in Theorem 1 is sharp in some settings, up to the sub-logarithmic factor $\log(T)^{1/3}$. These results show that the worst-case regret grows as $C\kappa_*^{1/3} T^{2/3} \log(T)^{1/3}$. This worst-case regret rate is higher than the typical rate $Cd \log(T) T^{1/2}$ obtained under unbiased feedback on the rewards. This increase in the regret corresponds to the cost of learning from unfair evaluations. It is due to the fact that the algorithm may need to sample actions that are sub-optimal in order to estimate the bias parameter. Note that this rate $\widetilde{\mathcal{O}}(T^{2/3})$ is typical for globally observable bandit problems with partial linear monitoring[2].

By contrast to previous results, Theorem 1 characterizes precisely the dependence of the worst-case regret on the geometry of the action set. The relevant constant κ_* is the minimal variance for estimating the bias, which appears when considering the related c-optimal design problem. The constant κ_* corresponds to the minimum number of samples required for estimating the bias with a variance equal to 1 (up to rounding issues). Intuitively, if the actions are very correlated with their sensitive attributes, more samples will be needed to estimate the bias with the same precision. This situation corresponds to cases where κ_* is large, and leads to a higher regret. The following Lemma relates κ_* to the margin between the two groups of actions.

Lemma ([1]). *κ_* is the largest constant $\kappa \geq 0$ such that, there exists an hyperplane H containing zero and separating the two groups, and such that, the margin to H is at least $\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ times the maximum distance of all points to the hyperplane H. When no such hyperplane exists, then $\kappa_* = 1$.*

Gap-dependent regret of the biased linear bandit. We characterize the worst-case regret of the biased linear bandit. Before stating our results, let us introduce further notations. We denote by $\Delta_x = (x^* - x)^\top \gamma^*$ the instantaneous regret (or *gap*) of action $x \in \mathcal{X}$, by $\Delta = (\Delta_x)_{x\in\mathcal{X}}$ the vector of gaps, and for any $\delta > 0$, by $\Delta \vee \delta = (\Delta_x \vee \delta)_{x\in\mathcal{X}}$. We write

$$\kappa(\Delta) = \min_{\mu \in \mathcal{M}_{\mathcal{X}}^{e_{d+1}}} \sum_x \mu(x)\Delta_x \text{ such that } e_{d+1}^\top \left(\sum_x \pi(x) \begin{pmatrix} x \\ z_x \end{pmatrix} \begin{pmatrix} x \\ z_x \end{pmatrix}^\top \right)^+ e_{d+1} \leq 1.$$

where $\mathcal{M}_{e_{d+1}}^{\mathcal{X}}$ denotes the measures on $\mathcal{X}$ such that $e_{d+1} \in \text{Span}\left\{ \begin{pmatrix} x \\ z_x \end{pmatrix} : \pi(x) > 0 \right\}$.

Theorem 2 ([1]). *Assume that $x^* \in \arg\max_{x\in\mathcal{X}} x^\top \gamma^*$ is unique. Then, there exists a numerical constant $C > 0$ such that the following bound on the regret of the* Fair Phased Elimination *algorithm* [1] *holds*

$$R_T \leq C \left(\frac{d}{\Delta_{\min}} \vee \frac{\kappa\big(\Delta \vee \Delta_{\neq} \vee \varepsilon_T\big)}{\Delta_{\neq}^2} \right) \log(T) \qquad \text{for} \quad T \geq T_{k,d,\Delta_{min}}$$

where $\Delta_{\min} = \min_{x\in\mathcal{X}\setminus x^*} \Delta_x$, $\Delta_{\neq} = \min_{x\in\mathcal{X}: z_x = -z_{x^*}} \Delta_x$, $\varepsilon_T = (\frac{\kappa_* \log(T)}{T})^{1/3}$, and $T_{k,d,\Delta_{min}}$ depends on k, d and $\Delta_{\min}$.

Lower bounds established in [1] show that the upper bound obtained in Theorem 2 is sharp in some settings up to a numerical constant. The term $\frac{d}{\Delta_{\min}} \vee \frac{\kappa(\Delta \vee \Delta_{\neq} \vee \varepsilon_T)}{\Delta_{\neq}^2}$ highlights the two sources of difficulty of the problem. On the one hand, the term $\frac{d \log(T)}{\Delta_{\min}}$ corresponds to the gap-dependent regret of a classical d-dimensional linear bandit. By contrast, the term $\frac{\kappa(\Delta \vee \Delta_{\neq} \vee \varepsilon_T)}{\Delta_{\neq}^2}$ is characteristic of the biased linear bandit problem. When $\frac{d}{\Delta_{\min}} \leq \frac{\kappa(\Delta \vee \Delta_{\neq} \vee \varepsilon_T)}{\Delta_{\neq}^2}$, the regret corresponds to the regret of this bias estimation phase. In other words, when both groups contain near-optimal actions, the difficulty of the problem is dominated by the price to pay for debiasing the unfair evaluations. Interestingly, when $\frac{d}{\Delta_{\min}} > \frac{\kappa(\Delta \vee \Delta_{\neq} \vee \varepsilon_T)}{\Delta_{\neq}^2}$, the difficulty of the linear bandit with systematic bias is dominated by that of the classical d-linear bandit. In this case, the algorithm is able to find the group containing the best action, and the problem reduces to a linear bandit in dimension d.

References

[1] Solenne Gaucher, Alexandra Carpentier, and Christophe Giraud. The price of unfairness in linear bandits with biased feedback. arXiv, 2022.

[2] Johannes Kirschner, Tor Lattimore, and Andreas Krause. Information Directed Sampling for Linear Partial Monitoring Proceedings of Thirty Third Conference on Learning Theory, PMLR 125:2328-2369, 2020.

Tensor PCA in high dimensional CP models

Cun-Hui Zhang
(joint work with Yuefeng Han)

The CP decomposition for high-dimensional non-orthogonal spiked tensors is an important problem with broad applications across many disciplines. However, previous works with theoretical guarantees typically assume restrictive incoherence conditions on the basis vectors for the CP components. In [2], we have proposed and studied composite PCA (CPCA) and iterative concurrent orthogonalization (ICO) algorithms for tensor CP decomposition with theoretical guarantees under much milder incoherence conditions. The CPCA applies the principal component or singular value decompositions twice, first to a matrix unfolding of the tensor data to obtain singular vectors and then to the matrix folding of the singular vectors obtained in the first step. It can be used as an initialization for any iterative optimization schemes for the tensor CP decomposition. The ICO iteratively estimates the basis vectors in each mode of the tensor by simultaneously applying projections to the orthogonal complements of the spaces generated by other CP components in other modes. It is designed to improve the alternating least squares (ALS) estimator and other forms of the high order orthogonal iteration for tensors with low or moderately high CP ranks, and it is guaranteed to converge rapidly when the error of any given initial estimator is bounded by a small constant. Both algorithms are applicable to deterministic tensor, its noisy version, and the order-$2K$ covariance tensor of order-K tensor data in a factor model with uncorrelated factors.

Tensor CP factor models. In the tensor CP factor model, we observe $d_1 \times \cdots \times d_K$ tensors X_i, $1 \le i \le n$, of the following form

$$X_i = \sum_{j=1}^{r} w_j f_{ij} a_{j1} \otimes a_{j2} \otimes \cdots \otimes a_{jK} + E_i,$$

where $\otimes$ denotes tensor product, f_{ij} are i.i.d $N(0,1)$, $w_j > 0$ represent signal strength, $a_{jk} \in \mathbb{R}^{d_k}$ are basis vectors with $\|a_{jk}\|_2 = 1$ for all $1 \le j \le r$, $1 \le k \le K$, and E_i are i.i.d. noise tensors each with i.i.d $N(0, \sigma^2)$ entries.

The covariance operator of the data, a tensor of order $2K$, can be written as

$$T = \frac{1}{n} \sum_{i=1}^{n} X_i \otimes X_i = \sum_{j=1}^{r} \lambda_j \otimes_{k=1}^{K} a_{jk}^{\otimes 2} + \Psi,$$

where $\lambda_j = w_j^2$ and Ψ is a random tensor with $\mathbb{E}[\Psi] = \sigma^2 \mathrm{Id}$. This is a spiked covariance tensor model as it is analogous to the spiked covariance matrix model in the study of matrix PCA in high dimensions [3]. However, here a_{jk}, $1 \le j \le r$, are not assumed orthogonal given mode k. The problem is to estimate λ_j and $a_{jk} a_{jk}^\top$. We note that a_{jk} is identifiable only up to $\pm a_{jk}$.

Among the most promising existing methods, [1] proposed to use clustering of power iterations of rank-1 random projections of T to obtain initial estimates of a_{jk} and the ALS to improve them. However, such methods require restrictive

incoherence conditions on each CP tensor basis, e.g. $\vartheta_{\max} \le \mathrm{polylog}(d_{\min})/\sqrt{d_{\min}}$ for 3-way tensors [1] where

$$\vartheta_{\max} = \max_{1\le k\le K} \vartheta_k, \quad \vartheta_k = \max_{1\le i<j\le r} |a_{ik}^\top a_{jk}|.$$

Our idea is to develop new methodologies to take advantages of the multiplicative nature of the co-linearity of the CP components. Let

$$\delta_{\max} = \max_{1\le k\le K} \delta_k, \quad \delta_k = \big\|A_k^\top A_k - I_r\big\|_{\mathrm{S}},$$

with the mode-k basis matrix $A_k = (a_{1k}, \ldots, a_{rk}) \in \mathbb{R}^{d_k\times r}$. If we vectorize the data points X_i, we would have a spiked covariance matrix

$$\mathrm{mat}_{[K]}\big(T\big) = \frac{1}{n}\sum_{i=1}^n \mathrm{vec}(X_i)\mathrm{vec}(X_i)^\top = \sum_{j=1}^r \lambda_j a_j^{\otimes 2} + \mathrm{mat}_{[K]}\big(\Psi\big),$$

with $a_j = \mathrm{vec}\big(\otimes_{k=1}^K a_{jk}\big)$. While $A = (a_1, \ldots, a_r)$ is still not orthonormal, a_j are much less co-linear than their counterparts in the individual modes.

Proposition 1. *Let $\vartheta = \max_{1\le i<j\le r} |a_i^\top a_j|$ and $\delta = \big\|A^\top A - I_r\big\|_{\mathrm{S}}$. Then,*

$$\vartheta \le \prod_{k=1}^K \vartheta_k, \quad \delta \le \Big((r-1)\vartheta\Big) \wedge \Big(\min_{k\le K}\delta_k\Big) \wedge \Big(\prod_{k\le K}\delta_k\Big).$$

When δ is small, a_j are not far from the j-th eigenvector of $\mathbb{E}\big[\mathrm{mat}_{[K]}\big(T\big)\big]$.

Proposition 2. *There exists an orthonormal matrix $U \in \mathbb{R}^{d\times r}$ such that $\|A\Lambda A^\top - U\Lambda U^\top\|_{\mathrm{S}} \le \delta\|\Lambda\|_{\mathrm{S}}$ for all nonnegative-definite matrices Λ in $\mathbb{R}^{r\times r}$.*

CPCA and ICO algorithms. In the tensor factor model, the CPCA and ICO algorithms are given in Tables 2 and 3 respectively.

TABLE 2. CPCA for pairwise symmetric tensors

Input:	$T = n^{-1}\sum_{i=1}^n X_i \otimes X_i$, CP rank r
	Formulate T to be a $d\times d$ matrix $\mathrm{mat}_{[K]}(T)$ with $d = \prod_{k=1}^K d_k$
	Compute $\{\widehat{\lambda}_j^{\mathrm{cpca}}, \widehat{u}_j\} = \mathrm{PCA}_j\big(\mathrm{mat}_{[K]}(T)\big)$
	Compute $\widehat{a}_{jk}^{\mathrm{cpca}} = \mathrm{LSVD}_1\big(\mathrm{mat}_k(\widehat{u}_j)\big) \in \mathbb{R}^{d_k}$
Output:	$\widehat{a}_{jk}^{\mathrm{cpca}},\ \widehat{\lambda}_j^{\mathrm{cpca}},\ j = 1, ..., r,\ k = 1, ..., K$

Here in CPCA, $\widehat{u}_j$ is the j-th eigenvector of the $d\times d$ matrix $\mathrm{mat}_{[K]}(T)$, $\widehat{u}_j$ is then reformatted into a $d_k \times (d/d_k)$ matrix $\mathrm{mat}_k(\widehat{u}_j)$, and $\widehat{a}_{jk}^{\mathrm{cpca}}$ is the leading left-singular vector of $\mathrm{mat}_k(\widehat{u}_j)$. According to Proposition 3 below, the second step of the CPCA is an contraction when the angular error of the first step is no more than 45 degrees. Thus, by Propositions 1 and 2, the error of CPCA is controlled by the much smaller, multiplicative incoherence measures ϑ and δ.

Proposition 3. *In the CPCA given in Table 2,*

$$\big(\|\widehat{a}_{jk}^{\mathrm{cpca}}(\widehat{a}_{jk}^{\mathrm{cpca}})^\top - a_{jk}a_{jk}^\top\|_{\mathrm{S}}^2\big) \wedge (1/2) \le \|\widehat{u}_j\widehat{u}_j^\top - a_j a_j^\top\|_{\mathrm{S}}^2.$$

TABLE 3. ICO for pairwise symmetric tensors

Input:	$T = n^{-1}\sum_{i=1}^n X_i \otimes X_i$, r, warm-start $\widehat{a}_{jk}^{(0)}$, $\epsilon > 0$, $M \geq 1$, $m = 0$
	Compute $(\widehat{b}_{1k}^{(1)}, ..., \widehat{b}_{rk}^{(1)}) = \big((\widehat{a}_{1k}^{(0)}, \ldots, \widehat{a}_{rk}^{(0)})^\top\big)^\dagger \in \mathbb{R}^{d_k \times r}$
	Repeat
	Set $m = m + 1$
	For $k = 1$ to K
	For $j = 1$ to r
	Compute $T_{jk}^{(m)} = T \times_{l \in [2K]\setminus\{k, K+k\}} (\widehat{b}_{jl}^{(m)})^\top$, $b_{j,K+l}^{(m)} = b_{jl}^{(m)}$
	Compute $\widehat{a}_{jk}^{(m)} = \mathrm{LPCA}_1 T_{jk}^{(m)} \in \mathbb{R}^{d_k}$
	End For
	Compute $(\widehat{b}_{1k}^{(m)}, ..., \widehat{b}_{rk}^{(m)}) = \big((\widehat{a}_{1k}^{(m)}, ..., \widehat{a}_{rk}^{(m)})^\top\big)^\dagger \in \mathbb{R}^{d_k \times r}$
	Set $(\widehat{b}_{1k}^{(m+1)}, ..., \widehat{b}_{rk}^{(m+1)}) = (\widehat{b}_{1k}^{(m)}, ..., \widehat{b}_{rk}^{(m)})$
	End For
	Until $m = M$ or $\max_{j,k} \|\widehat{a}_{jk}^{(m)}\widehat{a}_{jk}^{(m)\top} - \widehat{a}_{jk}^{(m-1)}\widehat{a}_{jk}^{(m-1)\top}\|_{\mathrm{S}} \leq \epsilon$
Output:	$\widehat{a}_{jk}^{\mathrm{ico}} = \widehat{a}_{jk}^{(m)}$, $\widehat{\lambda}_j^{\mathrm{ico}} = T \times_{k=1}^{2K} (\widehat{b}_{jk}^{(m)})^\top$, $j = 1, ..., r$, $k = 1, ..., K$

The ICO also takes advantage of the multiplicative nature of the CP basis, compared with ALS.

Proposition 4. *When the ICO in Table 3 is applied to* $T = \sum_{j=1}^r \lambda_j \otimes_{k=1}^r a_{jk}^{\otimes 2}$ *in the case of* $\sigma = 0$ *and* $n = \infty$, *for some numeric constant* C_0

$$\big\|\widehat{a}_{jk}^{\mathrm{new}}\widehat{a}_{jk}^{\mathrm{new}\top} - a_{jk}a_{jk}^\top\big\|_{\mathrm{S}} \leq C_0 \max_{jk} \big\|\widehat{a}_{jk}^{\mathrm{old}}\widehat{a}_{jk}^{\mathrm{old}\top} - a_{jk}a_{jk}^\top\big\|_{\mathrm{S}}^{2(K-1)}$$

Theoretical properties. Suppose $\lambda_1 \geq \cdots \geq \lambda_r > 0$. We summarize theoretical properties of CPCA and ICO in the tensor CP factor model as follows.

Theorem 1. *Suppose* $\sigma = 0$ *and* $n = \infty$. *The CPCA gives*

$$\big|\sin\theta(\widehat{a}_{jk}^{\mathrm{cpca}}, a_{jk})\big| = \sqrt{1 - \Big(a_{jk}^\top \widehat{a}_{jk}^{\mathrm{cpca}}\Big)^2} \leq \left(1 + \frac{2\lambda_1}{\lambda_{\mathrm{gap}}}\right)\delta =: \psi_0,$$

with $\lambda_{\mathrm{gap}} = \min_{j \leq r}\{\lambda_j - \lambda_{j+1}\}$. *When* $3(\lambda_1/\lambda_r)\psi_0^{2K-3} \leq \rho < 1$ *with the CPCA initialization, the ICO gives*

$$\sqrt{1 - \Big(a_{jk}^\top \widehat{a}_{jk}^{(m)}\Big)^2} \leq \psi_0\ \rho^{\gamma_K^{mK-1}} =: \psi_m,$$

and $\psi_m \leq \epsilon$ *within* $m = \lceil 1 + K^{-1}(\log\gamma_K)^{-1}\log(\log(\psi_0/\epsilon)/\log(1/\rho))\rceil$ *iterations for a certain constant* $\gamma_K \in [2, 3)$.

Theorem 2. *Let* $\tau \in [0, d]$. *With probability at least* $1 - e^{-\tau}$, *the CPCA gives*

$$\sqrt{1 - \Big(a_{jk}^\top \widehat{a}_{jk}^{\mathrm{cpca}}\Big)^2} \leq \underbrace{\left(1 + \frac{2\lambda_1}{\lambda_{\mathrm{gap}}}\right)\delta}_{bias} + \underbrace{C\left(\frac{\lambda_1}{\lambda_{\mathrm{gap}}}\right)\left(R^{(0)} + \sqrt{\tau/n}\right)}_{stochastic\ error}$$

with $R^{(0)} = \sqrt{(r_{\text{eff}}/n)(1+1/\text{SNR})(1+(r_{\text{eff}}/d)/\text{SNR})}$, *a constant* C *and*

$$\text{SNR} = \frac{\mathbb{E}\big\|\sum_{j=1}^r w_j f_{tj} \otimes_{k=1}^K a_{jk}\big\|_{\text{HS}}^2}{\mathbb{E}\|E_t\|_{\text{HS}}^2} = \frac{\sum_{j=1}^r \lambda_j}{\sigma^2 d} = \frac{r_{\text{eff}}\lambda_1}{\sigma^2 d}.$$

Theorem 3. *Suppose* $\psi_0 = (\delta + R^{(0)})\lambda_1/\lambda_{\text{gap}} < c_0/\sqrt{r}$ *where* c_0 *is a small constant. After at most* $O(\log\log(\psi_0/\psi_{\text{ideal}}))$ *ICO iterations, with probability at least* $1 - T^{-K} - \sum_k e^{-d_k}$, *the ICO with CPCA initialization gives*

$$\sqrt{1 - \left(\widehat{a}_{jk}^{\text{ico}\top} a_{jk}\right)^2} \le C_K \left(\sqrt{\frac{\sigma^2 d_{\max}}{\lambda_r n}} + \frac{\sigma^2}{\lambda_r}\sqrt{\frac{d_{\max}}{n}}\right) =: \psi_{\text{ideal}},$$

for all $1 \le j \le r$ *and tensor mode* $k \le K$, *where* C_K *depends on* K *only.*

Low-rank tensor denoising. In [2] we have also extended the CPCA and ICO algorithms and Theorems 1, 2 and 3 to the estimation of λ_j and a_{jk} in the following asymmetric noisy tensor CP model:

$$T = \sum_{j=1}^r \lambda_j \otimes_{k=1}^N a_{jk} + \Psi \in \mathbb{R}^{d_1 \times \cdots \times d_N}, \tag{1}$$

where $\lambda_j > 0$, $\|a_{j,k}\|_2 = 1$ and $\mathbb{E}[\Psi] = 0$. Again, $a_{jk}, 1 \le j \le r$, are allowed to have moderate correlations.

References

[1] Anandkumar, A., Ge, R., and Janzamin, M. (2014). Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *arXiv preprint arXiv:1402.5180*.

[2] Han, Y. and Zhang, C.-H. (2021). Tensor principal component analysis in high dimensional cp models. *arXiv preprint arXiv:2108.04428*.

[3] Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693.

From high-dimensional projection pursuit to interpolation in neural networks

Andrea Montanari

(joint work with Kangjie Zhou)

Given a cloud of n data points in $\mathbb{R}^d$, consider all projections onto m-dimensional subspaces of $\mathbb{R}^d$ and, for each such projection, the empirical distribution of the projected points. What does this collection of probability distributions look like when n, d grow large? We consider this question under the null model in which the points are i.i.d. standard Gaussian vectors, focusing on the asymptotic regime in which $n, d \to \infty$, with $n/d \to \alpha \in (0, \infty)$, while m is fixed. Denoting by $\mathscr{F}_{m,\alpha}$ the set of probability distributions in $\mathbb{R}^m$ that arise as low-dimensional projections in this limit, we establish several new results on this model:

Wasserstein radius for $m = 1$: Denoting by $W_2(P_1, P_2)$ the second Wasserstein distance between probability measures P_1 and P_2, we prove that $\sup\{W_2(P, \mathsf{N}(0,1)) : P \in \mathscr{F}_{1,\alpha}\} = 1/\sqrt{\alpha}$.

KL-Wasserstein outer bound: We show that, for any m, $\mathscr{F}_{m,\alpha}$ is contained in a W_2 neighborhood of the set of distributions P such that $D_{\mathrm{KL}}(P \| \mathsf{N}(\mathbf{0}, \boldsymbol{I}_m)) \le Cm\alpha^{-1}(1 \vee \log \alpha)$, with D_{KL} the Kullback-Leibler divergence.

Information dimension bound: Denoting by $\underline{d}(P)$ the lower information dimension of P, we prove that $\mathscr{F}_{m,\alpha}$ is contained in $\{P : \underline{d}(P) \ge m(1 - 1/\alpha)\}$ for $\alpha > 1$.

The previous question has application to unsupervised learning methods, such as projection pursuit and independent component analysis. We introduce a version of the same problem that is relevant for supervised learning, where the labels depend on k-dimensional projections of the covariates through a link function φ, and present the following results:

General ERM asymptotics: We consider a class of empirical risk minimization problems over functions $f : \mathbb{R}^d \to \mathbb{R}$ of the form $f(\boldsymbol{x}) = h(\boldsymbol{W}^\top \boldsymbol{x})$, and show that the asymptotics of the minimum empirical risk can be expressed in terms of the feasibility set $\mathscr{F}^{\varphi}_{m,\alpha}$.

Wasserstein bound for $m = 1$: We prove an outer bound on $\mathscr{F}^{\varphi}_{1,\alpha}$ for general $k = O(1)$, which generalizes the Wasserstein radius result obtained in the unsupervised setting. In fact, this outer bound characterizes the maximum W_2 distance between the empirical distribution of one-dimensional projections and the expected distribution.

Interpolation for two-layer networks: As a corollary to the previous result, we prove that a neural network with two-layers and m hidden neurons can separate n data points in d dimensions with margin κ only if $md \ge C\kappa^2 n$. Earlier bounds only required $md \ge Cn/\log(d/\kappa)$.

Margin distributions for linear classifier: We demonstrate the tightness of our W_2 bound by deriving the asymptotic distribution of the margins in linear max-margin classification.

References

[1] Peter J Bickel, Gil Kur, and Boaz Nadler. Projection pursuit in high dimensions. *Proceedings of the National Academy of Sciences*, 115(37):9151–9156, 2018.

[2] Jerome H Friedman and John W Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on computers*, 100(9):881–890, 1974.

[3] Persi Diaconis and David Freedman. Asymptotics of graphical projection pursuit. *The annals of statistics*, pages 793–815, 1984.

Perturbation bounds for (nearly) orthogonally decomposable Tensors with statistical applications

Ming Yuan
(joint work with Arnab Auddy)

We develop deterministic perturbation bounds for singular values and vectors of orthogonally decomposable tensors, in a spirit similar to classical results for matrices such as those due to Weyl, Davis, Kahan and Wedin. Our bounds demonstrate intriguing differences between matrices and higher-order tensors. Most notably, they indicate that for higher-order tensors perturbation affects each essential singular value/vector in isolation, and its effect on an essential singular vector does not depend on the multiplicity of its corresponding singular value or its distance from other singular values. Our results can be readily applied and provide a unified treatment to many different problems involving higher-order orthogonally decomposable tensors. In particular, we illustrate the implications of our bounds through connected yet seemingly different high dimensional data analysis tasks: the unsupervised learning scenario of tensor SVD and the supervised task of tensor regression, leading to new insights in both of these settings.

Laplace approximation in high dimension

Vladimir Spokoiny

High dimensional Laplace approximation has recently gained an increasing attention in connection with Bayesian inference for complicated nonlinear parametric models such as nonlinear inverse problems and Deep Neuronal Networks. The Laplace approximation is obtained by replacing a log density with its second order Taylor approximation around the point of maximum. This leads to a Gaussian measure centered at the maximum with a covariance corresponding to the Hessian of the negative log-density (see, e.g., [1, Section 4.4]). The asymptotic behavior of the parametric Laplace approximation in the small noise or large data limit has been studied extensively in the past (see, e.g., [18]). The asymptotic approximation of general integrals of the form $\int e^{\lambda f(x)}\, g(x)\, dx$ by Laplace's method is presented in [13, 18]. Non-asymptotic error bounds for the Laplace approximation can be found in [12] for the univariate case and in [8, 4] for the multivariate case. [5] studied the Laplace approximation error and its convergence in the limit $\lambda \to \infty$ in the multivariate case when the function f depends on λ. Coefficients appearing in the asymptotic expansion of the approximated integral are given in [10].

Laplace approximation is an important step in establishing the prominent Bernstein - von Mises (BvM) Theorem that quantifies the convergence of the scaled posterior distribution toward a Gaussian distribution in the large data or small noise limit. Parametric BvM theory is well-understood [17, 6]. Modern applications with a high dimensional parameter space and limited sample size pose new questions and identify new issues in study of applicability and accuracy of

Laplace approximation. We refer to [7] for a study a parametric BvM theorem for nonlinear Bayesian inverse problems with an increasing number of parameters. A number of papers discuss the BvM phenomenon for nonlinear inverse problems; see e.g. [11, 9, 2], where the convergence is quantified in a distance that metrizes the weak convergence. [16] showed that the Laplace approximation error in Hellinger distance converges to zero in the order of the noise level. The recent paper [3] provides a finite sample error of Laplace approximation for the total variation distance with an explicit dependence on the dimension and of the nonlinearity of the forward mapping. The Laplace approximation is also widely utilized for different purposes in computational Bayesian statistics; see e.g. [15].

Motivation: Gaussian approximation of the posterior. As one of the main motivation for this study, consider the problem of Bayesian inference for the log-likelihood function $L(\boldsymbol{\theta}) = L(\boldsymbol{Y}, \boldsymbol{\theta})$ with data $\boldsymbol{Y}$, a parameter $\boldsymbol{\theta} \in \mathbb{R}^p$ and a Gaussian prior $\pi \sim \mathcal{N}(\boldsymbol{\theta}_0, G^{-2})$. Here G^{-2} is a symmetric positive definite matrix in $\mathbb{R}^p$. Then the posterior density $\pi_G(\cdot)$ of $\boldsymbol{\theta}$ given $\boldsymbol{Y}$ is proportional to the product $\mathsf{e}^{L(\boldsymbol{\theta})}\, \mathsf{e}^{-\|G(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|^2/2}$:

$$\boldsymbol{\vartheta}_G | \boldsymbol{Y} \sim \pi_G(\boldsymbol{\theta}) \propto \exp\bigl\{ L(\boldsymbol{\theta}) - \|G(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|^2/2 \bigr\},$$

where the sign $\propto$ means equality up to a normalizing multiplicative constant. Assume that the penalized maximum likelihood estimator (pMLE) $\tilde{\boldsymbol{\theta}}_G$ is well defined: $\tilde{\boldsymbol{\theta}}_G = \arg\max_{\boldsymbol{\theta}} \bigl\{ L(\boldsymbol{\theta}) - \|G(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|^2/2 \bigr\}$. Clearly $\tilde{\boldsymbol{\theta}}_G$ is also maximizer of $\pi_G(\boldsymbol{\theta})$. That is why it is often referred to as maximum a posteriori (MAP) estimator. Let also the log-likelihood function $L(\boldsymbol{\theta})$ be twice differentiable and weakly concave. Define

$$F_G(\boldsymbol{\theta}) = -\nabla^2 L(\boldsymbol{\theta}) + G^2. \tag{1}$$

Assuming the latter expression to be positive definite for all considered $\boldsymbol{\theta}$, define also its square root $D_G(\boldsymbol{\theta}) = \sqrt{F_G(\boldsymbol{\theta})}$. We use the shortcut $\tilde{D}_G = D_G(\tilde{\boldsymbol{\theta}}_G)$. Laplace's approximation means that the posterior distribution π_G is close to the Gaussian distribution $\mathcal{N}(\tilde{\boldsymbol{\theta}}_G, \tilde{D}_G^{-2})$. A closely related Bernstein - von Mises phenomenon claims an approximation of the posterior by $\mathcal{N}(\tilde{\boldsymbol{\theta}}, D^{-2})$, where $\tilde{\boldsymbol{\theta}}$ is the MLE and $D^2 = F = -\nabla^2 L(\boldsymbol{\theta}^*)$ is the Fisher information matrix for the true parameter value $\boldsymbol{\theta}^*$; see e.g. [16] for a detailed discussion in context on nonlinear inverse problems. The mentioned results provide an efficient tool for Bayesian uncertainty quantification and constructing the elliptic credible sets as level sets of the approximating Gaussian distribution; see [3] or [14] for applications to drift and diffusion estimation.

This paper's contributions. This paper aims at reconsidering the classical results about Laplace approximation and to address the above issues. Below the list of the most important achievements in the paper.

Effective dimension and dimension free guarantees. We introduce the notion of *effective dimension* p_0 of the problem which can be small of moderate even for

huge parameter dimension p. The value p_0 is defined by an interplay between the information delivered by the data and information contained in the prior; see Section for more details. Further we establish explicit *non-asymptotic* and *dimension free* guarantees for the accuracy of Gaussian approximation of the posterior in *total variation* (TV) distance in terms of effective dimension; see Theorem 1. In the case when the non-penalized log-likelihood function grows linearly with the sample size n, the quality of Laplace approximation is of order $\sqrt{\mathsf{p}_0^3/n}$. It can be improved to p_0^3/n if instead of TV-distance, we limit ourselves to the class of centrally symmetric sets. The proofs combine classical variational arguments with sharp bounds for Gaussian quadratic forms. Conditions require that f is strongly concave and locally smooth with a uniform bound on the third Gateaux derivative of f in a local vicinity of the point of maximum.

Critical dimension. The result of Theorem 1 helps to address the issue of *critical dimension* for applicability of Laplace approximation: the relation $\mathsf{p}_0^3 \ll n$ between the sample size n and the effective dimension p_0 is sufficient for our main results. The result on concentration of the posterior only requires $\mathsf{p}_0 \ll n$.

Setup and conditions. Let $f(\boldsymbol{x})$ be a function in a high-dimensional Euclidean space $\mathbb{R}^p$ such that $\int \mathsf{e}^{f(\boldsymbol{x})}\,d\boldsymbol{x} = \mathsf{C} < \infty$, where the integral sign $\int$ without limits means the integral over the whole space $\mathbb{R}^p$. Then f determines a distribution P with the density $\mathsf{C}^{-1}\mathsf{e}^{f(\boldsymbol{x})}$. Let $\boldsymbol{x}^*$ be a point of maximum: $f(\boldsymbol{x}^*) = \sup_{\boldsymbol{u}\in\mathbb{R}^p} f(\boldsymbol{x}^*+\boldsymbol{u})$. We also assume that $f(\cdot)$ is smooth, more precisely, three or even four time differentiable. Introduce the negative Hessian $\mathsf{D}^2 = -\nabla^2 f(\boldsymbol{x}^*)$ and assume D^2 strictly positive definite. Given a function $g(\cdot)$, we consider the ratio of two integrals

$$\mathcal{I}(g) \triangleq \frac{\int g(\boldsymbol{u})\,\mathsf{e}^{f(\boldsymbol{x}^*+\boldsymbol{u})}\,d\boldsymbol{u}}{\int \mathsf{e}^{f(\boldsymbol{x}^*+\boldsymbol{u})}\,d\boldsymbol{u}}. \tag{2}$$

We aim at establishing a Gaussian approximation for $\mathcal{I}(g)$:

$$\mathcal{I}(g) \approx \mathcal{I}_{\mathsf{D}}(g) \triangleq \frac{\int g(\boldsymbol{u})\,\mathsf{e}^{-\|\mathsf{D}\boldsymbol{u}\|^2/2}\,d\boldsymbol{u}}{\int \mathsf{e}^{-\|\mathsf{D}\boldsymbol{u}\|^2/2}\,d\boldsymbol{u}} = \mathbb{E}g(\boldsymbol{\gamma}_{\mathsf{D}}), \qquad \boldsymbol{\gamma}_{\mathsf{D}} \sim \mathcal{N}(0, \mathsf{D}^{-2}).$$

The total variation distance between P and $\mathcal{N}(\boldsymbol{x}^*, \mathsf{D}^{-2})$ can be obtained as the supremum of $|\mathcal{I}(g) - \mathcal{I}_{\mathsf{D}}(g)|$ over all measurable functions $g(\cdot)$ with $|g(\boldsymbol{u})| \le 1$: $\mathrm{TV}\bigl(P, \mathcal{N}(\boldsymbol{x}^*, \mathsf{D}^{-2})\bigr) = \sup_{\|g\|_\infty \le 1} \bigl|\mathcal{I}(g) - \mathcal{I}_{\mathsf{D}}(g)\bigr|$.

Concavity. Below we implicitly assume the following condition.

$\boldsymbol{b}(\mathcal{C}_0)$: *There exists another operator* $\mathsf{D}_0^2 \le \mathsf{D}^2$ *in* $\mathbb{R}^p$ *such that the function* $f_0(\boldsymbol{u}) \triangleq f(\boldsymbol{x}^*+\boldsymbol{u}) + \frac{1}{2}\|\mathsf{D}\boldsymbol{u}\|^2 - \frac{1}{2}\|\mathsf{D}_0\boldsymbol{u}\|^2$ is concave. Equivalently, for all $\boldsymbol{x}$, $\nabla^2 f(\boldsymbol{x}) + \mathsf{D}^2 - \mathsf{D}_0^2 \le 0$.

Effective dimension. With $\mathsf{D}^2 = -\nabla^2 f(\boldsymbol{x}^*)$ and D_0^2 from $\boldsymbol{b}(\mathcal{C}_0)$, the *effective dimension* p_0 is defined as $\mathsf{p}_0 \triangleq \mathrm{tr}\bigl(\mathsf{D}_0^2\,\mathsf{D}^{-2}\bigr)$. Of course, $\mathsf{p}_0 \le p$ but the choice of

a proper penalty G^2 in (1) allows to avoid the "curse of dimensionality" issue and ensure a small or moderate effective dimension p_0 even for p large or infinite. The value p_0 determines the radius of the local vicinity $\mathcal{U}_0$ defined below. Namely, let us fix some $\nu < 1$, e.g. $\nu = 2/3$, and some $\boldsymbol{x} > 0$ ensuring that $\mathsf{e}^{-\boldsymbol{x}}$ is our significance level. Define

$$\mathsf{r}_0 = 2\sqrt{\mathsf{p}_0} + \sqrt{2\boldsymbol{x}}, \qquad \mathcal{U}_0 = \bigl\{\boldsymbol{u}\colon \|\mathsf{D}_0\boldsymbol{u}\| \le \nu^{-1}\mathsf{r}_0\bigr\}. \tag{3}$$

Local smoothness conditions. Let $p \le \infty$ and let $f(\cdot)$ be a three times continuously differentiable function on $\mathbb{R}^p$. We fix a reference point $\boldsymbol{x}$ and local region around $\boldsymbol{x}$ given by the local set $\mathcal{U}_0 \subset \mathbb{R}^p$ from (3). Also consider the second order Taylor approximation $f(\boldsymbol{x}+\boldsymbol{u}) \approx f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{u}\rangle + \frac{1}{2}\langle \nabla^2 f(\boldsymbol{x}), \boldsymbol{u}^{\otimes 2}\rangle$ and similarly the third order expansion and introduce the remainders $\delta_3(\boldsymbol{x},\boldsymbol{u}) = f(\boldsymbol{x}+\boldsymbol{u}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{u}\rangle - \frac{1}{2}\langle \nabla^2 f(\boldsymbol{x}), \boldsymbol{u}^{\otimes 2}\rangle$. Local smoothness of f or, equivalently, of f_0, at $\boldsymbol{x}$ will be measured by the quantity

$$\omega(\boldsymbol{x}) \triangleq \sup_{\boldsymbol{u}\in\mathcal{U}_0} \frac{1}{\|\mathsf{D}_0\boldsymbol{u}\|^2/2} \bigl|\delta_3(\boldsymbol{x},\boldsymbol{u})\bigr|. \tag{4}$$

We also set $\omega \triangleq \omega(\boldsymbol{x}^*)$. Our results apply under the condition $\omega \ll 1$. Local concentration of the measure P requires $\omega \le 1/3$. The main results about Gaussian approximation of P are valid under a stronger condition $\omega\,\mathsf{p}_0 \le 2/3$. Our main result describes the quality of approximation of the measure P by the Gaussian measure with mean $\boldsymbol{x}^*$ and covariance D^{-2} in total variation distance.

Theorem 1. *Let $\boldsymbol{b}X \sim P$. Suppose $\boldsymbol{b}(\mathcal{C}_0)$. Let $\mathcal{U}_0$ be defined by* (3)*. If ω from* (4) *satisfies $\omega \le 1/3$, then $P(\boldsymbol{b}X - \boldsymbol{x}^* \notin \mathcal{U}_0) \le \mathsf{e}^{-\boldsymbol{x}}$. If $\omega\,\mathsf{p}_0 \le 2/3$, then for any $g(\cdot)$ with $|g(\boldsymbol{u})| \le 1$, it holds for $\mathcal{I}(g)$ from* (2)

$$\bigl|\mathcal{I}(g) - \mathcal{I}_{\mathsf{D}}(g)\bigr| \le \frac{2(\diamondsuit + \mathsf{e}^{-\boldsymbol{x}})}{1 - \diamondsuit - \mathsf{e}^{-\boldsymbol{x}}} \le 4(\diamondsuit + \mathsf{e}^{-\boldsymbol{x}})$$

with $\diamondsuit = \diamondsuit_2 = \frac{0.75\,\omega\,\mathsf{p}_0}{1-\omega}$.

Corollary 1. *Under the conditions of Theorem 1, it holds*

$$\sup_{A\in\mathcal{B}(\mathbb{R}^p)} \bigl|P(\boldsymbol{b}X - \boldsymbol{x}^* \in A) - P(\boldsymbol{\gamma}_{\mathsf{D}} \in A)\bigr| \le 4(\diamondsuit_3 + \mathsf{e}^{-\boldsymbol{x}}),$$

References

[1] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[2] Giordano, M. and Kekkonen, H. (2020). Bernstein–von Mises theorems and uncertainty quantification for linear inverse problems. *SIAM/ASA Journal on Uncertainty Quantification*, 8(1):342–373.

[3] Helin, T. and Kretschmann, R. (2022). Non-asymptotic error estimates for the laplace approximation in bayesian inverse problems. *Numerische Mathematik*, 150(2).

[4] Inglot, T. and Majerski, P. (2014). Simple upper and lower bounds for the multivariate Laplace approximation. *Journal of Approximation Theory*, 186:1–11.
[5] Łapiński, T. M. (2019). Multivariate Laplace approximation with estimated error and application to limit theorems. *Journal of Approximation Theory*, 248:105305.
[6] Le Cam, L. (2012). *Asymptotic Methods in Statistical Decision Theory*. Springer Science & Business Media.
[7] Lu, Y. (2017). On the Bernstein-von Mises theorem for high dimensional nonlinear Bayesian inverse problems. https://arxiv.org/1706.00289.
[8] McClure, J. and Wong, R. (1983). Error bounds for multidimensional Laplace approximation. *Journal of Approximation Theory*, 37(4):372–390.
[9] Monard, F., Nickl, R., Paternain, G. P., et al. (2019). Efficient nonparametric Bayesian inference for X-ray transforms. *The Annals of Statistics*, 47(2):1113–1147.
[10] Nemes, G. (2013). An explicit formula for the coefficients in Laplace's method. *Constructive Approximation*, 38(3):471–487.
[11] Nickl, R. (2020). Bernstein–von Mises theorems for statistical inverse problems I: Schrödinger equation. *J. Eur. Math. Soc.*, 22:2697–2750.
[12] Olver, F. W. J. (1968). Error bounds for the Laplace approximation for definite integrals. *Journal of Approximation Theory*, 1(3):293–313.
[13] Olver, F. W. J. (1974). *Asymptotics and Special Functions*. Academic Press.
[14] Reich, S. and Rozdeba, P. J. (2020). Posterior contraction rates for non-parametric state and drift estimation. *Foundations of Data Science*, 2(3):333–349.
[15] Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.
[16] Schillings, C., Sprungk, B., and Wacker, P. (2020). On the convergence of the Laplace approximation and noise-level-robustness of Laplace-based Monte Carlo methods for Bayesian inverse problems. *Numerische Mathematik*, 145:915–971.
[17] Van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3. Cambridge university press.
[18] Wong, R. (2001). *Asymptotic Approximations of Integrals*. Society for Industrial and Applied Mathematics.

Meta-learning representations with contextual linear bandits

Karim Lounici
(joint work with Leonardo Cella, Massimiliano Pontil)

Meta-learning seeks to build algorithms that rapidly learn how to solve new learning problems based on previous experience. In this paper we investigate meta-learning in the setting of stochastic linear bandit tasks. We assume that the tasks share a low dimensional representation, which has been partially acquired from previous learning tasks. We aim to leverage this information in order to learn a new downstream bandit task, which shares the same representation. Our principal contribution is to show that if the learned representation estimates well the unknown one, then the downstream task can be efficiently learned by a greedy policy that we propose in this work. We derive an upper bound on the regret of this policy, which is, up to logarithmic factors, of order $r\sqrt{N}(1 \vee \sqrt{d/T})$, where N is the horizon of the downstream task, T is the number of training tasks, d the ambient dimension and $r \ll d$ the dimension of the representation. We highlight that our strategy does not need to know r. We note that if $T > d$ our bound

achieves the same rate of optimal minimax bandit algorithms using the true underlying representation (up to a logarithmic term).

References

[1] Leonardo Cella and Massimiliano Pontil. Multi-task and meta-learning with sparse linear bandits. In *The Conference on Uncertainty in Artificial Intelligence*, 2021.

[2] Leonardo Cella, Alessandro Lazaric, and Massimiliano Pontil. Meta-learning with stochastic linear bandits. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1360–1370. PMLR, 13–18 Jul 2020.

[3] Leonardo Cella, Karim Lounici, and Massimiliano Pontil. Multi-task representation learning with stochastic linear bandits. *arXiv preprint arXiv:2202.10066*, 2022.

[4] Jiachen Hu, Xiaoyu Chen, Chi Jin, Lihong Li, and Liwei Wang. Near-optimal representation learning for linear bandits and linear rl.

Shape-contrained thresholding bandit problem

Alexandra Carpentier

(joint work with James Cheshire, Maurilio Gutzeit, Andréa Locatelli, Pierre Ménard)

We consider the Thresholding Bandit Broblem (TBP), a sequential learning problem where the aim of the learner is to recover a set of actions such that their mean is above a given threshold τ - see [1] for the initial reference for this problem. More precisely, we consider a bandit problem with K distributions (arms) $(\nu)_{k\leq K}$, and we assume that ν_k is supported on $[0,1]$ for any k. For a given horizon T, at each time $t \leq T$, the learner is allowed to select an arm k_t and receives a sample $X_t \sim \nu_{k_t}$. Let us write $(\mu_k)_{k\leq K}$ for the means of the distributions, and let $\tau \in [0,1]$ be a threshold. the aim of the learner is to output at time T an estimator $\hat{Q}$ that encodes the set of arms such that their mean is above threshold, namely

$$(Q_k)_k = (2\mathbf{1}\{\mu_k \geq \tau\} - 1)_k,$$

where here $Q_k = 1$ if arm k is above the threshold and $Q_k = -1$ otherwise. In what follows, write

$$(\Delta_k)_k := (|\mu_k - \tau|)_k \ ,$$

for the vector of gaps to the threshold.

We define two measures of performances for algorithms in this setting, for any policy π and bandit problem ν. the simple regret is

$$\bar{r}_\nu^\pi(T) := \mathbb{E}_n u^\pi \max_{k:\hat{Q}_k \neq Q_k} |\tau - \mu_k| \ . \tag{1}$$

and the probability of error, i.e. the probability they misclassify at least one arm, is

$$\bar{e}_\nu^\pi(T) := \mathbb{P}_n u^\pi \exists k : \mu_k \neq \tau : \hat{Q}_k \neq Q_k \ .$$

In this report we only discuss the simple regret, but in the talk and the related work, the probability of error is also studied. For a subset of bandit problems $\mathcal{B}$, we define the associated minimax simple regret as:

$$r^*(T, \mathcal{B}) = \inf_{\pi} \sup_{\nu \in \mathcal{B}} \bar{r}_\nu^\pi(T), \tag{2}$$

where the infimum is taken on all policies and the supremum is taken on all bandit problems in $\mathcal{B}$. In what follows, and for positive sequences $(\phi_{K,T})_{K,T}, (\psi_{K,T})_{K,T}$ we write $\phi_{K,T} \asymp \psi_{K,T}$ if there exists two absolute constants $c, c' > 0$ such that $c'\psi_{K,T} \leq \phi_{K,T} \leq c\psi_{K,T}$.

We consider this bandit problem under several shape constraints on the means of the arms $(\mu_k)_k$ - namely for several set of problems $\mathcal{B}$:

- we consider the unconstrained case, namely $(\nu_k)_k$ are just assumed to be supported on $[0, 1]$ and no additional assumptions on the $(\mu_k)_k$ is made. In this case, one can prove that
$$r^*(T, \mathcal{B}) \asymp \sqrt{\frac{K \log K}{T}}.$$
- we consider the monotone case, namely we additionally assume that the sequence $(\mu_k)_k$ is monotone - w.l.g. $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_K$. In this case, one can prove that
$$r^*(T, \mathcal{B}) \asymp \sqrt{\frac{\log K}{T}}.$$
- we consider the concave case, namely we assume that the sequence $(\mu_k)_k$ is concave - for any k we have $\mu_{k-1} + \mu_{k+1} \leq \mu_k$. In this case, one can prove that
$$r^*(T, \mathcal{B}) \asymp \sqrt{\frac{\log \log K}{T}}.$$
- we finally consider the unimodal case, namely we assume that the sequence $(\mu_k)_k$ is unimodal - there exists $l \leq K$ such that for any $k < l$ we have $\mu_k \leq \mu_{k+1}$ and for any $k \geq l$ we have $\mu_k \geq \mu_{k+1}$. In this case, one can prove that
$$r^*(T, \mathcal{B}) \asymp \sqrt{\frac{K}{T}}.$$

These results appeared in the papers [2, 3, 4], along with some other related problem dependent results (on the probability of error e_ν^π). Other very relevant references to this line of work are to be found in the mentioned papers.

References

[1] Chen, Shouyuan, Tian Lin, Irwin King, Michael R. Lyu, and Wei Chen, *Combinatorial pure exploration of multi-armed bandits*, Advances in neural information processing systems 27 (2014).

[2] Cheshire, James, Pierre Ménard, and Alexandra Carpentier, *The influence of shape constraints on the thresholding bandit problem*, Conference on Learning Theory, PMLR (2020), 1228–1275.

[3] Cheshire, James, Pierre Ménard, and Alexandra Carpentier, *Problem Dependent View on Structured Thresholding Bandit Problems*, International Conference on Machine Learning, PMLR (2021), 1846–1854.
[4] Locatelli, Andréa, Maurilio Gutzeit, and Alexandra Carpentier, *An optimal algorithm for the thresholding bandit problem*, International Conference on Machine Learning, PMLR (2021), 1690–1698.

On the relationship between adaptive sampling and the suprema of empirical processes

Kevin Jamieson

(joint work with Romain Camilleri, Lalit Jain, Zohar Karnin, Julian Katz-Samuels)

We study different high-dimensional aspects of linear and combinatorial bandits. Given known finite subsets $\mathcal{X} \subset \mathbb{R}^d$,$\mathcal{Z} \subset \mathbb{R}^d$ and an unknown $\theta \in \mathbb{R}^d$, consider a sequential game where at each time the player chooses an $x_t in \mathcal{X}$ and then Nature reveals a noisy observation $y_t = \langle x_t, \theta \rangle + \epsilon_t$ where $\epsilon_t \sim \mathcal{N}(0,1)$. In as few time steps as possible, the learner's goal is to identify $\arg\max_{z \in \mathcal{Z}} \langle z, \theta \rangle$ with high probability. When $\mathcal{X}$ and $\mathcal{Z}$ are enumerable and the dimension d is small, we propose a simple algorithm that we show obtains a near-optimal sample complexity. This talk focuses on complications that arise when $\mathcal{Z}$ is intractably large to enumerate (such as all spanning trees) or when d is far larger than the desired sample complexity. In each case, we provide a computationally efficient algorithm that obtains a near-optimal sample complexity.

References

[1] J. Katz-Samuels, L. Jain, Z. Karnin, K. Jamieson, *An Empirical Process Approach to the Union Bound: Practical Algorithms for Combinatorial and Linear Bandits*, NeurIPS (2020).
[2] R. Camilleri, J. Katz-Samuels, K. Jamieson, *High-Dimensional Experimental Design and Kernel Bandits*, ICML (2021).

A gradient estimator for noisy zero-order optimization

Evgenii Chzhen

(joint work with Arya Akhavan, Massimiliano Pontil, Alexandre Tsybakov)

We consider the problem of online learning from sequentially observing noisy values of unknown functions $f_t : \mathbb{R}^d \to \mathbb{R}$, $t = 1, 2, \ldots$. The learning protocol is as follows. At each round t, the learner chooses $\boldsymbol{x}'_t, \boldsymbol{x}''_t \in \mathbb{R}^d$ and the adversary reveals

$$y'_t = f_t(\boldsymbol{x}'_t) + \xi'_t \qquad \text{and} \qquad y''_t = f_t(\boldsymbol{x}''_t) + \xi''_t,$$

where $\xi'_t, \xi''_t \in \mathbb{R}$ are noise variables, random or not. Based on the values $(\boldsymbol{x}_i, y'_i, y''_i)$ for $1 \le i \le t-1$, and on y'_t, y''_t, the learner outputs $\boldsymbol{x}_t \in \Theta$, where Θ is a given subset of $\mathbb{R}^d$, and suffers loss $f_t(\boldsymbol{x}_t)$. The goal of the learner is to find a strategy with small cumulative regret $\mathbf{E} \sum_{t=1}^T f_t(\boldsymbol{x}_t) - \inf_{\boldsymbol{x} \in \Theta} \sum_{t=1}^T f_t(\boldsymbol{x})$. To

achieve this goal, we consider the mirror descent algorithm, initializing at $\boldsymbol{z}_1 = 0$ and computing $\boldsymbol{x}_t$ according to the recursions

$$\boldsymbol{x}_t = \arg\max_{\boldsymbol{x}\in\Theta} \{\langle \boldsymbol{z}_t, \boldsymbol{x}\rangle - V(\boldsymbol{x})\}, \quad \boldsymbol{z}_{t+1} = \boldsymbol{z}_t - \eta_t \boldsymbol{g}_t, \tag{1}$$

where $V : \Theta \to \mathbb{R}$ is a given function, $(\eta_t)_{t\geq 1}$ a given sequence of positive numbers, and $\boldsymbol{g}_t$ is a gradient estimator. We propose to use a gradient estimator $\boldsymbol{g}_t$ based on ℓ_1-randomization defined as follows. Denote by B_1^d and ∂B_1^d the open unit ball and unit sphere in ℓ_1-norm, respectively. For $t \geq 1$, let $\boldsymbol{\zeta}_t$ be distributed uniformly on ∂B_1^d; r_t uniformly distributed on $[-1, 1]$; $h_t > 0$. We choose the query points $\boldsymbol{x}_t' = \boldsymbol{x}_t + h_t r_t \boldsymbol{\zeta}_t$, $\boldsymbol{x}_t'' = \boldsymbol{x}_t - h_t r_t \boldsymbol{\zeta}_t$ and define the gradient estimator as

$$\boldsymbol{g}_t \triangleq \frac{d}{2h_t}(y_t' - y_t'')\,\mathrm{sign}(\boldsymbol{\zeta}_t)K(r_t). \tag{2}$$

Here, $\mathrm{sign}(\cdot)$ is the coordinate-wise sign function and $K(\cdot)$ is a standard kernel for non-parametric estimation of the first derivative, cf. [1, 4].

The method defined above is analyzed in different setups yielding, in the particular cases previously studied in [3, 2, 4], either similar or better bounds. We state here selected results distinguishing between two possible assumptions on the noise.

Assumption 1 (Canceling noise)**.** *For $t \geq 1$, it holds that $\xi_t' = \xi_t''$ almost surely.*

Assumption 2 (Adversarial noise)**.** *For $t \geq 1$, it holds that: (i) $\mathbf{E}[(\xi_t')^2] \leq \sigma^2$ and $\mathbf{E}[(\xi_t'')^2] \leq \sigma^2$; (ii) $(\xi_t')_{t\geq 1}$ and $(\xi_t'')_{t\geq 1}$ are independent of $(\boldsymbol{\zeta}_t, r_t)_{t\geq 1}$.*

Consider first the setting where all f_t's are convex Lipschitz continuous functions. We assume that $p, q \in [1, \infty]$, $d \geq 3$, and set $p^*, q^* \in [1, \infty]$ such that $\frac{1}{p} + \frac{1}{p^*} = 1$ (respectively, for q). We denote by $\|\cdot\|_q$ the ℓ_q-norm on $\mathbb{R}^d$.

Assumption 3. *The following conditions hold:*

1. *The set $\Theta \subset \mathbb{R}^d$ is compact and convex.*
2. *There exists $V : \Theta \to \mathbb{R}$, which is 1-strongly convex on Θ w.r.t. the ℓ_p-norm and such that $\sup_{\boldsymbol{x}\in\Theta} V(\boldsymbol{x}) - \inf_{\boldsymbol{x}\in\Theta} V(\boldsymbol{x}) \leq R^2$, for some $R > 0$.*
3. *Each function $f_t : \mathbb{R}^d \to \mathbb{R}$ is convex on $\mathbb{R}^d$ for all $t \geq 1$.*
4. *For all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$, and all $t \geq 1$ we have $|f_t(\boldsymbol{x}) - f_t(\boldsymbol{x}')| \leq L\|\boldsymbol{x} - \boldsymbol{x}'\|_q$ for some constant $L > 0$.*

Set

$$\mathrm{b}_q(d) \triangleq \frac{1}{d+1} \cdot \begin{cases} q d^{\frac{1}{q}} & \text{if } q \in [1, \log(d)), \\ e\log(d) & \text{if } q \geq \log(d). \end{cases}$$

Theorem 1. *Let Assumptions 1 and 3 hold. Let $\boldsymbol{x}_t$ be defined by (1) - (2) with $\eta_t \equiv \eta = \frac{AL}{R}\sqrt{\frac{d^{-1-\frac{2}{q\wedge 2}+\frac{2}{p}}}{T}}$, $h_t \equiv h \leq \frac{R}{50\mathrm{b}_q(d)\sqrt{T}} d^{\frac{1}{2}+\frac{1}{q\wedge 2}-\frac{1}{p}}$, where $A = (3+3\sqrt{2})^{-1}$, and $K(\cdot) \equiv 1$. Then, for any $\boldsymbol{x} \in \Theta$ we have*

$$\mathbf{E}\left[\sum_{t=1}^{T}\big(f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x})\big)\right] \leq 14.5 \cdot RL\sqrt{T d^{1+\frac{2}{q\wedge 2}-\frac{2}{p}}}\ .$$

Theorem 2. *Let Assumptions 2 and 3 hold. Let $\boldsymbol{x}_t$ be defined by (1) - (2) with $\eta_t \equiv \eta = \frac{R}{\sqrt{TL}}\left(\frac{\sigma \mathrm{b}_q(d)}{\sqrt{2}R}\sqrt{Td^{4-\frac{2}{p}}}+\mathtt{A}Ld^{1+\frac{2}{q\wedge 2}-\frac{2}{p}}\right)^{-\frac{1}{2}}$, $h_t \equiv h = \left(\frac{\sqrt{2}R\sigma}{L\mathrm{b}_q(d)}\right)^{\frac{1}{2}} T^{-\frac{1}{4}}d^{1-\frac{1}{2p}}$, where $\mathtt{A} = 9(1+\sqrt{2})^2$, and $K(\cdot)\equiv 1$. Then, for any $\boldsymbol{x}\in\Theta$ we have*

$$\mathbf{E}\left[\sum_{t=1}^{T}\big(f_t(\boldsymbol{x}_t)-f_t(\boldsymbol{x})\big)\right] \le 14.5\cdot RL\sqrt{Td^{1+\frac{2}{q\wedge 2}-\frac{2}{p}}} + 2.4\cdot\sqrt{RL\sigma}T^{\frac{3}{4}}\cdot\begin{cases}\sqrt{qd^{1+\frac{1}{q}-\frac{1}{p}}} & \text{if } q\in[1,\log(d)),\\ \sqrt{e\log(d)d^{1-\frac{1}{p}}} & \text{if } q\ge\log(d).\end{cases}$$

We also propose a fully adaptive version of method (1) - (2), with η_t, h_t independent of L,σ, and show that it achieves bounds as in Theorems 1 and 2 with the same rates but slightly inflated constants.

Next, we consider the case where $f_t \equiv f$ for all t and a higher order smooth function $f:\mathbb{R}^d\to\mathbb{R}$. For a given horizon T, and a convex compact set Θ, we analyze the optimization error $\mathbf{E}f(\boldsymbol{x}_T)-f^\star$, where $f^\star=\min_{\boldsymbol{x}\in\Theta}f(\boldsymbol{x})$.

Definition 1 (Higher order smoothness)**.** *Fix some $\beta\ge 2$ and $L>0$. Denote by $\mathcal{F}_\beta(L)$ the set of all functions $f:\mathbb{R}^d\to\mathbb{R}$ that are $\ell=\lfloor\beta\rfloor$ times continuously differentiable and satisfy, for all $\boldsymbol{x},\boldsymbol{z}\in\mathbb{R}^d$ the Hölder condition*

$$\left\|f^{(\ell)}(\boldsymbol{x})-f^{(\ell)}(\boldsymbol{z})\right\| \le L\left\|\boldsymbol{x}-\boldsymbol{z}\right\|_2^{\beta-\ell},$$

where $\|\cdot\|$ is the standard ℓ_2-type norm on tensors of ℓth partial derivatives.

Assumption 4. *The function $f\in\mathcal{F}_\beta(L)\cap\mathcal{F}_2(\bar{L})$ for some $\beta\ge 2$ and $L,\bar{L}>0$.*

Definition 2 (α-gradient dominance)**.** *Let $\alpha>0$. Function $f:\mathbb{R}^d\to\mathbb{R}$ is called α-gradient dominant on $\mathbb{R}^d$ if f is differentiable on $\mathbb{R}^d$ and satisfies the Polyak-Łojasiewicz inequality: $2\alpha(f(\boldsymbol{x})-f^\star)\le\|\nabla f(\boldsymbol{x})\|_2^2$, $\forall\boldsymbol{x}\in\mathbb{R}^d$.*

Theorem 3. *Let $f:\mathbb{R}^d\to\mathbb{R}$ be an α-gradient dominant function and Assumptions 2, 4 hold. Let $\boldsymbol{x}_t$ be given by the gradient descent instance of (1) - (2) (that is, $V(\boldsymbol{x}):=\|\boldsymbol{x}\|_2^2/2$) with kernel K as in [4] and*

$$\eta_t=\min\left(\frac{c_d}{\bar{L}d},\frac{4}{\alpha t}\right),\qquad h_t=d^{\frac{\beta+2}{2\beta}}\cdot\begin{cases}t^{-\frac{1}{2\beta}} & \text{if } \eta_t=\frac{4}{\alpha t},\\ T^{-\frac{1}{2\beta}} & \text{if } \eta_t=\frac{c_d}{\bar{L}d}\end{cases}$$

for all $t=1,\dots,T$, where c_d is explicitly given and such that $c\le c_d\le c'$ for some absolute constants $c,c'>0$. If $\Theta=\mathbb{R}^d$, $\boldsymbol{x}_1$ is deterministic, and $T\ge d^{2-\frac{\beta}{2}}$, then

$$\mathbf{E}[f(\boldsymbol{x}_T)-f^\star]\le\frac{\mathtt{A}_1 d}{\alpha T}(f(\boldsymbol{x}_1)-f^\star)+\frac{\mathtt{A}_2}{\min\{\alpha,\alpha^2\}}\left(\frac{d^2}{T}\right)^{\frac{\beta-1}{\beta}},$$

where $\mathtt{A}_1,\mathtt{A}_2>0$ depend only on $\sigma^2,L,\bar{L},\beta$.

We show further that, for strongly convex functions f, the value $\min\{\alpha,\alpha^2\}$ in this bound can be replaced by α. We also analyze the non-convex case providing a stationary-point guarantee.

A key ingredient of our analysis is the following new Poincaré-type inequality.

Lemma 1. *Let $d \geq 3$. Assume that $G : \mathbb{R}^d \to \mathbb{R}$ is a continuously differentiable function and $\boldsymbol{\zeta}$ is distributed uniformly on ∂B_1^d. Then*

$$\mathrm{Var}(G(\boldsymbol{\zeta})) \leq \frac{4}{(d-2)(d-1)} \mathbf{E}\left[\|\nabla G(\boldsymbol{\zeta})\|_2^2 \left(1 + \sqrt{d}\|\boldsymbol{\zeta}\|_2\right)^2\right] .$$

Furthermore, if $G : \mathbb{R}^d \to \mathbb{R}$ is an L-Lipschitz function w.r.t. the ℓ_2-norm, then $\mathrm{Var}(G(\boldsymbol{\zeta})) \leq \frac{4L^2}{(d-2)(d-1)} \left(1 + \sqrt{\frac{2d}{d+1}}\right)^2$.

References

[1] Polyak, B.T. & Tsybakov, A.B. Optimal order of accuracy of search algorithms in stochastic optimization. *Problems of Information Transmission.* **26**, 45-53 (1990)

[2] Duchi, J. & Jordan, M. & Wainwright, M. & Wibisono, A. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory.* **61**, 2788–2806 (2015)

[3] Shamir, O. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research.* **18**, 1703–1713 (2017)

[4] Akhavan, A. & Pontil, M. & Tsybakov, A.B. Exploiting higher order smoothness in derivative-free optimization and continuous bandits. *Proceedings of NeurIPS-2020* (2020)

Sharp adaptive similarity testing with pathwise stability for ergodic diffusions

Angelika Rohde

(joint work with Johannes Brutsche)

Within the nonparametric diffusion model, we develop a multiple test to infer about similarity of an unknown drift b to some reference drift b_0: At prescribed significance, we simultaneously identify those regions where violation from *similarity* occurs, without a priori knowledge of their number, size and location. Here, a drift b is said to be similar to b_0 at tolerance $\eta \geq 0$ within some interval I if

$$b_0(x) - \eta \leq b(x) \leq b_0(x) + \eta \quad \text{for all } x \in I.$$

For $\eta > 0$, the null is a composite hypothesis. Although our test statistic is motivated by the idea of simultaneously testing $b \leq b_0 + \eta$ and $b \geq b_0 - \eta$ pointwisely, there is no evidence that the boundary cases are least favourable for the null hypothesis of similarity. Indeed, the stochastic order relation required for this purpose may be missing even for the corresponding local likelihood ratio statistics. The reason is that their distribution does not only depend on local values of the drift b, but on the entire drift function via the invariant density.

Our main results are the following.

(i) Based on a multiscale statistic and for any significance level $\alpha \in (0, 1)$, we construct a threshold level such that the resulting test ϕ_T^η for the similarity testing

problem satisfies

$$\limsup_{T\to\infty} \sup_{b\in H_0} \mathbb{E}_b \phi_t^\eta \leq \alpha, \tag{1}$$

where T denotes the time horizon of the diffusion's observation. Note that (1) is a substantially stronger statement than the pointwise relation $\limsup_{T\to\infty} \mathbb{E}_b \phi_t^\eta \leq \alpha$ for all $b \in H_0$. For the derivation of (1), we construct a random variable Y_η

- that provably dominates the test statistic uniformly on the similarity hypothesis in stochastic order asymptotically and
- whose distribution depends continuously on the level η of similarity, and Y_0 equals the limiting distribution of the test statistic under the simple null.

The cornerstone of the construction of Y_η is the identification of the weak limit of the multiscale test statistic *uniformly* in $b \in H_0$. Whereas weak limit results for supremum statistics like ours have been derived in various settings, the additional uniformity in the drift parameter accounting for the composite null hypothesis of similarity is new and considerably more involved on a mathematical level.

(ii) We prove optimality and adaptivity for the multiple similarity test in the minimax sense. We exemplarily consider the case of alternatives belonging to some Hölder ball where deviations are measured in weighted supremum norm which is the equivalent to weighted risk definitions in sharp adaptive drift estimation like in [1] or [3]. Our similarity test is shown to be rate-optimal in the minimax sense, adaptive in both the unknown Hölder exponent and the radius, optimal in the constant for the exponent ≤ 1 and here, even sharp adaptive in the radius. The hypotheses construction in the proof of the lower bound involves a delicate fixed point problem as the drift itself appears in the invariant density which pops up in the deviation measure between null and alternative.

Since diffusion models arise frequently as scaling limits of jump processes, another aspect is of particular importance: Small perturbations in the jump process modeling lead to a different drift in the limiting diffusion, but they might change the process class of the driving noise as well. In consequence, when testing for similarity of the drift to the drift of a reference model, the true noise may only be close to that of this reference model, and testing is only meaningful under some stability properties with respect to the driving noise specification. Besides, in applications where a nontrivial dependence structure in time is present, the driving noise is not given by Brownian motion. In such situations, fractional diffusion models have become increasingly popular. However, in contrast to standard diffusion, fractional diffusions neither belong to the process class of Markov processes nor semimartingales. Thus, the number of tools for statistical analysis is rather limited. In view of the non-availablity of efficient nonparametric testing procedures in such cases, stability justifies to use tests developed for the standard diffusion model at least in situations where the driving noise is close to Brownian motion in a suitable sense.

(iii) We address the problem of stability for fractional diffusion models where the driving Brownian motion is replaced by a fractional Brownian motion with Hurst

index $H \in (0,1)$. Note that $H = 1/2$ corresponds to standard Brownian motion. We prove that the test statistic built from observations in the fractional diffusion model has strong performance properties as the fractional driving noise approaches Brownian motion in the following sense:

- The test is uniformly over the hypothesis of similarity of approximate level α, i.e. (slightly simplified)

$$\limsup_{T\to\infty} \limsup_{H\to 1/2} \sup_{b\in H_0} \mathbb{E}_b^H \phi_t^\eta \leq \alpha \tag{2}$$

  where $\mathbb{E}_b^H$ denotes the expectation when applied to fractional diffusion with Hurst index H and drift b.
- We prove that the minimax optimality is preserved in a certain sense as the fractional driving process approaches Brownian mo-tion.

As our test statistic for ϕ_T^η involves a stochastic integral which is not even defined for fractional diffusion observations a priori, we first introduce a pathwise continuation of the statistic as a function of the data that is continuous with respect to the topology of uniform convergence. Then, uniformly over the similarity hypothesis, we prove that the test statistic built from observations for fractional driving noise converges for $H \to 1/2$ in probability to that built for standard Brownian motion. This uniformity, which is substantially harder to derive than the corresponding pointwise result for any fixed drift, is crucial in order to deduce (2). The preservation of minimax properties relies on $L_1(\mathbb{P})$-convergence of likelihood ratios of the fractional diffusion model to those of the standard model. This derivation is based on (deterministic) fractional calculus.

(iv) We outline how to extend our results to the multidimensional case. While the construction of a multiscale test statistic is possible in higher dimension, an identification of a dominating random variable as in (i) for $\eta > 0$ is not available a priori. However, in case of the simple null hy-pothesis $\eta = 0$, minimax rate-optimality can be attained. This extension is mainly of technical nature as local time does not exist in higher dimension. Our stability results for the fractional diffusion model do neither transfer straightforwardly. Instead, we present a pathwise sta- bility approach based on the theory of rough paths which was first introduced by Lyons. The presentation includes a rough path version of our test that is close in spirit to a similar rough path version of maximum likelihood estimators in [2].

References

[1] A. Dalalyan, *Sharp adaptive estimation of the drift function for ergodic diffusions*, Ann. Statist. **33** (2005),2500–2528.

[2] J. Diehl, P. Friz, and H. Mai, *Pathwise stability of likelihood estimators for diffusions via rough paths*, Ann. Appl. Probab. **16** (2016), 2169–2192.

[3] C. Strauch, *Exact adaptive pointwise drift estimation for multidimensional ergodic diffusions*, Probab. Theory Relat. Fields **164** (2016), 361–400.

Consistent classification in metric space

László Györfi
(joint work with Roi Weiss)

Let $(\mathbb{X}, \rho)$ be a separable metric space. Assume that the feature element X takes values in $\mathbb{X}$ and let its label Y take values in $\{1, \ldots, M\}$. The error probability of an arbitrary decision function $g : \mathbb{X} \to \{1, \ldots, M\}$ is

$$L(g) = \mathbb{P}\{g(X) \neq Y\},$$

while the error probability of the Bayes decision g^* is denoted by

$$L^* = \mathbb{P}\{g^*(X) \neq Y\}.$$

In the standard model of pattern classification, we are given labeled samples,

$$\mathbb{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\},$$

which are n independent copies of (X, Y). We assume that in addition to the labeled sample $\mathbb{D}_n$, we also have an independent unlabeled sample $\{X'_1, \ldots, X'_m\}$ where the X'_i's are independent copies of X. Introduce the data-driven partition $\mathbb{P}_m$ of $\mathbb{X}$ such that $\mathbb{P}_m$ is a Voronoi partition with the nucleus set $\{X'_1, \ldots, X'_m\}$, i.e.,

$$\mathbb{P}_m = \{A_{m,1}, A_{m,2}, \ldots, A_{m,m}\}$$

such that $A_{m,\ell}$ is the Voronoi cell around the nucleus X'_ℓ,

$$A_{m,\ell} = \left\{x \in \mathbb{X} : \ell = \text{argmin}_{1 \leq i \leq m} \rho(x, X'_i)\right\},$$

where tie breaking is done by indices, i.e., if X'_i and X'_j are equidistant from x, then X'_i is declared "closer" if $i < j$. Then, the prototype nearest neighbor classification rule is then defined by

$$g_n(x) = \text{argmax}_{1 \leq j \leq M} \sum_{i=1}^{n} \mathbb{I}_{\{Y_i = j, X_i \in A_{m,\ell}\}}, \qquad \text{if } x \in A_{m,\ell}.$$

Theorem 1 ([1]). *If $m = m_n \to \infty$ such that $m_n/n \to 0$, then the classification rule g_n is universally strongly consistent, that is, for any distribution of (X, Y),*

$$\lim_{n \to \infty} L(g_n) = L^* \qquad a.s.$$

References

[1] L. Györfi and R. Weiss, *Universal consistency and rates of convergence of multiclass prototype algorithms in metric spaces* , Journal of Machine Learning Research **22** (2021), 1–25.

On lower bounds for the bias-variance trade-off

Johannes Schmidt-Hieber
(joint work with Alexis Derumigny)

The summary below is an extension of the abstract in [2].

It is a common phenomenon that for high-dimensional and nonparametric statistical models, rate-optimal estimators balance squared bias and variance. Although this balancing is widely observed, little is known whether methods exist that could avoid this trade-off between bias and variance. Indeed for several machine learning methods such as neural networks, good generalization performance has been reported in the overparametrized regime. This behavior is highly counterintuitive as it suggests that the classical bias-variance trade-off does not hold [1, 3].

We propose a general strategy to obtain universal lower bounds on the variance that hold for any estimator with bias smaller than a prespecified bound. These bounds shows to which extent the bias-variance trade-off is unavoidable and allows us to quantify the loss of performance for methods that do not obey it. The approach is inspired by the Cramér-Rao lower bound, which lower bounds the variance by an expression involving the derivative of the bias. The underlying regularity conditions render this approach, however, impractical for nonparametric and highdimensional models. To circumvent these regularity conditions, our approach is based on a number of abstract inequalities for the variance involving the change of expectation with respect to different probability measures as well as information measures such as the Kullback-Leibler or chi-square-divergence. Some of these inequalities rely on a new concept of information matrices. We also show that these inequalities generalize the Cramér-Rao inequality in the sense that by taking appropriate limits and imposing the standard regularity conditions, the Cramér-Rao lower bound can be recovered.

The abstract change of expectation inequalities are applied to derive universal lower bounds for the bias-variance trade-off for several statistical models including the Gaussian white noise model, a boundary estimation problem, the Gaussian sequence model and the high-dimensional linear regression model. For these specific statistical applications, different types of bias-variance trade-offs occur that vary considerably in their strength.

For pointwise function estimation, we can prove that there is a U-shaped bias-variance curve in the sense that small bias or small variance will necessary inflate the mean squared error. More precisely, if the function space is a ball in the space of β-Hölder functions, we obtain a universal lower bound stating that the worst case bias B and the worst case variance V of any estimator must obey the inequality $B^{1/\beta}V \geq C/n$ with n the sample size and C a constant. This quantifies for instance by how much the worst case variance increases if the bias is forced to be small.

In the Gaussian sequence model, different phase transitions of the bias-variance trade-off occur. Although there is a non-trivial interplay between bias and variance, the rate of the squared bias and the variance do not have to be balanced in order to achieve the minimax estimation rate.

For the trade-off between integrated squared bias and integrated variance in the Gaussian white noise model, we combine the general strategy for lower bounds with a reduction technique. This allows us to link the original problem to the bias-variance trade-off for estimators with additional symmetry properties in a simpler statistical model.

References

[1] M. Belkin, D. Hsu, S. Ma, and S. Mandal *Reconciling modern machine-learning practice and the classical bias-variance trade-off*, Proceedings of the National Academy of Sciences **116** (2019), 15849–15854.

[2] A. Derumigny and J. Schmidt-Hieber, *On lower bounds for the bias-variance trade-off*, ArXiv preprint 2006.00278

[3] B. Neal, S. Mittal, A. Baratin, V. Tantia, M. Scicluna, S. Lacoste-Julien, and I. Mitliagkas, *A modern take on the bias-variance tradeoff in neural networks*, ArXiv preprint 1810.08591

Fundamental limits of generative deep neural networks

Helmut Bölcskei

(joint work with Dmytro Perekrestenko, Léandre Eberhard)

We show that every d-dimensional probability distribution of bounded support can be generated through deep ReLU networks out of a 1-dimensional uniform input distribution. What is more, this is possible without incurring a cost—in terms of approximation error measured in Wasserstein-distance—relative to generating the d-dimensional target distribution from d independent random variables. This is enabled by a vast generalization of the space-filling approach discovered in [1]. The construction we propose elicits the importance of network depth in driving the Wasserstein distance between the target distribution and its neural network approximation to zero. Finally, we find that, for histogram target distributions, the number of bits needed to encode the corresponding generative network equals the fundamental limit for encoding probability distributions as dictated by quantization theory.

References

[1] B. Bailey and M. J. Telgarsky, *Size-noise tradeoffs in generative networks*, Advances in Neural Information Processing Systems **31** (2018), 6489-6499.

Lower bounds for invariant statistical models with applications to PCA

Martin Wahl

We address the problem of deriving lower bounds for the estimation of principal components. A state-of-the-art result, obtained in [1], provides a non-asymptotic lower bound for the spiked covariance model with two groups of eigenvalues. To state their result, consider the statistical model defined by

$$(\mathbb{P}_U)_{U\in O(p)}, \qquad \mathbb{P}_U = \mathcal{N}(0, U\Lambda U^T)^{\otimes n}, \tag{1}$$

where $O(p)$ denotes the orthogonal group, $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_p)$ is a diagonal matrix with $\lambda_1 \geq \cdots \geq \lambda_p > 0$ and $\mathcal{N}(0, U\Lambda U^T)$ denotes a Gaussian distribution with expectation zero and covariance matrix $U\Lambda U^T$. This statistical model corresponds to observing n independent $\mathcal{N}(0, U\Lambda U^T)$-distributed random variables $X_1, \dots, X_n$, and we will write $\mathbb{E}_U$ to denote expectation with respect to $X_1, \dots, X_n$ having law $\mathbb{P}_U$. Moreover, in this model, the d-th principal subspace (resp. its corresponding orthogonal projection) is given by $P_{\leq d}(U) = \sum_{i\leq d} u_i u_i^T$, where $u_1, \dots, u_p$ are the columns of $U \in O(p)$.

Theorem 1 ([1])**.** *Consider the statistical model* (1) *with* $\lambda_1 = \cdots = \lambda_d > \lambda_{d+1} = \cdots = \lambda_p > 0$*. Then there is an absolute constant* $c > 0$ *such that*

$$\inf_{\hat{P}} \sup_{U\in O(p)} \mathbb{E}_U \|\hat{P} - P_{\leq d}(U)\|_2^2 \geq c \cdot \min\Big(\frac{d(p-d)}{n}\frac{\lambda_d\lambda_{d+1}}{(\lambda_d - \lambda_{d+1})^2}, d, p-d\Big),$$

where the infimum is taken over all estimators $\hat{P} = \hat{P}(X_1, \dots, X_n)$ *with values in the class of all orthogonal projections on* $\mathbb{R}^p$ *of rank* d *and* $\|\cdot\|_2$ *denotes the Hilbert-Schmidt (or Frobenius) norm.*

The proof of Theorem 1 is based on applying lower bounds under metric entropy conditions combined with the metric entropy of the Grassmann manifold. While Theorem 1 can be applied to different spiked structures, it is of limited use in settings with decaying eigenvalues, typically encountered in functional data analysis and kernel-based learning. To solve this problem, [2] and [3] developed a new approach based on a van Trees inequality (i.e. a Bayesian version of the Cramér-Rao inequality) tailored for invariant statistical models. The key ingredient was to explore the group equivariance of the model (1), saying that if $(X_1, \dots, X_n)$ has distribution $\mathbb{P}_U$, then $(VX_1, \dots, VX_n)$ has distribution $\mathbb{P}_{VU}$, $U, V \in O(p)$. For instance, a main consequence of the developed theory is the following extension of Theorem 1.

Theorem 2 ([3])**.** *Consider the statistical model* (1)*. Then, for each* $\delta > 0$*, we have*

$$\inf_{\hat{P}} \int_{O(p)} \mathbb{E}_U \|\hat{P} - P_{\leq d}(U)\|_2^2 \, dU \geq I_\delta$$

with infimum taken over all $\mathbb{R}^{p\times p}$*-valued estimators* $\hat{P} = \hat{P}(X_1, \dots, X_n)$ *and*

$$I_\delta = \tfrac{1}{1+2\delta} \max \Big\{ \sum_{i\le d} \sum_{j>d} x_{ij} \ : 0 \le x_{ij} \le \tfrac{2}{n} \tfrac{\lambda_i \lambda_j}{(\lambda_i - \lambda_j)^2} \quad \textit{for all } i \le d, j > d,$$
$$\sum_{i\le d} x_{ij} \le \delta \quad \textit{for all } j > d,$$
$$\sum_{j>d} x_{ij} \le \delta \quad \textit{for all } i \le d \Big\}.$$

The lower bound in Theorem 2 is characterized by doubly substochastic matrices whose entries are bounded by the inverses of the different Fisher information directions. Similar results can be stated for the matrix denoising problem and the group synchronization problem.

References

[1] T. Cai and Z. Ma and Y. Wu, *Sparse PCA: optimal rates and adaptive estimation*, Ann. Statist. **41** (2013), 3074–3110.

[2] M. Wahl, *Lower bounds for invariant statistical models with applications to principal component analysis*, Ann. Inst. Henri Poincaré Probab. Stat., to appear.

[3] M. Wahl, *Van Trees inequality, group equivariance, and estimation of principal subspaces*, Foundations of Modern Statistics. On the occasion of Volodia Spokoiny's 60th birthday. Springer Proceedings in Mathematics & Statistics, to appear.

Reporter: Arya Akhavan

Participants

Arya Akhavan
Istituto Italiano di Tecnologia (IIT), and
ENSAE, IP Paris
Via Morego, 30
16163 Genova
ITALY

Prof. Dr. Peter Bartlett
Computer Science Division
University of California, Berkeley
Soda Hall
Berkeley, CA 94720
UNITED STATES

Prof. Dr. Pierre Bellec
Department of Statistics and
Biostatistics
RUTGERS
The State University of New Jersey
501 Hill Center, Busch Campus
110 Frelinghuysen Road
Piscataway NJ 08854
UNITED STATES

Prof. Dr. Mike Bing
Department of Statistical Science
Cornell University
1188 Comstock Hall
Ithaca, NY 14853-2601
UNITED STATES

Prof. Dr. Helmut Bölcskei
Mathematical Information Sciences
ETH Zürich
Room: ETF E 122
Sternwartstrasse 7
8092 Zürich
SWITZERLAND

Prof. Dr. Victor-Emmanuel Brunel
École Nationale de la Statistique
et de l'Adm. Economique
ENSAE
5 Avenue Le Chatelier
91120 Palaiseau Cedex
FRANCE

Prof. Dr. Peter Bühlmann
Seminar für Statistik
ETH Zürich (HG G 17)
Rämistrasse 101
8092 Zürich
SWITZERLAND

Prof. Dr. Florentina Bunea
Department of Statistics and Data
Science
Cornell University
Comstock Hall
Ithaca NY 14853-2601
UNITED STATES

Prof. Dr. Cristina Butucea
CREST - ENSAE
5, Avenue Henry Le Chatelier
91120 Palaiseau Cedex
FRANCE

Prof. Dr. Alexandra Carpentier
Institut für Mathematik
Universität Potsdam
Postfach 601553
14415 Potsdam
GERMANY

Dr. Leonardo Cella
Istituto Italiano di Tecnologia
Via Morego, 30,
16163 Genova
ITALY

Prof. Dr. Kamalika Chaudhuri
Department of Computer Science and Engineering
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0112
UNITED STATES

Julien Chhor
École Nationale de la Statistique et de l'Adm. Economique
ENSAE
5, avenue Henry Le Chatelier
91764 Palaiseau
FRANCE

Dr. Evgenii Chzhen
Institut de Mathématique d'Orsay
CNRS, Université Paris-Saclay
Orsay
91405 Orsay Cedex
FRANCE

Prof. Dr. Arnak Dalalyan
ENSAE / CREST
École Nationale de la Statistique et de l'Administration Économique
5, Avenue Henry Le Châtelier
91120 Palaiseau Cedex
FRANCE

Prof. Dr. Mathias Drton
Technische Universität München
Lehrstuhl für Mathematische Statistik
Boltzmannstr. 3
85748 Garching bei München
GERMANY

Corinne Emmenegger
Seminar für Statistik
ETH Zürich (HG G 17)
Rämistrasse 101
8092 Zürich
SWITZERLAND

Prof. Dr. Rina Foygel Barber
Department of Statistics
The University of Chicago
5747 S. Ellis Avenue
Chicago, IL 60637-1514
UNITED STATES

Solenne Gaucher
Laboratoire de Mathématiques
Université Paris Sud (Paris XI)
Batiment 307
rue Michel Magat
91405 Orsay Cedex
FRANCE

Prof. Dr. Christophe Giraud
Laboratoire de Mathématiques d'Orsay
Université de Paris-Saclay
Bat. 307
91405 Orsay Cedex
FRANCE

Prof. Dr. László Györfi
Department of Computer Science and Information Theory
Budapest University of Technology and Economics
Stoczek u. 2
1521 Budapest
HUNGARY

Dr. Kevin Jamieson
Department of Computer Science & Engineering
University of Washington
Box 352350
Seattle WA 98195-2350
UNITED STATES

Dr. Olga Klopp
ESSEC Business School
CS 50105 Cergy
3, Avenue Bernard Hirsch
95021 Cergy-Pontoise / Cedex
FRANCE

Prof. Dr. Vladimir Koltchinskii
School of Mathematics
Georgia Institute of Technology
686 Cherry Street
Atlanta, GA 30332-0160
UNITED STATES

Gil Kur
Laboratory for Information and Decision Systems
Massachusetts Institut of Technology
77 Massachusetts Avenue
Cambridge MA 02139
UNITED STATES

Prof. Dr. Elizaveta Levina
Department of Statistics
University of Michigan
323 West Hall, 1085 S. University Ave
Ann Arbor MI 48109-1107
UNITED STATES

Dr. Karim Lounici
Centre de Mathématiques
École Polytechnique
Plateau de Palaiseau
91128 Palaiseau Cedex
FRANCE

Prof. Dr. Enno Mammen
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 205
69120 Heidelberg
GERMANY

Prof. Dr. Andrea Montanari
Department of Electrical Engineering and Department of Statistics
Stanford University
Stanford CA 94305-4065
UNITED STATES

Prof. Dr. Jaouad Mourtada
École Nationale de la Statistique e de l'Adm. Economique
ENSAE
5, avenue Henry Le Chatelier
91120 Palaiseau
FRANCE

Prof. Dr. Boaz Nadler
Department of Computer Science and Applied Mathematics
The Weizmann Institute of Science
234 Herzl Street
P.O. Box 26
Rehovot 76100
ISRAEL

Dr. Mohamed Ndaoud
École Nationale de la Statistique et de l'Administration Économique
CREST - UMR 9194
5, Avenue Henry le Châtelier
91764 Palaiseau Cedex
FRANCE

Prof. Dr. Richard Nickl
Centre for Mathematical Sciences
Wilberforce Road
Cambridge CB3 0WA
UNITED KINGDOM

Prof. Dr. Robert Nowak
Department of Electrical and Computer Engineering
University of Wisconsin-Madison
1415 Engineering Drive
Madison WI 53706
UNITED STATES

Dr. Sofia Olhede
EPFL
MA-C1-573
Station 8
1015 Lausanne
SWITZERLAND

Dr. Dimitris Papailiopoulos
Department of Electrical and Computer Engineering
University of Wisconsin-Madison
Madison, WI 53706-1685
UNITED STATES

Rahul Parhi
Department of Electrical and Computer Engineering
University of Wisconsin-Madison
Engineering Hall # 3627
1550 Engineering Drive
Madison WI 53706
UNITED STATES

Prof. Dr. Philippe Rigollet
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge MA 02139-4307
UNITED STATES

Prof. Dr. Alessandro Rinaldo
Department of Statistics and Data Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh MA 15213
UNITED STATES

Prof. Dr. Angelika Rohde
Fakultät für Mathematik
Albert-Ludwigs-Universität Freiburg
LST für Stochastik
Ernst-Zermelo-Straße 1
79104 Freiburg i. Br.
GERMANY

Prof. Dr. Johannes Schmidt-Hieber
Department of Applied Mathematics
University of Twente
P.O.Box 217
7500 AE Enschede
NETHERLANDS

Prof. Dr. Vladimir G. Spokoiny
Weierstrass Institute for Applied Analysis
and Stochastics (WIAS) and Humboldt University
Mohrenstraße 39
10117 Berlin
GERMANY

Julia Struwe
Ludwigstr. 99
04315 Leipzig
GERMANY

Prof. Dr. Ryan Tibshirani
Departments of Statistics and Machine Learning
Carnegie Mellon University
229B Baker Hall
Pittsburgh PA 15213
UNITED STATES

Prof. Dr. Alexandre B. Tsybakov
CREST - ENSAE
5, Avenue Henry Le Châtelier
91120 Palaiseau Cedex
FRANCE

Prof. Dr. Sara van de Geer
Seminar für Statistik
ETH Zürich (HG G 24.1)
Rämistrasse 101
8092 Zürich
SWITZERLAND

Dr. Nicolas Verzelen
UMR 729, MISTEA
SUPAGRO
Bat. 29
2, Place Viala
34060 Montpellier Cedex 1
FRANCE

Dr. habil. Martin Wahl
Institut für Mathematik
Humboldt Universität zu Berlin
Unter den Linden 6
10099 Berlin
GERMANY

Dr. Kaizheng Wang
Department of IEOR
Columbia University
500 W 120th St Room 315
New York NY 10027
UNITED STATES

Prof. Dr. Marten Wegkamp
Department of Mathematics
Cornell University
Malott Hall
Ithaca, NY 14853-7901
UNITED STATES

Prof. Dr. Ming Yuan
Department of Statistics
Mailcode 2377
Columbia University
1255 Amsterdam Avenue
New York NY 10027
UNITED STATES

Prof. Dr. Anru Zhang
2424 Erwin Road, Durham NC 27710,
USA
Department of Biostatistics &
Bioinformatics
Duke University
2424 Erwin Road
Durham NC 27710
UNITED STATES

Prof. Dr. Cun-Hui Zhang
Department of Statistics
Rutgers University
110 Frelinghuysen Road
Piscataway NJ 08854-8019
UNITED STATES

Prof. Dr. Huibin Zhou
Department of Statistics and Data
Science
Yale University
24 Hillhouse Ave
P.O. Box 208290
New Haven CT 06520-8290
UNITED STATES