Mathematisches Forschungsinstitut Oberwolfach

Report No. 30/2024

DOI: 10.4171/OWR/2024/30

Statistics and Learning Theory in the Era of Artificial Intelligence

Organized by
Florentina Bunea, Ithaca
Arnak Dalalyan, Palaiseau
Robert Nowak, Madison
Sara van de Geer, Zürich

23 June – 28 June 2024

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): 68T07.

Introduction by the Organizers

The workshop highlighted recent theoretical advances and views on statistics, learning theory and artificial intelligence (AI) based on the interplay of techniques from mathematical statistics, machine learning, theoretical computer science and related areas. The workshop brought together researchers from these areas in order to exchange ideas and explore open mathematical problems at the intersection of mathematical statistics and machine learning. The anticipated long-term intellectual impact of the workshop is the enrichment of each of the represented disciplines through collaborations born during the meeting. The practical outcomes are the development of new theory and computationally efficient algorithms for machine learning based on a strong mathematical foundation.

The practice of AI has advanced at an unprecedented pace in recent years. However, progress in developing a rigorous mathematical framework for understanding

advanced AI models, such as deep neural networks, has lagged behind. Establishing solid mathematical foundations is crucial for creating principled engineering practices that ensure future AI systems will be reliable, interpretable, and safe.

The research lectures and discussions in this workshop mark the beginning of this foundational effort. Our understanding of high-dimensional statistical models, particularly for pattern classification and regression, is already well-developed. Many talks illustrated how rigorous mathematics can offer profound insights into the design, strengths, and limitations of such models. Ongoing work seeks to extend this understanding from classical models, such as linear models, to the more complex modern models used in AI.

Several lectures delved into the emerging theory of deep learning, which concerns neural networks with many layers. These models construct function spaces through compositions of functions, where each function is expressed as a linear combination of atoms from continuous dictionaries – the so-called neurons of the network. Unlike traditional nonparametric methods in statistics and data science, often grounded in reproducing kernel Hilbert spaces (RKHS), the compositional nature of neural network function spaces represents a significant departure. The talks and collaborations during the workshop addressed many challenges in this new domain.

A major difficulty in deep learning and related machine learning methods is the non-convexity of the optimization problems involved in training neural networks, a stark contrast to the typically convex problems in RKHS-based methods. Many lectures explored the challenges posed by non-convex optimization in machine learning and AI. The research presented paves the way for a deeper understanding of both the methods and models used in practice and improved strategies for optimizing them.

The lectures can be roughly categorized into three groups:

Regularization in High-Dimensional Models: This group of lectures focuses on advanced techniques and methods for regularizing models in high-dimensional spaces, a key challenge in modern machine learning. Speakers, including *Schmidt-Hieber*, *Yang*, *Skalski*, *Schneider*, *Nadler*, *Wegkamp*, *Kato*, *Rohde*, *Sen*, *Bradic*, *Neubert*, *Blum*, *Suvorikova*, and *Katsevich*, explored the theoretical foundations and practical applications of regularization to enhance model performance and generalization in complex, high-dimensional datasets.

Theory for Understanding Deep Neural Networks: These lectures provide insights into the theoretical underpinnings of deep neural networks. Experts such as *Jacot*, *Bach*, *Belkin*, *Willett*, *Rigollet*, and *Hundrieser* discussed recent developments in fundamental theory related to the learning dynamics, expressivity, and generalization of neural networks, offering deeper comprehension of how these models function and their limitations.

Theory for Optimization in AI and Machine Learning: This set of lectures centers on the mathematical and algorithmic theory behind optimization techniques crucial for artificial intelligence and machine learning. Researchers like *Bellec*, *Vinayak*, *Chamon*, *Brunel*, *Niles-Weed*, *Tibshirani*, *Manole*, *Hucker*, and

Spokoiny discuss cutting-edge approaches to improving optimization methods, which are vital for training models efficiently and effectively in both theory and practice.

Acknowledgement: The MFO and the workshop organizers would like to thank the National Science Foundation for supporting the participation of junior researchers in the workshop by the grant DMS-2230648, "US Junior Oberwolfach Fellows".

Workshop: Statistics and Learning Theory in the Era of Artificial Intelligence

Table of Contents

Abstracts

Statistical analysis of dropout in the linear model

Johannes Schmidt-Hieber
(joint work with Gabriel Clara, Sophie Langer)

Applying gradient descent for the loss function $\boldsymbol{\theta} \mapsto L(\boldsymbol{\theta})$ leads to the iterates

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} - \alpha_k \nabla L\big(\boldsymbol{\theta}_{k-1}\big), \quad k = 1, 2, \dots$$

with $\boldsymbol{\theta}_0$ the initialization, $\alpha_k > 0$ the learning rate, and ∇L the gradient of L with respect to $\boldsymbol{\theta}$.

DropConnect samples in every iteration a diagonal $d \times d$ random matrix D_k with diagonal entries independently drawn from a Bernoulli distribution with success probability p. We refer to a matrix with this distribution as a *dropout matrix (with success probability p)*. The gradient descent parameter updates with dropout are then given by

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} - \alpha_k \nabla L\big(D_k \boldsymbol{\theta}_{k-1}\big), \quad k = 1, 2, \dots \tag{1}$$

It is important to notice that DropConnect first applies the dropout matrix and then takes the gradient. This means that $\nabla L\big(D_k \boldsymbol{\theta}_{k-1}\big)$ has to be understood as taking the gradient with respect to the function $\boldsymbol{\theta} \mapsto L(D_k \boldsymbol{\theta})$, which is different from evaluating the gradient of ∇L at $D_k \boldsymbol{\theta}$.

The dropout matrices randomly select subsets of the parameters that are updated. For the k-th round, the j-th parameter is in this subset if and only if the Bernoulli variable on the j-th diagonal entry of D_k is 1.

DropConnect forces the model to still perform well if arbitrary subsets of the parameters are deactivated. One can then argue that this makes the model more robust and possibly also prevents overfitting.

Dropout is a similar procedure that is more specifically designed for training deep neural networks. In every gradient step, dropout randomly drops units in the network (except the output units). This is equivalent to randomly dropping columns in the weight matrices.

The effect of DropConnect can be decomposed in two parts that separates the variability of the dropout matrices from the mean behavior. If the matrix D has the same distribution as $D_{\ell+1}$ and $\mathbb{E}_D$ is the expectation with respect to D, define

$$\overline{\boldsymbol{\theta}}_k = \overline{\boldsymbol{\theta}}_{k-1} - \alpha_k \mathbb{E}_D\big[\nabla L\big(D \overline{\boldsymbol{\theta}}_{k-1}\big)\big].$$

Since the gradient is taken with respect to $\boldsymbol{\theta} \mapsto L(D\boldsymbol{\theta})$, this means that the gradient descent scheme $(\overline{\boldsymbol{\theta}}_k)_k$ aims to minimize the averaged loss function

$$\boldsymbol{\theta} \mapsto \mathbb{E}_D\big[L(D\boldsymbol{\theta})\big]. \tag{2}$$

In some cases, $\mathbb{E}_D\big[L(D\boldsymbol{\theta})\big]$ can be rewritten as penalized loss L. This shows that $(\overline{\boldsymbol{\theta}}_k)_k$ aims to minimize a regularized objective. All the regularization that is induced by $(\overline{\boldsymbol{\theta}}_k)_k$ is called *explicit regularization*.

The differences $(\boldsymbol{\theta}_k - \overline{\boldsymbol{\theta}}_k)_k$ add noise to the gradient scheme. This can have a regularizing effect that is then called the *implicit regularization* of dropout.

For dropout, [3] compared the generalization behavior of the dropout iterates with the dynamics of $(\overline{\boldsymbol{\theta}}_k)_k$. They observe a difference and dropout performs better. It is also shown that by adding suitable random noise to $(\overline{\boldsymbol{\theta}}_k)_k$, one recovers a very similar performance than the dropout scheme. This suggests that the differences $(\boldsymbol{\theta}_k - \overline{\boldsymbol{\theta}}_k)_k$ just adds extra noise to the gradient descent. We will show this theoretically for the linear model.

Assume that we want to fit parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)^\top$ to the data $(\mathbf{X}_1, Y_1)$, $\dots, (\mathbf{X}_n, Y_n)$ via the least squares functional $L(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^n \left(Y_i - \boldsymbol{\beta}^\top \mathbf{X}_i\right)^2$. This can also be written as $L(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{Y} - X\boldsymbol{\beta}\|_2^2$ with response vector $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ and X the $n \times d$ matrix with $\mathbf{X}_i^\top$ as the i-th row. The linear model can be viewed as a neural network without hidden layers. In the linear model, dropout and DropConnect are the same.

The explicit regularizer has been studied in the original dropout article [2]. We investigate the implicit regularizer for fixed learning rate. For that we rewrite the gradient descent iterates as vector autoregressive process with random coefficients. If $\boldsymbol{\beta}_k$ denotes the k-th gradient descent iterate, we use an extension of the Gauss-Markov theorem to show that the implicit effect of dropout in the linear model is characterized by the covariance $\mathrm{Cov}(\boldsymbol{\beta}_k - \widehat{\boldsymbol{\beta}})$ with $\widehat{\boldsymbol{\beta}}$ the estimator that minimizes the averaged loss function $\boldsymbol{\beta} \mapsto \mathbb{E}_D[L(D\boldsymbol{\beta})]$ defined in (2). An analysis of this covariance reveals that there is no implicit effect of dropout if the Gram matrix $X^\top X$ is diagonal, but an implicit effect occurs whenever $X^\top X$ has at least two non-zero entries in every column. In the latter case we can quantify how much extra noise is added by the implicit regularization. All details can be found in [1].

References

[1] G. Clara, S. Langer, J. Schmidt-Hieber *Dropout regularization versus ℓ^2-penalization in the linear model*, Journal of Machine Learning Research, *to appear*.

[2] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*, Journal of Machine Learning Research, **15** (2014), 1929–1958.

[3] C. Wei, S. Kakade, T. Ma *The implicit and explicit regularization effects of dropout*, 37th International Conference on Machine Learning, (2020), 10181–10192.

How DNNs break the Curse of Dimensionality: Compositionality and Symmetry Learning

Arthur Jacot

(joint work with Seok Hoan Choi, Yuxiao Wen)

We show that deep neural networks (DNNs) can efficiently learn any composition of functions with bounded F_1-norm, which allows DNNs to break the curse of dimensionality in ways that shallow networks cannot. More specifically, we derive a

generalization bound that combines a covering number argument for compositionality, and the F_1-norm (or the related Barron norm) for large width adaptivity. We show that the global minimizer of the regularized loss of DNNs can fit for example the composition of two functions $f^* = h \circ g$ from a small number of observations, assuming g is smooth/regular and reduces the dimensionality (e.g. g could be the modulo map of the symmetries of f^*), so that h can be learned in spite of its low regularity. The measures of regularity we consider is the Sobolev norm with different levels of differentiability, which is well adapted to the F_1 norm. We compute scaling laws empirically and observe phase transitions depending on whether g or h is harder to learn, as predicted by our theory.

References

[1] A. Jacot, S.H. Choi, Y. Wen, *How DNNs break the Curse of Dimensionality: Compositionality and Symmetry Learning*, https://arxiv.org/abs/2407.05664 (2024).

Surprising phenomena of max-ℓ_p-margin classifiers in high dimensions

Fanny Yang

In recent years, the analysis of max-ℓ_p-margin classifiers has gained attention from the theory community not only due to the implicit bias of first-order methods, but also due to the observation of harmless interpolation for neural networks. Here, interpolation refers to the classifier achieving zero classification error on the training data. Our work contributes to this active area of research with two results for linear classification in ambient dimension d given n i.i.d. samples. We first show that surprisingly, in the noiseless case, while minimizing the ℓ_1-norm achieves minimax-optimal rates for *regression* for hard-sparse ground truths [1], this adaptivity does not directly apply analogously to max-ℓ_1-margin classification [2]. In particular, while known interpolating estimators that are based on minimizing non-convex optimization problems achieve an $O(1/n)$ rate, the max-ℓ_1-margin estimator (minimizing a relaxed convex objective) only achieves rates of order $O(1/n^{1/3})$.

Further, when the observations are noisy, we prove how max-ℓ_p-margin classifiers can achieve $O(1/\sqrt{n})$ rates for p slightly larger than one when $d > n^2$, while the maximum ℓ_1-margin classifier only achieves rates of order $\frac{1}{\sqrt{\log(d/n)}}$ for all $d > \omega(n)$ [3, 2]. Notably, for $p > 1$, max-ℓ_p-margin classifiers can achieve a faster, minimax-optimal $O(1/\sqrt{n})$ rate in the noisy case, than the max-ℓ_1-margin estimator in the noiseless case. Together, these two results suggest a number of new open problems for future work: What is the underlying mechanism that $p > 1$ can lead to a better noiseless recovery? The minimizer of a non-interpolating max-average-margin objective subject to an ℓ_1-norm constraint, also achieves minimax-optimal rates for noisy recovery. Could this regularized max-average-margin estimator achieve the optimal $O(1/n)$ rate in the noiseless case, while interpolating max-margin interpolators can only achieve $O(1/\sqrt{n})$?

References

[1] G. Wang, K. Donhauser, and F. Yang, *Tight bounds for minimum ℓ_1-norm interpolation of noisy data.*, Proceesdings of the International Conference on Artificial Intelligence and Statistics **2022**, 10572-10602.

[2] S. Stojanovic, and K. Donhauser, and F. Yang, *Tight bounds for maximum ℓ_1-margin classifiers*, Proceedings of the International Conference on Algorithmic Learning Theory **2024**, 1055–1112.

[3] K. Donhauser, N. Ruggeri, S. Stojanovic, & F. Yang, *Fast rates for noisy interpolation require rethinking the effect of inductive bias.*, Proceedings of the International Conference on Machine Learning **2022**, 5397–5428.

Pattern Recovery by SLOPE

Tomasz Skalski

(joint work with Małgorzata Bogdan, Xavier Dupuis, Piotr Graczyk, Bartosz Kołodziejek, Patrick Tardivel, Maciej Wilczyński)

Sorted L-One Penalized Estimator (SLOPE), a generalization of the LASSO estimator, was introduced by Bogdan, van den Berg, Sabatti, Su and Candès in 2015. It is a convex regularization method for fitting high-dimensional regression models. While LASSO can eliminate redundant predictors by setting the corresponding regression coefficients to zero, SLOPE can also identify clusters of variables with the same absolute values of regression coefficients.

In this talk I discuss sufficient and necessary conditions for the proper identification of the SLOPE pattern, i.e. of the proper sign and of the proper ranking of the absolute values of individual regression coefficients, including a proper clustering. I also mention the strong consistency of pattern recovery by SLOPE in an asymptotic case when the number of columns in the design matrix is fixed, but the sample size diverges to infinity.

References

[1] M. Bogdan, X. Dupuis, P. Graczyk, B. Kołodziejek, T. Skalski, P. Tardivel, M. Wilczyński, *Pattern Recovery by SLOPE*, ArXiv **2203.12086**.

[2] P. Graczyk, U. Schneider, T. Skalski, P. Tardivel, *A Unified Framework for Pattern Recovery in Penalized and Thresholded Estimation and its Geometry*, ArXiv **2307.10158**.

[3] T. Skalski, P. Graczyk, B. Kołodziejek, M. Wilczyński, *Pattern recovery and signal denoising by SLOPE when the design matrix is orthogonal*, Probability and Mathematical Statistics **42(2)** (2022), 283–302.

A unified framework for pattern recovery in penalized estimation and its geometry

Ulrike Schneider

(joint work with Patrick Tardivel, Tomasz Skalski, Piotr Graczyk)

We consider the framework of penalized estimation where the penalty term is given by a polyhedral norm, or more generally, a polyhedral gauge, which encompasses methods such as LASSO and generalized LASSO, SLOPE, OSCAR, PACS and others. Each of these estimators can uncover a different structure or "pattern" of the unknown parameter vector. We define a novel and general notion of patterns based on subdifferentials and formalize an approach to measure pattern complexity. For pattern recovery, we provide a minimal condition for a particular pattern to be detected with positive probability, the so-called accessibility condition. We make the connection to estimation uniqueness by showing that uniqueness holds if and only if no pattern with complexity exceeding the rank of the X-matrix is accessible. Subsequently, we introduce the noiseless recovery condition which is a stronger requirement than accessibility and which can be shown to play exactly the same role as the well-known irrepresentability condition for the LASSO – in that the probability of pattern recovery is bounded by $1/2$ if the condition is not satisfied. Through this, we unify and extend the irrepresentability condition to a broad class of penalized estimators using an interpretable criterion. We also look at the "gap" between accessibility and the noiseless recovery condition and discuss that our criteria show that it is more pronounced for simple patterns. Finally, we prove that the noiseless recovery condition can indeed be relaxed when turning to so-called thresholded penalized estimation: in this setting, the accessibility condition is already sufficient (and necessary) for sure pattern recovery provided that the signal of the pattern is large enough. We demonstrate how our findings can be interpreted through a geometrical lens throughout the talk and illustrate our results for the Lasso as well as other estimation procedures.

References

[1] K. Ewald and U. Schneider, *Model selection properties and uniqueness of the Lasso estimator in low and high dimensions*, Electronic Journal of Statistics **14** (2020), 944–969.

[2] U. Schneider and P. Tardivel, *The geometry of uniqueness, sparsity and clustering in penalized estimation*, Journal of Machine Learning Research **23** (2022), 1–36.

[3] P. Gracyk, U. Schneider, T. Skalski and P. Tardivel, *A unified framework for pattern recovery in penalized and thresholded estimation and its geometry*, **arxiv:2307.10158** (2023).

Uncertainty quantification for iterative algorithms in linear models

Pierre Bellec

This paper investigates the iterates $\hat{b}^1, \ldots, \hat{b}^T$ obtained from iterative algorithms in high-dimensional linear regression problems, in the regime where the feature dimension p is comparable with the sample size n, i.e., $p \asymp n$. The analysis and proposed estimators are applicable to Gradient Descent (GD), proximal GD and

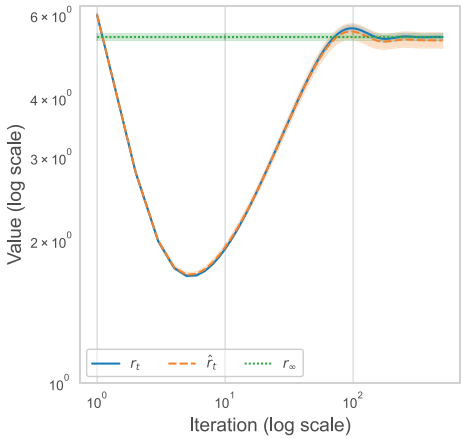

FIGURE 1. Generalization error r_t along the trajectory of accelerated gradient descent and its estimate $\hat{r}_t$, over 100 repetitions

their accelerated variants such as Fast Iterative Soft-Thresholding (FISTA). The paper proposes novel estimators $\hat{r}_t$ for the generalization error r_t of the iterate $\hat{b}^t$ for any fixed iteration t along the trajectory. These estimators are proved to be $\sqrt{n}$-consistent under Gaussian designs. Applications to early-stopping are provided: when the generalization error of the iterates is a U-shape function of the iteration t, the estimates allow to select from the data an iteration $\hat{t}$ that achieves the smallest generalization error along the trajectory. Additionally, we provide a technique for developing debiasing corrections and valid confidence intervals for the components of the true coefficient vector from the iterate $\hat{b}^t$ at any finite iteration t.

Semi-Supervised Sparse Gaussian Classification. Provable Benefit of Unlabeled Data

BOAZ NADLER

The premise of semi-supervised learning (SSL) is that combining labeled and unlabeled data yields significantly more accurate models. Despite empirical successes, the theoretical understanding of SSL is still far from complete. In our work, we study SSL for high dimensional sparse Gaussian classification. To construct an accurate classifier, a key task is feature selection, detecting the few variables that separate the two classes. For this SSL setting, we analyze information-theoretic lower bounds for accurate feature selection as well as computational lower bounds, assuming the low-degree polynomial likelihood hardness conjecture. Our key contribution is the identification of a regime in the problem parameters where SSL is guaranteed to be advantageous for classification. Specifically, there is a

regime where it is possible to construct in polynomial time an accurate SSL classifier. However, any computationally efficient supervised or unsupervised learning schemes, that separately use only the labeled or unlabeled data would fail.

Our work highlights the provable benefits of combining labeled and unlabeled data for classification and feature selection in high dimensions.

Linear Discriminant Analysis and Regularized Regression

Marten Wegkamp
(joint work with Xin Bing, Bingqing Li)

Linear Discriminant Analysis (LDA) is an important classification approach. Its simple linear form makes it easy to interpret and it is capable to handle multi-class responses. It is closely related to other classical multivariate statistical techniques, such as Fisher's discriminant analysis, canonical correlation analysis and linear regression.

In this talk we strengthen its connection to multivariate response regression by characterizing the explicit relationship between the discriminant directions and the regression coefficient matrix. This key characterization leads to a new regression-based multi-class classification procedure that is flexible enough to deploy any existing structured, regularized, and even non-parametric, regression methods. Moreover, our new formulation is amenable to analysis: we establish a general strategy of analyzing the excess misclassification risk of the proposed classifier for all aforementioned regression techniques. We provide complete theoretical guarantees for using the widely used ℓ_1-regularization as well as for using the reduced-rank regression, neither of which has yet been fully analyzed in the LDA context. Our theoretical findings are corroborated by extensive simulation studies and real data analysis.

References

[1] X. Bing, B. Li and M. Wegkamp, *Linear Discriminant Regularized Regression*, arXiv: 2402.14260 (2024).
[2] X. Bing and M. Wegkamp, *Interpolating Discriminant Functions in High-Dimensional Gaussian Latent Mixtures*, Biometrika **111(1)** (2024), 291–308.
[3] X. Bing and M. Wegkamp, *Optimal Discriminant Analysis in High-Dimensional Latent Factor Models*, Annals of Statistics **51(3)** (2023), 1232–1257.

An alternative view of diffusion models

Francis Bach
(joint work with Saeed Saremi, Ji-Won Park)

We introduce a theoretical framework for sampling from unnormalized densities based on a smoothing scheme that uses an isotropic Gaussian kernel with a single fixed noise scale. We prove one can decompose sampling from a density (minimal assumptions made on the density) into a sequence of sampling from log-concave

conditional densities via accumulation of noisy measurements with equal noise levels. Our construction is unique in that it keeps track of a history of samples, making it non-Markovian as a whole, but it is lightweight algorithmically as the history only shows up in the form of a running empirical mean of samples. Our sampling algorithm generalizes walk-jump sampling [1]. The "walk" phase becomes a (non-Markovian) chain of (log-concave) Markov chains. The "jump" from the accumulated measurements is obtained by empirical Bayes. We study our sampling algorithm quantitatively using the 2-Wasserstein metric and compare it with various Langevin MCMC algorithms. We also report a remarkable capacity of our algorithm to "tunnel" between modes of a distribution.

References

[1] S. Saremi, J.-W. Park, F. Bach. Chain of Log-Concave Markov Chains. International Conference on Learning Representations (ICLR), 2024.
[2] Saremi, S., Hyvärinen, A. Neural empirical bayes. Journal of Machine Learning Research, 20(181), 1–23, 2019.

The puzzle of dimensionality and feature learning in neural networks and kernel machines

Misha Belkin

Remarkable progress in AI has far surpassed expectations of just a few years ago. At their core, modern models, such as transformers, implement traditional statistical models – high order Markov chains. Nevertheless, it is not generally possible to estimate Markov models of that order given any possible amount of data. Therefore these methods must implicitly exploit low-dimensional structures present in data. Furthermore, these structures must be reflected in high-dimensional internal parameter spaces of the models. Thus, to build fundamental understanding of modern AI, it is necessary to identify and analyze these latent low-dimensional structures. In this talk, we discuss how deep neural networks of various architectures learn low-dimensional features and how the lessons of deep learning can be incorporated in non-backpropagation-based algorithms that we call Recursive Feature Machines. We provide a number of experimental results on different types of data, as well as some connections to classical sparse learning methods, such as Iteratively Reweighted Least Squares. The discussion is based on papers [1, 2].

References

[1] A Radhakrishnan, D Beaglehole, P Pandit, M Belkin, *Mechanism for feature learning in neural networks and backpropagation-free machine learning models*, Science **383** (2024), 1461-1467.
[2] A. Radhakrishnan, M Belkin, D Drusvyatskiy, *Linear Recursive Feature Machines provably recover low-rank matrices*, arXiv:2401.04553.

ReLU Neural Networks with Linear Layers are Biased Towards Single- and Multi-Index Models

Rebecca Willett

(joint work with Suzanna Parkinson, Greg Ongie)

Neural networks often operate in the overparameterized regime, in which there are far more parameters than training samples, allowing the training data to be fit perfectly. That is, training the network effectively learns an interpolating function, and properties of the interpolant affect predictions the network will make on new samples. This manuscript explores how properties of such functions learned by neural networks of depth greater than two layers. Our framework considers a family of networks of varying depths that all have the same capacity but different representation costs. The representation cost of a function induced by a neural network architecture is the minimum sum of squared weights needed for the network to represent the function; it reflects the function space bias associated with the architecture. Our results show that adding additional linear layers to the input side of a shallow ReLU network yields a representation cost favoring functions with low mixed variation - that is, it has limited variation in directions orthogonal to a low-dimensional subspace and can be well approximated by a single- or multi-index model. Such functions may be represented by the composition of a function with low two-layer representation cost and a low-rank linear operator. Our experiments confirm this behavior in standard network training regimes. They additionally show that linear layers can improve generalization and the learned network is well-aligned with the true latent low-dimensional linear subspace when data is generated using a multi-index model.

References

[1] S. Parkinson, G. Ongie, and R. Willett, *ReLU Neural Networks with Linear Layers are Biased Towards Single- and Multi-Index Models*, arXiv:2305.15598 (2024).

Emergence of clusters in self-attention dynamics

Philippe Rigollet

Our goal in this talk was to study a simple model for the self-attention mechanism, which is the main innovation behind the transformer architecture in deep learning. We study dynamics of the form:

$$\dot{x}(t) = P_{x_i(t)} \frac{\sum_{j=1}^n x_j(t) e^{\beta \langle x_j(t), x_i(t) \rangle}}{\sum_{j=1}^n e^{\beta \langle x_j(t), x_i(t) \rangle}}, \quad i = 1, \dots n,$$

or the unnormalized version

$$\dot{x}(t) = P_{x_i(t)} \frac{1}{n} \sum_{j=1}^n x_j(t) e^{\beta \langle x_j(t), x_i(t) \rangle}, \quad i = 1, \dots n.$$

Here, P_x denotes the projection onto the tangent space of the unit sphere $\mathbb{S}^{d-1}$ at $x \in \mathbb{S}^{d-1}$.

$$P_x z = z - \langle z, x \rangle x.$$

The above-mentioned dynamics describe the dynamics of the particles on the sphere, and we study their asymptotic configuration.

First, it can be shown that unnormalized dynamics converge to a single cluster, reminiscent of synchronization of Kuramoto oscillators, $\forall \beta \geq 0$ and $d \geq 3$ (the $d = 2$ question is still open today).

Then, we studied *metastable* states, where particles are tightly clustered into $k \geq 2$ clusters. These states are reached in time $\propto \beta$, and particles get stuck there for time $\propto e^{c\beta}$.

Finally, we showed that $k = \tilde{O}(\sqrt{\beta})$ by studying the number of modes of a kernel density estimator.

Towards Plurality: Foundations for learning from Diverse Human Preferences

Ramya Korlakai Vinayak

(joint work with Greg Canal, Blake Mason, Zhi Wang, Geelon So, Daiwei Chen, Yi Chen, Aniket Rege, Rob Nowak)

Large pre-trained models trained on internet-scale data are often not ready for safe deployment out-of-the-box. They are heavily fine-tuned and *aligned* using large quantities of human preference data, usually elicited using pairwise comparisons of outputs for a given input context. While aligning these artificial intelligent and/or machine learning models to human preferences, it is important to ask whose preferences are we aligning them to? The current approaches used for alignment are severely limited due to their inherent assumption uniformity of preferences and the need for *plurality*, i.e., capturing the diverse or heterogeneous human preferences, is getting recognised as an important challenge to address in this arena. We aim to overcome the limitations of current approaches by building mathematical foundations for learning from heterogeneous human preferences.

In this talk, I discuss a series of recent results with my collaborators that focus on how we can reliably capture diverse preferences while pooling together data from a large number of individuals in a given population. In the first part, we focus on fundamental questions pertaining to simultaneous metric and preference learning from pairwise comparisons [1]. Under the ideal point model [2], we characterize the sufficient conditions for identifiability and provide sample complexity bounds for learning an shared unknown Mahalanobis metric and different unknown preferences of individuals from pairwise comparisons. In particular, in $\mathbb{R}^d$, we show that if we have $\Omega(d)$ individuals, then with $\tilde{\mathcal{O}}(d)$ comparison queries per individual, we can simultaneously learn the unknown metric as well as the individual user preferences. We note that even when the metric is known, $\tilde{\Omega}(d)$ queries per individual are necessary to learn their prefrence point in $\mathbb{R}^d$. We then address the question

of whether it is possible to learn the unknown metric without having to localize user preference [3] and show a general impossibility result with $o(d)$ queries per individual. We then consider the setting with structure in the data, particularly, union of low-dimisional subspaces, and provide a divide-and-conquer approach for learning the mertric with $o(d)$ queries per individual. Using the insights gained from these theoretical understanding, we then propose a practical framework for pluralistic alignment using preference queries [4]. We apply our framework on both vision and language generation tasks and show that we can obtain state-of-the-art results with simple 2-layer multi-layer perceptron (MLP) learned on top of the pre-trained models compared to previous approaches that need fine-tuning of very large models.

References

[1] G. Canal, B. Mason, R. K. Vinayak, and R. Nowak. *One for All: Simultaneous Metric and Preference Learning over Multiple Users*, In Proceedings of Neural Information Processing Systems (NeurIPS). 2022.

[2] C. H. Coombs. *Psychological scaling without a unit of measurement*, In Psychological review, 57(3):145, 1950.

[3] Z. Wang, G. So, and R. K. Vinayak. *Metric learning from limited pairwise preference comparisons*, In Proceedings of Uncertainty in Aritificial Intelligence (UAI). 2024.

[4] D. Chen, Y. Chen, A. Rege, and R. K. Vinayak. *PAL: Pluralistic Alignment Framework for Learning from Heterogeneous Preferences*, arxiv pre-print. 2024.

Learning large softmax mixtures

Jonathan Niles-Weed

(joint work with Xin Bing, Florentina Bunea, Martin Wegkamp)

We study the softmax mixture model, which has come to occupy an important role in machine learning, discrete choice theory, and text analysis. In this model, observations are drawn according to a discrete mixture given by convex combination of vectors of the form

$$A_{\boldsymbol{\theta}}(x_j \mid \mathbf{x}_p) =: \mathsf{softmax}(x_j^\top \boldsymbol{\theta}) = \frac{\exp(x_j^\top \boldsymbol{\theta})}{\sum_{i=1}^p \exp(x_i^\top \boldsymbol{\theta})}, \qquad \text{for each } j \in [p]\,,$$

where $\mathbf{x}_p = \{x_1, \ldots, x_p\}$ denotes a collection of observed "embedding vectors."

Explicitly, we work in the setting where $x_1, \ldots, x_p$ are the observed values of a collection $X_1, \ldots, X_p$ of i.i.d. $\mathcal{N}(0, \Sigma)$ vectors. We view the data as arising from a discrete mixture on $\mathbf{x}_p$, with masses given by

$$\pi_\omega(x_j \mid \mathbf{x}_p) =: \sum_{k=1}^K \alpha_k A_{\theta_k}(x_j \mid \mathbf{x}_p), \qquad \text{for each } j \in [p]\,.$$

The goal is to estimate the parameters $\omega := (\theta_1, \ldots, \theta_K, \alpha)$.

We study:

(1) Method of Moments (MoMMS) parameter estimation in softmax mixtures,

(2) EM-based parameter estimation with MoMMS warm-start in softmax mixtures.

We show that the MoMMS procedure as a warm start followed by our EM-estimator provably recovers the parameters at a nearly minimax rate in polynomial time.

Probably Approximately Correct Constrained Learning

Luiz F. O. Chamon

(joint work with Miguel Calvo-Fullana, Santiago Paternain, Alejandro Ribeiro)

Requirements are integral to systems that are always defined as compromises between competing specifications such as accuracy, robustness, safety, and efficiency. As data plays an increasingly central role in systems design, requirements have also become of growing interest in machine learning (ML). Learning to satisfy requirements is, however, antithetical to the standard ML practice of minimizing individual losses. Constrained learning overcomes this challenge by incorporating requirements as statistical constraints rather than modifying the training objective. Explicitly, constrained learning is defined as

$$\text{(P-CSL)} \qquad \begin{aligned} P^\star = \min_{\theta\in\Theta} \quad & \mathbb{E}_{(x,y)\sim\mathfrak{D}_0}\Big[\ell_0\big(f_\theta(x),y\big)\Big] \\ \text{subject to} \quad & \mathbb{E}_{(x,y)\sim\mathfrak{D}_i}\Big[\ell_i\big(f_\theta(x),y\big)\Big] \leq c_i, \quad i=1,\dots,m, \end{aligned}$$

where $\mathfrak{D}_i$, $i=0,\dots,m$, denote probability distributions over data pairs $(x,y)\in\mathbb{R}^d\times\mathbb{R}$; the $\ell_i:\mathbb{R}^k\times\mathcal{Y}\to[0,B]$, $i=0,\dots,m$, together with the c_i, encode the performance metric and the desired statistical properties of the solution; and $f_\theta:\mathbb{R}^d\to\mathbb{R}^k$ is a function associated with the parameter vector $\theta\in\Theta\subseteq\mathbb{R}^p$. Observe that (P-CSL) considers statistical constraints rather than parameter constraints, such as quadratic requlatization ($\|\theta\|_2\leq c$) or sparsity ($\|\theta\|_1\leq c$). The latter, embedded in Θ, are deterministic and can be directly imposed using projections [1, 2].

In this talk, I develop a theory of constrained learning that establishes when and how (P-CSL) can be solved using only samples from the $\mathfrak{D}_i$. This is akin to what classical learning theory does for unconstrained learning problems (i.e., (P-CSL) with $m=0$). Our main result shows that the solution of (P-CSL) can be approximated by solving the empirical dual problem

$$\hat{D}^\star = \max_{\mu_i\geq 0}\ \min_{\theta\in\Theta}\ \frac{1}{N_0}\sum_{n_0=1}^{N_0}\ell_0\big(f_\theta(x_{n_0}),y_{n_0}\big) + \sum_{i=1}^{m}\mu_i\left[\frac{1}{N_i}\sum_{n_i=1}^{N_i}\ell_i\big(f_\theta(x_{n_i}),y_{n_i}\big) - c_i\right],$$

which uses on N_i samples $(x_{n_i},y_{n_i})\sim\mathfrak{D}_i$. Indeed, assume that the ℓ_i are convex and M-Lipschitz continuous functions and that the hypothesis class is probably approximately correct (PAC) learnable with respect to each ℓ_i and is ν-*universal*, i.e., there exists $\nu\geq 0$ and a (convex) function space $\mathcal{H}$ such that for each $\phi\in\mathcal{H}$ there exists $\theta\in\Theta$ such that $\mathbb{E}_{\mathfrak{D}_i}[|\phi(x)-f_\theta(x)|]\leq\nu$. Then, we show that $|P^\star-\hat{D}^\star|$

is bounded with high probability over draws of the samples [1, 2]. This result can be extended to non-convex ℓ_i using techniques developed to tackle sparsity on the continuum [3]. Additionally, if the ℓ_i are *smooth*, i.e., have M-Lipschitz gradients, and ℓ_0 is strongly convex, we can provide a bound on the infeasibility of the solutions, namely $|\mathbb{E}_{\mathfrak{D}_i}\big[\ell_i(f_{\hat{\theta}^\star}(x),y)-\ell_i(f_{\theta^\star}(x),y)\big]|$, where $\theta^\star$ is a solution of (P-CSL) and $\hat{\theta}^\star$ achieves $\hat{D}^\star$ [4]. This result enables a practical constrained learning rule that uses dual ascent methods to tackle $\hat{D}^\star$. Hence, under mild conditions, it is possible to tackle constrained learning tasks by solving only unconstrained empirical risk minimization (ERM) problems [1, 2, 4].

These advances can be used to directly tackle challenging problems in robust learning [1, 5], fair learning [1, 2], learning under invariance [6], and semi-supervised learning [7]. These contributions suggest how we can go beyond the current objective-centric learning paradigm towards a constraint-driven learning one.

References

[1] L. F. O. Chamon and A. Ribeiro. *Probably approximately correct constrained learning*, Conference on Neural Information Processing Systems (2020).

[2] L. F. O. Chamon, S. Paternain, M. Calvo-Fullana, and A. Ribeiro. *Constrained learning with non-convex losses*, IEEE Trans. on Inf. Theory **69[3]** (2023), 1739–1760.

[3] L. F. O. Chamon, Y. C. Eldar, and A. Ribeiro. *Functional nonlinear sparse models*, IEEE Trans. on Signal Process. **68[1]** (2020), 2449–2463.

[4] J. Elenter, L. F. O. Chamon, and A. Ribeiro. *Near-optimal solutions of constrained learning problems*, International Conference on Learning Representations (2024).

[5] A. Robey*, L. F. O. Chamon*, G. J. Pappas, H. Hassani, and A. Ribeiro. *Adversarial robustness with semi-infinite constrained learning*, Conference on Neural Information Processing Systems (2021). (* equal contribution)

[6] I. Hounie, L. F. O. Chamon, and A. Ribeiro. *Automatic data augmentation via invariance-constrained learning*, International Conference on Machine Learning (2023).

[7] J. Cervino, L. F. O. Chamon, B. D. Haeffele, R. Vidal, and A. Ribeiro. *Learning globally smooth functions on manifolds*, International Conference on Machine Learning (2023).

Nonparametric bootstrap of high-dimensional sample covariance matrices

Angelika Rohde

(joint work with Holger Dette)

Let $Y_1,\dots,Y_n$ be iid p-dimensional centered random vectors with covariance matrix Σ_n and corresponding sample covariance matrix $\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n Y_iY_i^\top$. We denote by $\hat{\lambda}_1,\dots,\hat{\lambda}_p$ its eigenvalues, by $\mu^{\hat{\Sigma}_n} = \frac{1}{p}\sum_{i=1}^p \delta_{\hat{\lambda}_i}$ its (normalized) spectral measure and by $m_{\hat{\Sigma}_n}$ its Stieltjes transform. Our goal is to provide a fully nonparametric and computationally tractable tool to obtain accurate approximations for the distribution of particular eigenvalue statistics of the sample covariance matrix in the high-dimensional context where the dimension grows proportionately with the sample size.

Model assumptions. Aligning with the common framework in random matrix theory, we shall work under the same type of conditions and study a triangular array of $p = p(n)$-dimensional observations $Y_1, \ldots, Y_n$ of the form

$$Y_i = A_n X_i, \quad i = 1, \ldots, n. \tag{1}$$

Here, $X_i = (X_{i1}, X_{i2}, \ldots)^\top$ $(i \in \mathbb{N})$ are iid infinite dimensional random vectors and A_n is a $p \times \infty$ matrix such that the following assumptions are satisfied:

(A1) The matrix A_n has square summable rows and $\sup_{n\in\mathbb{N}} \|A_n\|_{S_\infty} < \infty$.
(A2) $p/n \to c$ for some real constant $c > 0$ as $n \to \infty$.
(A3) The vector X_1 has iid entries X_{1k}, $k \in \mathbb{N}$, with $\mathbb{E}X_{11} = 0$ and $\mathbb{E}X_{11}^2 = 1$.

Under these conditions, $Y_1 = A_n X_1$ is well defined as limit in $L^2(\mathbb{P})$ with covariance matrix

$$\Sigma_n = \mathbb{E}[Y_1 Y_1^\top] = A_n A_n^\top \ .$$

As concerns normal approximation of linear spectral statistics, the existence of the fourth moment $\mathbb{E}X_{11}^4 < \infty$ is known to be necessary. Therefore, we shall impose in that case the stronger assumption

(A3+) In addition to assumption (A3), $\mathbb{E}X_{11}^3 = 0$ and $\mathbb{E}X_{11}^4 = 3$.

Coincidence of the third and fourth moment with those of the standard normal distribution can be avoided in the CLT for linear spectral statistics of high-dimensional covariance matrices at the expense of additional regularity assumptions on the eigenvectors, see [1]. We refrain from this generalization to keep the technical expenditure as small as possible.

Results of [2, 3] indicate that the classical bootstrap for the LSD is untrustworthy when the problem is genuinely high-dimensional. In Theorem S2.2 in the supplementary material of their paper, [2] showed that the limiting spectral distribution (LSD) of the bootstrapped covariance matrix is completely different from that of $\widehat{\Sigma}_n$. The traditionally in a wider range applicable m out of n bootstrap does not even preserve the limiting ratio c of dimension and sample size if $m \ll n$, which appears already explicitly in the characterizing Marčenko-Pastur equation for the Stieltjes transform of the LSD.

Condition 1 (Representative Subpopulation Condition)**.** *The triangular array of p-dimensional vectors $Y_1, \ldots, Y_n$ in model* (1) *is said to satisfy the Representative Subpopulation Condition, if the following conditions are satisfied.*

(1) For every $q \le p$, there exists a possibly random strategy (independent of $Y_1, \ldots, Y_n$) of selecting q out of p coordinates such that the covariance matrix $\tilde{\Sigma}_n$ of the resulting q-dimensional subvectors $Y_{i,sub}$ $(i = 1, \ldots, n)$ satisfies

$$\mu^{\tilde{\Sigma}_n} - \mu^{\Sigma_n} \Rightarrow 0 \quad \text{as } q, n \to \infty \quad \text{in probability.} \tag{2}$$

(2) If $\Pi_n = \Pi_{n,q}$ denotes the projection corresponding to the possibly random selection strategy, that is $Y_{i,sub} = \Pi_n Y_i$ $(i = 1, \ldots, n)$, then there exists for almost all realizations a decomposition of the form

$$\Pi_n A_n = L_n + R_n \ , \tag{3}$$

where the sets of non-zero entries of the matrices L_n and R_n are disjoint, the matrix L_n has a most q' non-zero columns with q' a possibly random integer of deterministic order $O(q)$, and $\mathbb{E}_{\Pi_n}\left[||R_n||_{S_2}^2\right] = o(1)$ as $q, n \to \infty$. Here, $\mathbb{E}_{\Pi_n}$ denotes expectation with respect to the random projection Π_n.

The Representative Subpopulation Condition being granted, we propose the following resampling scheme.

Algorithm 2 ($(m, mp/n)$ out of (n, p) Bootstrap)**.**

(i) For $m = o(n)$, draw an iid sample $Y_1^*, \ldots, Y_m^*$ from the empirical distribution $\hat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$.

(ii) Define the bootstrap sample

$$Z_i^* = \Pi_n Y_i^* = (Y_{ij_1}^*, \ldots, Y_{ij_q}^*)^\top, \quad i = 1, \ldots, m,$$

using the $q = \lfloor mp/n \rfloor$ coordinates $j_1, \ldots, j_q$ selected according to the Representative Subpopulation Condition.

(ii) Output: the estimator

$$\widehat{\Sigma}_n^* = \frac{1}{m}\sum_{i=1}^m Z_i^* Z_i^{*\top} = \Pi_n\Big(\frac{1}{m}\sum_{i=1}^m Y_i^* Y_i^{*\top}\Big)\Pi_n^\top$$

and its corresponding spectral distribution $\mu^{\widehat{\Sigma}_n^*}$.

Our first result demonstrates that $\widehat{\Sigma}_n^*$ mimics the sample covariance matrix in terms of spectral distributions. Besides being of interest in its own, this is a necessary ingredient for the CLT for linear spectral statistics as the limiting spectral distribution of the $\widehat{\Sigma}_n$ explicitly enters the limiting variance expression of linear spectral statistics.

Theorem 3 (Spectral distribution)**.** *Grant assumptions (A1)–(A3). Suppose that the triangular array of p-dimensional vectors $Y_1, \ldots, Y_n$ in model* (1) *satisfies the Representative Subpopulation Condition 1. If $m = o(n)$, then*

$$\mu^{\widehat{\Sigma}_n} - \mu^{\widehat{\Sigma}_n^*} \Longrightarrow 0 \quad \textit{in probability.}$$

A further important step in the proof of the CLT for linear spectral statistics is the following result.

Theorem 4. *Grant assumptions (A1)–(A3) and $\mathbb{E}X_{11}^4 < \infty$. Suppose that the triangular array of p-dimensional vectors $Y_1, \ldots, Y_n$ in model* (1) *satisfies Condition 1. Let $c' = \limsup(q'/m)$.*

(a) If $m = o(\sqrt{n})$, then there exists a constant $K_r > 0$ such that

$$\mathbb{P}\Big(\big\|\widehat{\Sigma}_n^*\big\|_{S_\infty} > K_r\Big) = o(m^{-l}) \quad \textit{for every } l \in \mathbb{N}. \tag{4}$$

If $m = o(\log n)$, then (4) *holds even for every $K_r > \limsup_{n\in\mathbb{N}} \|\Sigma_n\|_{S_\infty}(1+\sqrt{c'})^2$.*

(b) If $m = o(\sqrt{n})$, then we have for any $K_l < \liminf_{n\in\mathbb{N}} \|\Sigma_n\|_{S_\infty} \max\{(1 - \sqrt{c'})^2, 0\}$

$$\mathbb{P}\Big(\lambda_{\min}\big(\widehat{\Sigma}_n^*\big) < K_l\Big) = o(m^{-l}) \quad \textit{for every } l \in \mathbb{N}\ .$$

Finally, we study linear spectral statistics

$$\hat{T}_n^*(f) = \sum_{j=1}^{q} f\big(\hat{\lambda}_j^*\big) = q \int f(x) d\mu^{\widehat{\Sigma}_n^*}(x), \tag{5}$$

where $\hat{\lambda}_1^*, \ldots, \hat{\lambda}_q^*$ denote the eigenvalues of the matrix $\widehat{\Sigma}_n^*$. We set f equal to 0 outside its domain and $f_m := fI\{|f| \le m^\ell\}$ for some arbitrary $\ell \in \mathbb{N}$. This definition ensures the existence of $\mathbb{E}\hat{T}_n^*(f)$ for functions that grow faster than any polynomial.

Theorem 5 (Linear spectral statistics)**.** *Grant assumptions (A1)–(A3+) and suppose that the triangular array of p-dimensional vectors $Y_1, \ldots, Y_n$ in model (1) satisfies Condition 1. Let f be a real-valued function which is analytic in a region of the complex plane containing the interval*

$$I = \big[K_l - \limsup_{n\to\infty} \|\Sigma_n\|_{S_\infty}, K_r + \limsup_{n\to\infty} \|\Sigma_n\|_{S_\infty}\big], \tag{6}$$

where K_l and K_r are the constant in Theorem 4. Furthermore, assume that $m = o(\sqrt{n})$. If $\mathbb{E}|X_{11}|^6 < \infty$, then

$$d_{BL}\Big[\mathcal{L}\big(\hat{T}_n^*(f) - \mathbb{E}^*\hat{T}_n^*(f_m)\big) \,\big|\, Y_1, \ldots, Y_n\big), \mathcal{L}\big(\hat{T}_n(f) - \mathbb{E}\hat{T}_n(f_n)\big)\Big] \longrightarrow_{\mathbb{P}} 0, \tag{7}$$

where d_{BL} denotes the dual bounded Lipschitz metric. Moreover, the conditional distribution of $\hat{T}_n^*(f) - \mathbb{E}^*\hat{T}_n^*(f_m)$ is asymptotically centered normal with variance

$$-\frac{1}{2\pi^2} \oint\oint \frac{f(z_1)f(z_2)}{(\underline{m}_{\mu^{\widehat{\Sigma}_n}}(z_1) - \underline{m}_{\mu^{\widehat{\Sigma}_n}}(z_2))^2} \underline{m}'_{\mu^{\widehat{\Sigma}_n}}(z_1)\underline{m}'_{\mu^{\widehat{\Sigma}_n}}(z_2) dz_1 dz_2 + o_{\mathbb{P}}(1).$$

References

[1] J. Najim and J. Yao, *Gaussian fluctuations for linear spectral statistics of large random covariance matrices*, The Annals of Applied Probability, **26** (2016),1837–1887.

[2] N. El Karoui and E. Purdom, *The bootstrap, covariance matrices and PCA in moderate and high- dimensions*, arXiv:1608.00948.

[3] N. El Karoui and E. Purdom, *The non-parametric bootstrap and spectral analysis in moderate and high- dimension*, in Chaudhuri, K. and Sugiyama, M., editors, Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, volume 89 of Proceedings of Machine Learning Research, (2019), 2115–2124. PMLR.

Estimation of barycenters in metric spaces

Victor-Emmanuel Brunel
(joint work with Jordan Serres)

In metric spaces that lack a linear structure, barycenters provide a canonical extension of linear averaging. In this talk, we are interested in the problem of estimating the barycenter of a distribution, given iid data. We work under a geometric assumption on the underlying space, ensuring that barycenters are defined as solutions to (geodesically) convex optimization problems and we present statistical guarantees for several estimators, some of which that can be computed efficiently from streamed data.

Stability and statistical inference for semidiscrete optimal transport maps

Kengo Kato
(joint work with Ritwik Sadhu, Ziv Goldfeld)

We study statistical inference for the optimal transport (OT) map (also known as the Brenier map) from a known absolutely continuous reference distribution onto an unknown finitely discrete target distribution. We derive limit distributions for the L^p-error with arbitrary $p \in [1, \infty)$ and for linear functionals of the empirical OT map, together with their moment convergence. The former has a non-Gaussian limit, whose explicit density is derived, while the latter attains asymptotic normality. For both cases, we also establish consistency of the nonparametric bootstrap. The derivation of our limit theorems relies on new stability estimates of functionals of the OT map with respect to the dual potential vector, which may be of independent interest. We also discuss applications of our limit theorems to the construction of confidence sets for the OT map and inference for a maximum tail correlation. Finally, we show that, while the empirical OT map does not possess nontrivial weak limits in the L^2 space, it satisfies a central limit theorem in a dual Hölder space, and the Gaussian limit law attains the asymptotic efficiency bound.

Empirical partially Bayes multiple testing and compound χ^2 decisions

Bodhisattva Sen
(joint work with Nikolaos Ignatiadis)

A common task in high-throughput biology is to screen for associations across thousands of units of interest, e.g., genes or proteins. Often, the data for each unit are modeled as Gaussian measurements with unknown mean and variance and are summarized as per-unit sample averages and sample variances. The downstream goal is multiple testing for the means. In this domain, it is routine to "moderate" (that is, to shrink) the sample variances through parametric empirical Bayes methods before computing p-values for the means. Such an approach is asymmetric in that a prior is posited and estimated for the nuisance parameters (variances)

but not the primary parameters (means). Our work initiates the formal study of this paradigm, which we term "empirical partially Bayes multiple testing". In this framework, if the prior for the variances were known, one could proceed by computing p-values conditional on the sample variances–a strategy called partially Bayes inference by Sir David Cox. We show that these conditional p-values satisfy an Eddington/Tweedie-type formula and are approximated at nearly-parametric rates when the prior is estimated by nonparametric maximum likelihood. The estimated p-values can be used with the Benjamini-Hochberg procedure to guarantee asymptotic control of the false discovery rate. Even in the compound setting, wherein the variances are fixed, the approach retains asymptotic type-I error guarantees.

Conformal PID control for time series prediction

Ryan J. Tibshirani

(joint work with Anastasios N. Angelopoulos, Emmanuel J. Candès)

Machine learning models run in production systems regularly encounter data distributions that change over time. This can be due to factors such as seasonality and time-of-day, continual updating and re-training of upstream machine learning models, changing user behaviors, and so on. These distribution shifts can degrade a model's predictive performance. They also invalidate standard techniques for uncertainty quantification, such as *conformal prediction* [1].

To address the problem of shifting distributions, we consider a (possibly) adversarial time series of deterministic covariates $x_t \in \mathcal{X}$ and responses $y_t \in \mathcal{Y}$, for $t \in \mathbb{N} = \{1, 2, 3, \dots\}$. As in standard conformal prediction, we are free to define any *conformal score function* $s_t : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which we can view as measuring the accuracy of our forecast at time t. We will assume with a loss of generality that s_t is negatively oriented (lower values mean greater forecast accuracy). For example, we may use the absolute error $s_t(x, y) = |y - f_t(x)|$, where f_t is a forecaster trained on data up to but not including data at time t.

The challenge in the sequential setting is as follows. We seek to invert the score function to construct a *conformal prediction set*,

$$\mathcal{C}_t = \{y \in \mathcal{Y} : s_t(x_t, y) \leq q_t\}, \tag{1}$$

where q_t is an estimated $1 - \alpha$ quantile for the distribution of the score $s_t(x_t, y_t)$ at time t. Recall, in standard conformal prediction, we would take q_t to be a level $1 - \alpha$ sample quantile (up to a finite-sample correction) of $s_t(x_i, y_i)$, $i < t$; if the data sequence (x_i, y_i), $i \in \mathbb{N}$ were i.i.d. or exchangeable, then this would yield $1 - \alpha$ coverage [1] at each time t. However, in the sequential setting, which does not assume exchangeability (or any probabilistic model for the data for that matter), choosing q_t in (1) to yield coverage is a formidable task. If we are not willing to make any assumptions about the sequence, then a coverage guarantee at time t would only be possible with trivial methods, which construct prediction intervals of infinite sizes.

Therefore, our goal is to achieve *long-run coverage* in time. That is, letting $\text{err}_t = 1\{y_t \notin \mathcal{C}_t\}$, we would like to achieve, for large integers T,

$$\frac{1}{T}\sum_{t=1}^{T} \text{err}_t = \alpha + o(1) \tag{2}$$

under few or no assumptions, where $o(1)$ denotes a quantity that tends to zero as $T \to \infty$. Furthermore, going beyond (2), we also seek to design flexible strategies to produce the sharpest prediction sets possible, which not only adapt to, but also *anticipate* distribution shifts.

We call our proposed solution *conformal PID control*. It treats the system for producing prediction sets as a proportional-integral-derivative (PID) controller. In the language of control, the prediction sets take a *process variable*, q_t, and then produce an output, err_t. We seek to anchor err_t to a *set point*, α. To do so, we apply corrections to q_t based on the error of the output, $g_t = \text{err}_t - \alpha$. By reframing the problem in this language, we are able to build algorithms that have more stable coverage while also prospectively adapting to changes in the score sequence, much in the same style as a control system.

Three design principles underlie our methods:

(1) *Quantile tracking (P control).* Running online gradient descent on the quantile loss (summed over all past scores) gives rise to a method that we call *quantile tracking*, which achieves long-run coverage (2) under no assumptions except boundedness of the scores. This bound can be unknown. Unlike adaptive conformal inference (ACI) [2], quantile tracking does not return infinite sets after a sequence of miscoverage events. This can be seen as equivalent to proportional (P) control.

(2) *Error integration (I control).* By incorporating the running sum $\sum_{i=1}^{t} g_t$ of the coverage errors into the online quantile updates, we can further stabilize the coverage. This *error integration* scheme achieves long-run coverage (2) under no assumptions whatsoever on the scores (they can be unbounded). This can be seen as equivalent to integral (I) control.

(3) *Scorecasting (D control).* To account for systematic trends in the scores—this may be due to aspects of the data distribution, fixed or changing, which are not captured by the initial forecaster—we train a second model, namely, a *scorecaster*, to predict the quantile of the next score. While quantile tracking and error integration are merely reactive, scorecasting is forward-looking. It can potentially residualize out systematic trends in the errors and lead to practical advantages in terms of coverage and efficiency (set sizes). This can be seen as equivalent to derivative (D) control.

These three modules combine to make our final iteration, the *conformal PID controller*:

$$q_{t+1} = \underbrace{\eta g_t}_{\text{P}} + \underbrace{r_t\left(\sum_{i=1}^{t} g_t\right)}_{\text{I}} + \underbrace{g'_t}_{\text{D}} . \tag{3}$$

In traditional PID control, one would take $r_t(x)$ to be a linear function of x. Here, we allow for nonlinearity and take r_t to be a *saturation function* obeying

$$x \geq c \cdot h(t) \implies r_t(x) \geq b, \quad \text{and} \quad x \leq -c \cdot h(t) \implies r_t(x) \leq -b, \tag{4}$$

for constants $b, c > 0$, and a sublinear, nonnegative, nondecreasing function h—we call a function h satisfying these conditions *admissible*. An example is the *tangent integrator* $r_t(x) = K_{\mathrm{I}} \tan(x \log(t)/(tC_{\mathrm{sat}}))$, where we set $\tan(x) = \mathrm{sign}(x) \cdot \infty$ for $x \notin [-\pi/2, \pi/2]$, and $C_{\mathrm{sat}}, K_{\mathrm{I}} > 0$ are constants. The choice of integrator r_t is a design decision for the user, as is the choice of scorecaster g'_t.

We find it convenient to reparametrize (3), to produce a sequence of quantile estimates q_t, $t \in \mathbb{N}$ used in the prediction sets (1), as follows:

$$\begin{gathered} \text{let } \hat{q}_{t+1} \text{ be any function of the past: } x_i, y_i, q_i, \text{ for } i \leq t, \\ \text{then update } q_{t+1} = \hat{q}_{t+1} + r_t\left(\sum_{i=1}^{t}(\mathrm{err}_i - \alpha)\right). \end{gathered} \tag{5}$$

Taking $\hat{q}_{t+1} = \eta g_t + g'_t$ recovers (3), Now we view $\hat{q}_{t+1}$ as the scorecaster, which directly predicts q_{t+1} using past data. Our main result [3] is that the conformal PID controller (5) achieves long-run coverage for any choice of integrator r_t that satisfies the appropriate saturation condition, and any scorecaster $\hat{q}_{t+1}$.

Theorem 6. *Let $\{\hat{q}_t\}_{t\in\mathbb{N}}$ be any sequence of numbers in $[-b/2, b/2]$ and let $\{s_t\}_{t\in\mathbb{N}}$ be any sequence of score functions with outputs in $[-b/2, b/2]$. Here $b > 0$, and may be infinite. Assume that r_t satisfies* (4)*, for an admissible function h. Then the iteration* (5) *achieves long-run coverage* (2)*.*

References

[1] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*, Springer (2005).

[2] I. Gibbs and E. Candès, *Adaptive conformal inference under distribution shift*, Advances in Neural Information Processing Systems (2021).

[3] A. N. Angelopoulos, E. J. Candès, and R. J. Tibshirani, *Conformal PID control for time series prediction*, Advances in Neural Information Processing Systems (2023).

Adaptive Split Balancing for Optimal Random Forest

Jelena Bradic

(joint work with Yuqian Zhang, Weijie Ji)

Random forests are widely used for regression problems, but existing methods often lack adaptability in complex scenarios and fail to maintain optimality in simple, smooth cases. In this study, we introduce the adaptive split balancing forest (ASBF), a novel approach that learns tree representations from data while achieving minimax optimality under the Lipschitz class. To further enhance performance, we propose a localized version that attains the minimax rate under the Hölder class $\mathcal{H}^{q,\beta}$ for any $q \in \mathbb{N}$ and $\beta \in (0, 1]$.

Our contributions are threefold: 1) While many methods achieve minimax rates of estimation, uniform minimax optimality for all Hölder classes $\mathcal{H}^{q,\beta}$ for any $q \in \mathbb{N}$ and $\beta \in (0,1]$ has remained elusive for random forests, except in purely random forests. The ASBF effectively addresses this gap. 2) We demonstrate that excessive randomness in selecting the splitting variable can negatively impact estimation rates and the model's ability to adapt to the function's underlying smoothness. To mitigate this, each time a leaf is split, we only randomly select a direction from one of the sides that has been split the least times. In other words, the splitting directions are chosen in a balanced fashion – we have to split once in each direction before proceeding to the next round. This approach helps reduce the impact of auxiliary randomness and enables more efficient splitting, enhancing both estimation accuracy and adaptability. 3) Our primary motivation is to justify and enable the use of random forests for computing root-n confidence sets for average treatment effects, representing a significant advancement in the practical application of random forests.

Additionally, through extensive simulation studies and real-data applications, we demonstrate the superior empirical performance of the proposed methods compared to existing random forest techniques.

References

[1] Gérard Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1):1063–1095, 2012.
[2] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.
[3] Torsten Hothorn and Achim Zeileis. Predictive distribution modeling using transformation forests. *Journal of Computational and Graphical Statistics*, 30(4):1181–1196, 2021.
[4] Jason Klusowski. Sharp analysis of a simple model for random forests. In *International Conference on Artificial Intelligence and Statistics*, pages 757–765. PMLR, 2021.

Quadratic functional estimation from observations with multiplicative measurement error

Bianca Neubert
(joint work with Fabienne Comte, Jan Johannes)

We consider a multiplicative deconvolution problem, in which a quadratic functional of the density of a strictly positive random variable X is estimated nonparametrically based on an iid. sample from a noisy observation $Y = X \cdot U$ of X. The multiplicative measurement error U is supposed to be independent of X. The objective of this work is to construct a fully data-driven estimation procedure of quadratic functionals of the density of X when the error density is known. The proposed estimation procedure is based on the estimation of the Mellin transformation of the density. The main issue addressed in this work is the data-driven choice of the cut-off parameter using an approach in the spirit of Goldenshluger and Lepski. We discuss conditions under which the fully data-driven estimator can attain the oracle-risk up to a logarithmic deterioration. We compute convergence rates under classical smoothness assumptions.

Central Limit Theorems for Smooth Optimal Transport Maps

Tudor Manole

(joint work with Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman)

One of the central objects in the optimal transport framework is the quadratic optimal transport map: the unique monotone transformation which pushes forward an absolutely continuous probability law onto any other given law. Several recent works have analyzed the L^2 risk of plugin estimators of optimal transport maps, which are defined as the unique optimal transport map between density estimates of the underlying distributions. In this work, we show that such estimators enjoy pointwise central limit theorems. These results provide a first step toward the problem of performing statistical inference for smooth optimal transport maps in general dimension. Our proofs hinge upon a quantitative linearization of a Monge-Ampère equation, which allows us to reduce our problem to that of deriving limit laws for the solution of a uniformly elliptic partial differential equation with a stochastic right-hand side.

Consistency for a general class of Random Forest type algorithms

Ricardo Blum

(joint work with Munir Hiabu, Enno Mammen, Joseph Theo Meyer)

We present a unifying consistency theorem for a broad class of tree-based algorithms by introducing a probabilistic sufficient impurity decrease condition. Our theory can be applied to algorithms that vary from traditional Random Forests due to additional randomness for choosing splits, allowing partitions into more than two cells in a single iteration step, and combinations of these. For example, our theory can be used to derive consistency of Extremely Randomized Trees and Interaction Forests. Furthermore, we demonstrate consistency for a larger function class compared to previous results on Random Forests if one allows for additional random splits.

References

[1] R. Blum, M. Hiabu, E. Mammen, J.T. Meyer *Consistency of Random Forest Type Algorithms under a Probabilistic Impurity Decrease Condition*, arXiv Preprint, arXiv:2309.01460, 2024.

[2] R. Blum, M. Hiabu, E. Mammen, J.T. Meyer *Hidden Variables unseen by Random Forests*, arXiv Preprint, arXiv:2406.15500, 2024.

Early stopping for conjugate gradients in statistical inverse problems

Laura Hucker
(joint work with Markus Reiß)

The conjugate gradient (CG) algorithm is arguably one of the computationally most efficient off-the-shelf methods for solving systems of linear equations. Like for standard gradient descent methods, stopping the algorithm *early*, that is before terminating at the solution, induces a regularisation. In the context of deterministic ill-posed inverse problems the seminal work by Nemirovskii [5] and Hanke [3] has shown that a stopping criterion based on the discrepancy principle, which monitors when the residual norm reaches a given threshold depending on the noise level, can lead to optimal convergence rates under bounded noise. Despite its importance in applications, the case of statistical noise has been understood only partially for CG so far. Major reasons are the nonlinearity of the CG algorithm and that canonical Gaussian white noise ξ on an infinite-dimensional Hilbert space has infinite norm $\|\xi\| = \infty$ almost surely. Therefore, the discrepancy principle cannot be well-defined.

In this talk, we consider estimators obtained by iterates of the standard CGNE (*conjugate gradients for the normal equation*) algorithm, also popular under the name *partial least squares*, applied to a prototypical statistical inverse problem under Gaussian white noise. When implemented, the CGNE algorithm is necessarily finite-dimensional, and we analyse it carefully, keeping explicitly track of the underlying dimension D. This is in line with the approach by Blanchard, Hoffmann and Reiß [1, 2] for linear spectral methods. Due to the nonlinear dependence on the noise, however, the analysis must be undertaken for every noise realisation and often requires more sophisticated arguments. The main contributions of our work [4] presented in this talk are the following:

- We identify two random quantities, called *stochastic error* and *approximation error*, which share important properties of variance and bias and allow for a precise nonasymptotic CG error control.
- Interpolating linearly between CG iterates allows us to equilibrate stochastic and approximation error. The oracle stopping time defined in this way achieves optimal prediction error control under minimal assumptions on the unknown signal and on the noise.
- We construct a data-driven residual-based stopping rule τ, depending on previous iterates $t \leq \tau$ only, that satisfies an oracle-type inequality for the prediction error and achieves optimal convergence rates whenever the error in estimating the noise level $\|\xi\|^2$ is not dominant. The choice of τ circumvents the computational drawback of classical model selection criteria, which make use of the entire iteration path.
- The convergence rates transfer to the reconstruction error, thus establishing minimax optimality of the best iterate along the CG iteration path and – under restrictions on the dimension D – of the iterate selected by our early stopping criterion τ.

References

[1] G. Blanchard, M. Hoffmann, M. Reiß, *Early stopping for statistical inverse problems via truncated SVD estimation*, Electronic Journal of Statistics **12** (2018), 3204–3231.

[2] G. Blanchard, M. Hoffmann, M. Reiß, *Optimal adaptation for early stopping in statistical inverse problems*, SIAM/ASA Journal on Uncertainty Quantification **6** (2018), 1043–1075.

[3] M. Hanke, *Conjugate gradient type methods for ill-posed problems*, Chapman and Hall/CRC, New York (1995).

[4] L. Hucker, M. Reiß, *Early stopping for conjugate gradients in statistical inverse problems*, arXiv:2406.15001 (2024).

[5] A.S. Nemirovskii, *The regularizing properties of the adjoint gradient method in ill-posed problems*, USSR Computational Mathematics and Mathematical Physics **26** (1986), 7–16.

Low intrinsic dimensionality is all you need

Shayan Hundrieser

(joint work with Thomas Staudt, Michel Groppe, Axel Munk)

The theory of optimal transport (OT) offers versatile tools for the comparison of probability measures in a geometrically faithful way. Formally, given a measurable cost function $c\colon \mathbb{R}^d \times \mathbb{R}^d \to [0,\infty)$ and two probability measures P, Q on $\mathbb{R}^d$, the OT cost is defined

$$T_c(P,Q) := \inf_{\pi \in \Pi(P,Q)} \int c(x,y) \mathrm{d}\pi(x,y),$$

where $\Pi(P,Q)$ denotes the collection of couplings between P and Q. This quantity has found various applications, e.g., in economics, machine learning and biology. Oftentimes, for computational benefits, practitioners relies a regularized variant, namely the entropy penalized OT (EOT) cost with regularization parameter $\varepsilon > 0$,

$$T_c^\varepsilon(P,Q) := \inf_{\pi \in \Pi(P,Q)} \int c(x,y) \mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi \| P \otimes Q),$$

$$\text{with} \quad \mathrm{KL}(\pi \| P \otimes Q) := \begin{cases} \int \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}P \otimes Q}(x,y)\right) & \text{if } \pi \ll P \otimes Q, \\ +\infty & \text{else.} \end{cases}$$

In applied contexts, one often relies on estimating the OT and EOT cost by an empirical plug-in approach. Assuming i.i.d. random variables $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_n \sim Q$, with corresponding empirical measures $\hat{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\hat{Q}_n := \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$, the resulting plug-in estimators for the OT and EOT cost are given by $T_c(\hat{P}_n, \hat{Q}_n)$ and $T_c^\varepsilon(\hat{P}_n, \hat{Q}_n)$. The convergence behavior of the empirical OT and EOT cost for increasing sample size is dictated by various aspects and affected by the curse of dimensionality. Indeed, when only imposing boundedness of the support of the population measures it follows for costs $c(x,y) = \|x-y\|^p$ with $p \geq 1$ and $n \geq 1$ according to [3] that

$$\inf_{\hat{T}_c} \sup_{P,Q \in \mathcal{P}([0,1]^d)} \mathbb{E}\left[\left|\hat{T}_c - T_c(P,Q)\right|\right] \geq K\left[(n\log(n+1))^{-\min(p,2)/d} + n^{-1/2}\right],$$

where the infimum is taken over all estimators based on $X_1, \dots, X_n, Y_1, \dots, Y_n$ and $K > 0$ is a positive constant that depends on d and p. Likewise, for the EOT cost we show in our work [1] for small $\varepsilon > 0$ that

$$\begin{aligned}&\inf_{\hat{T}_c^\varepsilon} \sup_{P,Q\in\mathcal{P}([0,1]^d)} \mathbb{E}\left[\left|\hat{T}_c^\varepsilon - T_c^\varepsilon(P,Q)\right|\right]\\ &\geq K\left[(n\log(n+1))^{-\min(p,2)/d} + n^{-1/2}\right] - 2d\varepsilon|\log(\varepsilon)|,\end{aligned}$$

confirming when $\varepsilon > 0$ is small that estimation of the EOT cost is also affected by the curse of dimensionality .

The central contribution of our works [1, 2] concerns the statistical convergence rate of the empirical OT and EOT cost under structural assumptions on the intrinsic dimensions of the population measures. Under distinct population measures with different intrinsic dimensions, we establish that the convergence rate for the empirical OT cost adapts to the population measures in the most favorable way, being determined by the lower dimensional measure. More precisely, we show in [2] for the empirical OT cost based on a cost function that is α-Hölder regular and assuming P and Q are concentrated on compact submanifolds of dimension s and t, respectively, that

$$\mathbb{E}\left[\left|\hat{T}_c(\hat{P}_n,\hat{Q}_n) - T_c(P,Q)\right|\right] \leq K \cdot \begin{cases} n^{-\min(\alpha,2)/\min(s,t)} & \text{if } \min(s,t) > 2\min(\alpha,2),\\ n^{-1/2}\log(n+1) & \text{if } \min(s,t) = 2\min(\alpha,2),\\ n^{-1/2} & \text{if } \min(s,t) < 2\min(\alpha,2),\end{cases}$$

where K only depends on the cost function and the manifold with the lower dimension. This phenomenon represents a hallmark feature of empirical optimal transport and we term it *lower complexity adaptation*. In addition, we also confirm the empirical EOT cost to benefit from low intrinsic dimensionality of one measure. More precisely, we show in [1] for an α-Hölder regular cost function with $\alpha > \min(s,t)/2$ that

$$\mathbb{E}\left[\left|\hat{T}_c^\varepsilon(\hat{P}_n,\hat{Q}_n) - T_c^\varepsilon(P,Q)\right|\right] \leq K(1+\varepsilon^{-\min(s,t)/2})n^{-1/2},$$

which highlights that the adaptation to lower complexity manifests in the dependency of the regularization parameter. Overall, our works establish that low intrinsic dimensionality of a single population measure is sufficient in order to expect fast convergence rates of the empirical OT and EOT cost.

References

[1] M. Groppe, S. Hundrieser, *Lower complexity adaptation for empirical entropic optimal transport*, Preprint arXiv:2306.13580 (2023).

[2] S. Hundrieser, T. Staudt, A. Munk, *Empirical optimal transport between different measures adapts to lower complexity*, Ann. inst. Henri Poincaré (B) Probab. Stat. **60**(2) (2024), 824–846.

[3] T. Manole, J. Niles-Weed, *Sharp convergence rates for empirical optimal transport with smooth costs*, Ann. Appl. Probab. **34**(1B) (2024), 1108–1135.

Bernstein-type inequalities for unbounded martingales

Alexandra Suvorikova
(joint work with Alexey Kroshnin)

Nowadays, concentration inequalities are pivotal in many areas, offering crucial insights into the non-asymptotic analysis of the behavior of random variables. Among the various forms of concentration inequalities, Bernstein-type ones have garnered significant attention due to their scope of applicability. They are essential in numerous fields, including statistical learning theory, empirical process theory, and high-dimensional statistics. Their applicability ranges from deriving error bounds for machine learning algorithms to ensuring the reliability of high-dimensional data analysis.

The classical Bernstein inequality, derived in the early 20th century by Sergei Bernstein, provides a probabilistic bound for the sum of independent, centered random variables with bounded absolute values [1]. Over the decades, the scope and utility of Bernstein-type inequalities have expanded, leading to various generalizations and refinements. It is worth noting that Sergei Bernstein himself relaxed the assumption on bounded random variables and replaced it with the assumption on bounded moments.

George Bennett obtained a tighter bound for bounded independent observations [2]. Several years later, Vadim Yurinskii generalized Bernstein's result to the case of random variables in Banach spaces, assuming Bernstein's moment condition on the norm [3]. Around the same time, David Freedman relaxed the independence assumption and derived the concentration inequality for bounded martingales [4].

In the 2000s, there was increased interest in the matrix case: Joel Tropp generalized Freedman's result to the case of matrix-valued martingales satisfying Bernstein's moment condition [5]. At the same time, Vladimir Koltchinskii obtained a Bernstein-type result for independent random Hermitian matrices with bounded Orlicz norm [6].

This work suggests a novel result in the same direction: we consider matrix-valued martingale differences, assuming their Orlicz norm to be bounded. Furthermore, we illustrate the applicability of the result by deriving a McDiarmid-type inequality.

References

[1] S. Bernstein, *Theory of probability*, Moscow, (1927).
[2] G. Bennett, *Probability inequalities for the sum of independent random variables*, Journal of the American Statistical Association **57** (1962), no.297, 33–45.
[3] V. Yurinskii, *Exponential inequalities for sums of random vectors*, Journal of Multivariate Analysis **6** (1976), no. 4, 473–499.
[4] D. A. Freedman, *On tail probabilities for martingales*, The Annals of Probability, (1975), 100–118.
[5] J. Tropp, *Freedman's inequality for matrix martingales*, Electronic Communications in Probability **16** (2011), 262–270.

[6] V. Koltchinskii, *Oracle inequalities in empirical risk minimization and sparse recovery problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008*, vol. 2033, Springer Science & Business Media, (2011).

Non-asymptotic and non-minimax estimation of a smooth functionals

Vladimir Spokoiny

The talk discusses a general framework of statistical estimation by a quasi maximum likelihood method without specifying any particular structure of the data. The obtained results are stated first for linear models and then extended to so-called stochastically linear models with a linear stochastic component (SLS).

Later it is shown how a general model can be transformed to fit the SLS framework by extending the parameter space. The approach is illustrated by the case of estimation of a smooth functional.

Laplace asymptotics in high-dimensional Bayesian inference

Anya Katsevich

Developing cheap and accurate computational techniques for Bayesian inference is an important goal, as Bayesian inference tasks can be very computationally intensive. These tasks include computing posterior credible sets, posterior mean and covariance, and the evidence (marginal likelihood) of the data. Computing all of these quantities involves either sampling from the posterior π, taking integrals $\int g d\pi$ against the posterior, or integrating the unnormalized posterior. When the dimensionality d of the parameter is large, these tasks can be very expensive.

A popular approach to simplify such computations is to find a simple distribution $\hat{\gamma}$ which approximates π, and to use this distribution as a proxy for π to do all of one's inference tasks. In particular, we approximate $\int f d\pi$ by $\int f d\hat{\gamma}$ and in the ideal scenario, many integrals against $\hat{\gamma}$ are computable in closed form. The idea of using an approximation $\hat{\gamma}$ to π is at the heart of approximate Bayesian inference methods such as variational inference, expectation propagation, and the Laplace approximation (LA), the focus of our studies. The idea of the LA is to exploit large sample properties of the posterior. Namely if certain conditions are met (e.g. if the statistical model is well-specified), then the uncertainty in the posterior decreases as more and more samples are collected. Therefore as sample size $n \to \infty$, the mass of the posterior π concentrates in a small neighborhood of the mode, which we call $\hat{x}$. Since most of the mass of π is near $\hat{x}$, we should incur only a small error by replacing the log posterior with its second order Taylor expansion about $\hat{x}$. This gives rise to the LA, the Gaussian $\hat{\gamma}$ given by

$$\hat{\gamma} = \mathcal{N}\left(\hat{x},\ \nabla^2 V(\hat{x})^{-1}\right), \quad \hat{x} = \arg\min_{x \in \Theta} V(x) \tag{1}$$

where $\pi \propto e^{-V}$ is a density on $\Theta \subseteq \mathbb{R}^d$. The concentration of π about the mode can also be explained by the fact that we can write

$$V = nv, \qquad \pi \propto e^{-V} = e^{-nv}$$

for a function v which only weakly depends on n as $n \to \infty$. Thus $\pi \propto e^{-nv}$ concentrates around the global minimizer $\hat{\theta}$ of v.

The LA has proved to be an invaluable tool for Bayesian inference in applications ranging from deep learning to inverse problems to variable selection in high-dimensional regression. Quantifying the LA's error as a function of dimension d, sample size n, and model parameters, is a worthy task given its widespread use. It is also a challenging theoretical endeavor when dimension d is large, and currently a very active research area. Major contributions have been made e.g. by [5, 12, 3]. But arguably, it is even more important to go *beyond* the LA to develop new, more accurate approximations which better capture the complexity of the posterior π. For example, a known downside of the LA $\hat{\gamma}$ is that it is symmetric about the mode and therefore cannot capture skewness of π. Instead of constructing an entirely new kind of approximation, a natural idea is to correct the LA in some way to get a higher-order accuracy approximation. Only a single work [2] rigorously derives a higher-order accurate LA. However, it is only shown to be accurate in constant dimension d. So far, no prior work has obtained a higher-accuracy LA which is rigorously justified in high dimensions.

In the work [6], we develop a powerful technique to analyze the Laplace approximation more precisely than was possible before. This technique leads us to derive the *first ever correction to the LA which provably improves its accuracy by an order of magnitude, in high dimensions.* At the same time, the more accurate, skew-corrected LA — which we call $\hat{\gamma}_S$ — retains the useful property of $\hat{\gamma}$ that integrals $\int f d\hat{\gamma}_S$ can be computed in closed form when f is a polynomial.

Our approach allows us to prove error bounds on the approximation $\pi \approx \hat{\gamma}_S$ in terms of a variety of error metrics. It also improves our understanding of the accuracy of the uncorrected LA itself: we prove both tighter upper bounds and the first ever lower bounds on the standard LA in high dimensions. In particular, we prove that $d^2 \ll n$ is in general necessary for accuracy of the LA.

If $\pi \propto e^{-nv}$ then $\int f d\pi = \int f e^{-nv} / \int e^{-nv}$, which is the ratio of two Laplace-type integrals (i.e. integrals involving an exponential, with a large parameter n in the exponent). In the second part of the talk we consider directly approximating Laplace-type integrals by an asymptotic expansions in powers of n^{-1}, whose coefficients are given in terms of derivatives of f and v at $\hat{\theta} = \operatorname{argmin}_\theta v(\theta)$. This allows us to a) directly approximate $\int f d\pi$ by a ratio of two numbers, rather than by a second integral which still may not be so simple to compute, and b) compute the normalizing constant $\int e^{-nv}$, which is inaccessible using only an approximation of the density π by a second density $\hat{\gamma}$ or $\hat{\gamma}_S$. The normalizing constant, or evidence, is of central importance in Bayesian model selection.

Asymptotic expansions of Laplace-type integrals are a classical subject in asymptotic analysis [14], but most results in this area consider dimension d to be constant relative to the large parameter n. Works obtaining remainder bounds depending explicitly on dimension either have exponential dimension dependence, consider particular forms of f and v, or only consider the expansion of $\int e^{-nv}$ to zeroth order. See [10, 4, 9, 1, 11, 13] for partial results on Laplace expansions.

In [7], we derive the asymptotic expansion of Laplace-type integrals in high dimension. Namely, we show that if $0 = \operatorname{argmin}_\theta v(\theta)$ and $\nabla^2 v(0) = I_d$ (the general case can be derived through a change of variables), then under appropriate regularity conditions it holds

$$\frac{e^{nv(0)}}{(2\pi/n)^{d/2}} \int_{\mathbb{R}^d} f(x) e^{-nv(x)} dx = f(0) + \sum_{k=1}^{L-1} a_k n^{-k} + \mathrm{Rem}_L.$$

The coefficients a_k coincide with those derived explicitly for the first time in [8]. The main contribution lies in our remainder error bound, which takes the form

$$|\mathrm{Rem}_L| \lesssim_L C \left(\frac{d}{\sqrt{n}}\right)^{2L},$$
$$C = C\left(d, |f|, \|\nabla f\|, \ldots, \|\nabla^{2L} f\|, \|\nabla^3 v\|, \ldots, \|\nabla^{2L+2} v\|\right).$$

Here, C is a fully explicit function, and $\|\nabla^k g\|$ is shorthand for $\sup_{\|x\| \le R\sqrt{d/n}} \|\nabla^k g(x)\|$, for some absolute constant R. The dependence of C on the terms $\|\nabla^k g\|$ is polynomial, and the dependence on d is only beneficial: d appears raised to various negative powers in front of the terms $\|\nabla^k g\|$. Our result is a complete extension of the classical Laplace expansion to the high-dimensional setting. The expansion can now be used in applications where dimension cannot be considered constant relative to the large parameter n, both in Bayesian inference and beyond.

References

[1] Rina Foygel Barber, Mathias Drton, and Kean Ming Tan. Laplace approximation in high-dimensional Bayesian regression. In *Statistical Analysis for High-Dimensional Data: The Abel Symposium 2014*, pages 15–36. Springer, 2016.

[2] Daniele Durante, Francesco Pozza, and Botond Szabo. Skewed Bernstein-von Mises theorem and skew-modal approximations. *arXiv preprint arXiv:2301.03038*, 2023.

[3] Tapio Helin and Remo Kretschmann. Non-asymptotic error estimates for the Laplace approximation in Bayesian inverse problems. *Numerische Mathematik*, 150(2):521–549, 2022.

[4] Tadeusz Inglot and Piotr Majerski. Simple upper and lower bounds for the multivariate Laplace approximation. *Journal of Approximation Theory*, 186:1–11, 2014.

[5] Mikolaj J Kasprzak, Ryan Giordano, and Tamara Broderick. How good is your Gaussian approximation of the posterior? Finite-sample computable error bounds for a variety of useful divergences. *arXiv preprint arXiv:2209.14992*, 2022.

[6] Anya Katsevich. The Laplace approximation accuracy in high dimensions: a refined analysis and new skew adjustment. *arXiv preprint arXiv:2306.07262*, 2023

[7] Anya Katsevich. The Laplace asymptotic expansion in high dimensions: a nonasymptotic analysis. *arXiv preprint arXiv:2406.12706*, 2024

[8] William D Kirwin. Higher asymptotics of laplace's approximation. *Asymptotic Analysis*, 70(3-4):231–248, 2010.

[9] Vassili N Kolokoltsov. Rates of convergence in Laplace's integrals and sums and conditional central limit theorems. *Mathematics*, 8(4):479, 2020.

[10] Tomasz M Łapiński. Multivariate Laplace's approximation with estimated error and application to limit theorems. *Journal of Approximation Theory*, 248:105305, 2019.

[11] Helen Ogden. On the error in Laplace approximations of high-dimensional integrals. *Stat*, 10(1):e380, 2021.

[12] Vladimir Spokoiny. Dimension free nonasymptotic bounds on the accuracy of high-dimensional laplace approximation. *SIAM/ASA Journal on Uncertainty Quantification*, 11(3):1044–1068, 2023.
[13] Yanbo Tang and Nancy Reid. Laplace and saddlepoint approximations in high dimensions. *arXiv preprint arXiv:2107.10885*, 2021.
[14] R. Wong. *Asymptotic Approximations of Integrals*. Society for Industrial and Applied Mathematics, 2001.

Reporter: Arshak Minasyan

Participants

Dr. Francis Bach
INRIA
Departement d'Informatique
École Normale Supérieure
Voie DQ 12
2, rue Simone Iff
75012 Paris Cedex
FRANCE

Misha Belkin
Halicioglu Data Science Institute
University of California, San Diego
10100 Hopkins Drive
La Jolla, CA 92093-0112
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

Nayel Bettache
CREST
5, Avenue Henry Le Chatelier
91120 Palaiseau
FRANCE

Ricardo Blum
Institut für Mathematik
Universität Heidelberg
Im Neuenheimer Feld 205
69120 Heidelberg
GERMANY

Dr. Claire Boyer
Laboratoire de Probabilités, Statistique
et Modélisation (LPSM), BP 158
Sorbonne Université
Campus Pierre et Marie Curie
4, place Jussieu
75252 Paris Cedex 05
FRANCE

Dr. Jelena Bradic
Department of Mathematics
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0112
UNITED STATES

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. 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

Dr. Luiz Chamon
Universität Stuttgart
Pfaffenwaldring 5a
70569 Stuttgart
GERMANY

Prof. Dr. Arnak Dalalyan
CREST – ENSAE
École Nationale de la Statistique et de l'Administration Économique
5, Avenue Henry Le Châtelier
91120 Palaiseau 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. Niao He
Institute of Machine Learning
ETH Zürich
Andreasstrasse 5
8050 Zürich
SWITZERLAND

Laura Hucker
Institut für Mathematik
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin
GERMANY

Dr. Shayan Hundrieser
Institut für Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstraße 7
37077 Göttingen
GERMANY

Dr. Arthur Jacot
Courant Institute of Mathematical Sciences
New York University
251, Mercer Street
New York, NY 10012
UNITED STATES

Prof. Dr. Kengo Kato
Department of Statistics and Data Science
Cornell University
1194 Comstock Hall
Ithaca, NY 14853-2601
UNITED STATES

Dr. Anya Katsevich
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge 02139-4307
UNITED STATES

Dr. Yu Lu
Department of Statistics
ENSAE/CREST/IP Paris
5 Avenue Henry Le Chatelier
91120 Palaiseau
FRANCE

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

Tudor Manole
Department of Statistics and Data Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213-3890
UNITED STATES

Dr. Arshak Minasyan
Department of Statistics
ENSAE/CREST/IP Paris
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

Bianca Neubert
Institut für Mathematik
Universität Heidelberg
Im Neuenheimer Feld 205
69120 Heidelberg
GERMANY

Dr. Jonathan Niles-Weed
Courant Institute of Mathematical
Sciences and Center for Data Science,
New York University
New York, NY 10011
UNITED STATES

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

Prof. Dr. Markus Reiß
Institut für Mathematik
Humboldt-Universität Berlin
Unter den Linden 6
10117 Berlin
GERMANY

Prof. Dr. Philippe Rigollet
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139-4307
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
Drienerlolaan 5
7522 NB Enschede
NETHERLANDS

Prof. Dr. Ulrike Schneider
Technische Universität Wien
Wiedner Hauptstraße 8 - 10
1040 Wien
AUSTRIA

Dr. Bodhisattva Sen
Department of Statistics
Columbia University
1255 Amsterdam Avenue
New York NY 10027
UNITED STATES

Tomasz Skalski
Wydział Matematyki
Politechnika Wrocławska
ul. Hoene-Wrońskiego 13C
50-376 Wrocław
POLAND

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

Dr. Alexandra Suvorikova
Weierstrass Institute for Applied Analysis and Stochastics
Mohrenstr. 39
10117 Berlin
GERMANY

Prof. Dr. Ryan Tibshirani
Department of Statistics
University of California, Berkeley
Berkeley, CA 94707
UNITED STATES

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

Prof. Dr. em. Sara van de Geer
Seminar für Statistik
ETH Zürich
Rämistrasse 101
8092 Zürich
SWITZERLAND

Dr. Ramya Vinayak
Department of Electrical and Computer Engineering
University of Wisconsin-Madison
1415 Engineering Dr
53706 Madison, WI 53706-1685
UNITED STATES

Prof. Dr. Martin Wahl
Fakultät für Mathematik
Universität Bielefeld
Postfach 100131
33501 Bielefeld
GERMANY

Y. Samuel Wang
Department of Statistics and Data Science
Cornell University
129 Garden Ave
Ithaca, NY 14853
UNITED STATES

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

Prof. Dr. Rebecca Willett
Department of Statistics
The University of Chicago
5747 S. Eillis Avenue
Chicago, IL 60637
UNITED STATES

Prof. Dr. Fanny Yang
Department of Computer Science
ETH Zürich (CAB G 68)
Universitätsstrasse 6
8092 Zürich
SWITZERLAND

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