Mathematisches Forschungsinstitut Oberwolfach

Report No. 17/2025

DOI: 10.4171/OWR/2025/17

Frontiers of Statistics and Machine Learning

Organized by
Marc Hoffmann, Paris
Richard J. Samworth, Cambridge UK
Johannes Schmidt-Hieber, Enschede
Claudia Strauch, Heidelberg

23 March – 28 March 2025

Abstract. AI is currently the central theme in science. Whereas the underlying algorithms rely on rather simple mathematical operations such as matrix-vector multiplications and applying non-linearities componentwise, deriving a theoretical understanding proves to be extremely challenging. To identify synergies between the fields of mathematical statistics and theoretical machine learning, the workshop brought together leading researchers and rising stars who are tackling core challenges at the intersection of these fields. We have identified the topics of robustness and model misspecification, statistical theory for neural networks and statistics for stochastic processes as three key themes that underpin increasingly many current developments. These topics were the focus of the talks and research that was carried out during the Oberwolfach week.

Mathematics Subject Classification (2020): 62G05, 62G08, 62G20.

Introduction by the Organizers

The workshop *Frontiers of Statistics and Machine Learning* was attended by 47 participants (46 on site and one online). The workshop brought together researchers with diverse backgrounds. The participants came from universities in the US, Japan, and Europe. The event featured around 22 talks and we organized an evening session with short presentations by junior participants (their abstracts are also included in this report). The talks sparked numerous questions and (follow-up) discussions.

Until recently, statistics and machine learning were developed by nearly disjoint communities. Due to these independent developments, data science/machine learning and statistics differ in their approaches to data problems. This distinction is highlighted in Leo Breiman's "Two cultures" [1]. While data science starts with specific benchmark data sets and data competitions, statistics begins with the modelling of the sampling process. The more pragmatic, engineering-oriented approach of data scientists has a particular advantage in dealing with complex data structures where statistical modelling is unclear, often leading to better procedures. Conversely, statistics can squeeze out more information if the data distribution can be modelled. In this case one can often say more about uncertainty quantification, whether Bayesian or frequentist, which remains one of the challenging problems in data science.

Unifying these fields with the goal to combine the different strengths is an ongoing and very active branch within statistics and machine learning. The workshop aimed to summarize the current state-of-the-art and push the frontiers of both statistics and machine learning.

Within this field, we have identified three highly relevant subjects that are currently experiencing tremendous developments. These selected subjects are robustness, theory for neural networks and statistical theory for stochastic processes. All of them are intricately interconnected. While robustness to outliers is a classical topic within statistics with a well-developed mathematical theory, new ideas and concepts are currently developed to deal with very different forms of robustness, such as robustness of machine learning methods to a new distribution of the covariates during test time (covariate shift) or robustifying neural networks against adversarial attacks. Theory for neural networks has become a very active subject in the past years and combines elements from various areas in mathematics. Statistical theory for stochastic processes has a long tradition within mathematical statistics and the challenge is to extend the theory to machine learning applications, such as mean-field limits of neural networks [2] and the recently-developed stable diffusion sampling procedures for generative AI [3].

Below we provide a more detailed description of these three key subjects for the workshop.

Robustness and model misspecification

For contemporary AI applications, simple statistical models that underpin the way we typically think of small data sets having been generated may no longer be fit for purpose. Large-scale data are usually messy: data may be collected under different conditions, data may be missing and data may be corrupted. Basic model checks that are effective in traditional, low-dimensional settings, may become completely infeasible when there are so many possible departures from an assumed model.

A classical approach to modeling robustness assumes that a fraction of the data are outliers. Chao Gao demonstrated in his talk that if an ϵ-fraction of the sample is arbitrarily perturbed and ϵ is unknown then it provably becomes much harder to construct confidence intervals. Tengyao Wang explored testing of regression coefficients in highdimensional settings with heavy-tailed noise. Working also in the linear regression model, Min Xu showed optimality statements for a data-driven convex loss function.

Rajen Shah introduced some new methods linking robustness and efficiency in semiparametric models. Missing data were discussed in the talk by Kabir Verchand. Transfer learning refers to statistical problems in which we wish to make inference about a target data population, but where some (typically, the large majority) of our training data come from a related but distinct distribution. In his talk, Martin Wainwright explored the specific case of covariate shift.

Data privacy can be considered as a strong notion of robustness concerning the values of individual data points. Yi Yu considered in her talk privacy for functional data and Tom Berrett developed a theoretical framework addressing scenarios where an individual might be associated with multiple data points in a dataset. Chiara Amorino worked out the minimax rate for multivariate data under privacy constraints.

Statistical theory for neural networks and theory of machine learning
Artificial neural networks (ANNs) are at the core of the AI revolution. In the past years, enormous efforts have been made to unravel their mathematical properties, leading to fundamental insights and mathematical guarantees on when and why deep learning works well.

Regarding the energy landscape, Andrea Montanari derived bounds on the expected number of local minima via the Kac-Rice formula. In his talk, Peter Bartlett revealed the existence of different regimes when training a logistic regression model with a fixed learning rate. Optimisation was also addressed by Alexandra Carpentier in her talk, where she proposed and analyzed a zeroth-order optimization method in the strongly convex regime.

Classification is a crucial learning task in modern machine learning and was addressed in the talks by Henry Reeve and Holger Dette. In classification tasks, neural networks often output probabilities that tend to be overconfident. To address this issue, Francis Bach proposed a method and developed theoretical foundations to calibrate these outputs.

Nicolai Meinshausen proposed a neural network-based distributional regression methodology called ‘engression’. Nicole Mücke’s talk explored the interface between robustness and theory of machine learning, treating ridge regression with heavy-tailed noise.

Mathias Trabs gave a talk on confidence bands for random forests.

Regarding generative AI, Yuting Wei summarized her recent contributions to the theory for diffusion models.

Statistics for stochastic processes
Understanding the statistical properties of stochastic processes is crucial for analyzing machine learning procedures because many real-world phenomena, including data generation and model dynamics, can be modeled as stochastic processes. It is, however, challenging to develop statistical procedures and statements that are sufficiently general and robust to be relevant for investigations in the ML context.

In his talk, Sven Wang explored likelihood-based methods for low-frequency diffusion data. Mark Podolskij's presentation covered interacting particle systems, focusing on the statistical estimation of McKean-Vlasov stochastic differential equations. Additionally, Arnak Dalalyan discussed a parallelizable sampling scheme for Langevin Monte Carlo.

The workshop also featured talks on adaptive density estimation under low-rank constraints (Olga Klopp), bandits and online learning (Alexandre Tsybakov) and the emerging topic of e-values (Wouter Koolen).

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".

References

[1] L. Breiman, *Statistical modeling: The two cultures.*, Statistical Science, **16** (2001), 199–215.

[2] S. Mei, A. Montanari, P-M. Nguyen, *A mean field view of the landscape of two-layer neural networks.*, Proc. Natl. Acad. Sci. U.S.A., **115** (2018), E7665–E7671.

[3] K. Oko, S. Akiyama, T. Suzuki, *Diffusion models are minimax optimal distribution estimators.*, Proceedings of the 40th International Conference on Machine Learning (ICML'23), **202** (2023), 26517–26582.

Workshop: Frontiers of Statistics and Machine Learning

Table of Contents

Abstracts

Statistical and Computational Scaling Low in Test Time Inference

Taiji Suzuki

(joint work with Juno Kim, Jason D. Lee, Naoki Nishikawa, Yujin Song,
Kazusato Oko, and Denny Wu)

In this talk, I introduce recent theoretical developments that elucidate the learning capabilities of Transformers, where we analyze statistical and computational complexities in test-time inference methods such as chain-of-thought and in-context learning.

(1) Analysis of chain-of-thought. In the first half, I discuss theoretical guarantees of chain-of-thought (CoT) that recursively generates intermediate states to solve complex problems. We consider training a one-layer transformer to solve the fundamental k-parity problem, extending the work on RNNs by [7]. We establish three key results: (i) any finite-precision gradient-based algorithm, without intermediate supervision, requires substantial iterations to solve parity with finite samples. (ii) In contrast, when intermediate parities are incorporated into the loss function, our model can learn parity in one gradient update when aided by *teacher forcing*, where ground-truth labels of the reasoning chain are provided at each generation step. (iii) Even without teacher forcing, where the model must generate CoT chains end-to-end, parity can be learned efficiently if augmented data is employed to internally verify the soundness of intermediate steps. Our findings, supported by numerical experiments, show that task decomposition and stepwise reasoning naturally arise from optimizing transformers with CoT; moreover, self-consistency checking can improve multi-step reasoning ability, aligning with empirical studies of CoT. This part is mainly based on [3].

(2) Analysis of in-context learning. In the second half, I discuss in-context (IC) learning. Transformers can efficiently learn in-context from a few numbers of example demonstrations. Most existing theoretical analyses studied the ICL ability of transformers for linear function classes, where it is typically shown that the minimizer of the pretraining loss implements one gradient descent step on the least squares objective. However, this simplified linear setting arguably does not demonstrate the statistical efficiency of ICL, since the pretrained transformer does not outperform directly solving linear regression on the test prompt. In this work, we study ICL of a nonlinear function class via transformer with nonlinear MLP layer: given a class of *single-index* target functions

$$f_*(\boldsymbol{x}) = \sigma_*(\langle \boldsymbol{x}, \boldsymbol{\beta} \rangle),$$

where the index features $\boldsymbol{\beta} \in \mathbb{R}^d$ are drawn from a r-dimensional subspace and $\boldsymbol{x}$ follows i.i.d. standard normal distribution $N(\mathbf{0}, I_d)$. We show that a nonlinear transformer optimized by gradient descent learns f_* in-context with a prompt length that only depends on the dimension of the distribution of target functions r; in contrast, any algorithm that directly learns f_* on test prompt yields a statistical complexity that scales with the ambient dimension d. Interestingly, the pretraining sample complexity is characterized by the *information exponent* of the link functions σ_* instead of the maximum degree of it. Our result highlights the adaptivity of the pretrained transformer to low-dimensional structures of the function class, which enables sample-efficient ICL that outperforms estimators that only have access to the in-context data. This part is based on [6].

In addition to that, we investigate how transformers learn features in-context – a key mechanism underlying their inference-time adaptivity. We again consider the single-index models but assume the soft-max attention is employed in the transformer architecture here. We prove that transformers with soft-max attention pretrained by gradient-based optimization can perform *inference-time feature learning*, i.e., extract information of $\boldsymbol{\beta}$ solely from test prompts (despite $\boldsymbol{\beta}$ varying across different prompts), hence achieving inference-time statistical efficiency that surpasses any non-adaptive (fixed-basis) algorithms such as kernel methods. Moreover, we show that the inference-time sample complexity surpasses the Correlational Statistical Query (CSQ) lower bound, owing to nonlinear label transformations naturally induced by the self-attention mechanism. This part is based on [5].

This result arises from the fact the nonlinear transformation achieved by the soft-max attention lower the information exponent of the target function to the *generative exponent.* More precisely, the information exponent and the generative exponent are defined as follows. Considering the Hermite expansion of a function f, i.e., $f(z) = \sum_{i\geq 0} \frac{c_i}{i!}\mathrm{He}_i(z)$ where He_i is the degree-i (probablists') Hermite polynomia, we define $\mathrm{H}(f,i) := c_i$ as its degree-i coefficient.

- The *information exponent* [1] of σ_*, denoted by $\mathrm{ie}(\sigma_*)$, is defined as
$$\mathrm{ie}(\sigma_*) := \min\{i \mid \mathrm{H}(\sigma_*, i) \neq 0\}.$$
- The *generative exponent* [2] of σ_*, written as $\mathrm{ge}(\sigma_*)$, is defined as
$$\mathrm{ge}(\sigma_*) := \min_{h\in L^2} \min\{i \mid \mathrm{H}(h \circ \sigma_*, i) \neq 0\},$$
where L^2 denotes the set of all $L^2(P_Y)$-measurable transformations from $\mathbb{R}$ to $\mathbb{R}$ for $P_Y = \sigma_{*\#}N(0,1)$ where $\sigma_{*\#}$ is the push-forward by σ_*.

Then, [4] showed in their Proposition 6 that a polynomial σ_* satisfies
$$\mathrm{ge}(\sigma_*) = \begin{cases} 1 \text{ (if } \sigma_* \text{ is not even)}, \\ 2 \text{ (if } \sigma_* \text{ is even)}. \end{cases}$$

The key lemma to show effectiveness of the soft-max attention is the following one:

Lemma (Informal [5]) The information exponent of $\exp(\bar{\sigma}_*)$ and $\bar{\sigma}_* \exp(\bar{\sigma}_*)$ is equal to $\mathrm{ge}(\sigma_*)$, where $\bar{\sigma}_*$ is a *clipped version* of σ_*.

This lemma enables us to show that the transformers with soft-max attention can perform a inference-time feature learning with lower inference-time sample complexity.

References

[1] G. Ben Arous, R. Gheissari, and A. Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *The Journal of Machine Learning Research*, 22(1):4788–4838, 2021.

[2] A. Damian, L. Pillaud-Vivien, J. D. Lee, and J. Bruna. The computational complexity of learning gaussian single-index models. *arXiv preprint arXiv:2403.05529*, 2024.

[3] J. Kim and T. Suzuki. Transformers provably solve parity efficiently with chain of thought. In *The Thirteenth International Conference on Learning Representations*, 2025.

[4] J. D. Lee, K. Oko, T. Suzuki, and D. Wu. Neural network learns low-dimensional polynomials with sgd near the information-theoretic limit. *arXiv preprint arXiv:2406.01581*, 2024.

[5] N. Nishikawa, Y. Song, K. Oko, D. Wu, and T. Suzuki. Nonlinear transformers can perform inference-time feature learning: a case study of in-context learning on single-index models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. to appear.

[6] K. Oko, Y. Song, T. Suzuki, and D. Wu. Pretrained transformer efficiently learns low-dimensional target functions in-context. In *Advances in Neural Information Processing Systems*, volume 37, pages 77316–77365, 2024.

[7] N. Wies, Y. Levine, and A. Shashua. Sub-task decomposition enables learning in sequence to sequence tasks. In *The Eleventh International Conference on Learning Representations*, 2023.

Are Robust Confidence Intervals Possible?

Chao Gao

(joint work with Yuetian Luo)

We study the construction of confidence intervals under Huber's contamination model. When the contamination proportion is unknown, we characterize the necessary adaptation cost of the problem. For Gaussian location model, the optimal length of an adaptive confidence interval is proved to be exponentially wider than that of a non-adaptive one. In particular, with data $X_1, \cdots, X_n$ independently generated from $(1-\epsilon)N(\theta,1)+\epsilon Q$, we show that the optimal length of an adaptive confidence interval scales as

$$\frac{1}{\sqrt{\log n}} + \frac{1}{\sqrt{\log(1/\epsilon)}}.$$

An explicit optimal construction is given by

$$\begin{aligned}\widehat{\mathrm{CI}} \quad = \quad & \left[\max_{t\in[1,\log n]}\left(F_n^{-1}\left(2\left(1-\Phi(t)\right)\right)+t-\frac{2}{t}\right),\right. \\ & \left.\min_{t\in[1,\log n]}\left(F_n^{-1}\left(1-2\left(1-\Phi(t)\right)\right)-t+\frac{2}{t}\right)\right].\end{aligned}$$

Results for general location models will be discussed. In addition, we also consider the same problem in a network setting for an Erdos-Renyi graph with node contamination. It will be shown that the hardness of the adaptive confidence interval construction is implied by the detection threshold between Erdos-Renyi model and stochastic block model.

On Robustness in Semiparametric Statistics

Rajen Shah
(joint work with Elliot Young)

Given that all models are wrong, it is important to understand the performance of methods when the settings for which they have been designed are not met, and to modify them where possible so they are robust to these sorts of departures from the ideal. We present two examples with this broad goal in mind.

We first look at a classical case of model misspecification in (linear) mixed effect models for grouped data. Existing approaches estimate linear model parameters through weighted least squares, with optimal weights (given by the inverse covariance of the response, conditional on the covariates) typically estimated by maximising a (restricted) likelihood from random effects modelling or by using generalised estimating equations. We introduce a new 'sandwich loss' whose population minimiser coincides with the weights of these approaches when the parametric forms for the conditional covariance are well-specified, but can yield arbitrarily large improvements when they are not.

The starting point of our second vignette is the recognition that semiparametric efficient estimation can be hard to achieve in practice: estimators that are in theory efficient may require unattainable levels of accuracy for the estimation of complex nuisance functions. As a consequence, estimators deployed on real datasets are often chosen in a somewhat ad hoc fashion, and may suffer high variance. We study this gap between theory and practice in the context of a broad collection of semiparametric regression models that includes the generalised partially linear model. We advocate using estimators that are robust in the sense that they enjoy $\sqrt{n}$-consistent estimation uniformly over a sufficiently rich class of distributions characterised by certain conditional expectations being estimable by user-chosen machine learning methods. We show that even asking for locally uniform estimation within such a class narrows down possible estimators to those parametrised by certain weight functions. Conversely, we show that such estimators do provide the desired uniform consistency and introduce a novel random forest-based procedure for estimating the optimal weights. We prove that the resulting estimator recovers a notion of **ro**bust **s**emiparametric **e**fficiency (ROSE) and provides a practical alternative to semiparametric efficient estimators. We demonstrate the effectiveness of our ROSE random forest estimator in a variety of semiparametric settings on simulated and real-world data.

References

[1] Elliot Young, Rajen Shah, *ROSE Random Forests for Robust Semiparametric Efficient Estimation*, arXiv:2410.03471, 2024.

[2] Elliot Young, Rajen Shah, *Sandwich Boosting for Accurate Estimation in Partially Linear Models for Grouped Data*, J. Roy. Statist. Soc., Ser. B. **86**(5) (2024), 1286–1311

Regularized Empirical risk Minimization under Heavy Tailed Noise

Nicole Mücke
(joint work with Mattes Mollenhauer, Dimitri Meunier, Arthur Gretton)

Given two random variables X and Y, we seek to empirically minimize the expected squared error

$$R(f) := \mathbb{E}\left[(Y - f(X))^2\right] \tag{1}$$

over functions f in a reproducing kernel Hilbert space $\mathcal{H}$ consisting of functions from a topological space $\mathcal{X}$ to $\mathbb{R}$.

We consider the standard $L^2(P)$-orthogonal decomposition of Y with respect to the closed subspace $L^2(P, \sigma(X)) \subset L^2(P)$ of $\sigma(X)$-measurable functions, given by

$$Y = f^\star(X) + \varepsilon \tag{2}$$

with the regression function $f^\star(X) = \mathbb{E}[Y \mid X] \in L^2(P)$ and noise $\varepsilon \in L^2(P)$ satisfying $\mathbb{E}[\varepsilon \mid X] = 0$.

Given n independent sample pairs (X_i, Y_i) from the joint distribution of X and Y, we investigate the classical ridge regression estimate

$$\hat{f}_\alpha := \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \|Y_i - f(X_i)\|^2 + \alpha \|f\|_{\mathcal{H}}^2 \right\} \tag{3}$$

with regularization parameter $\alpha > 0$.

We adopt the well-known perspective going back to pathbreaking work by [1, 2, 3], which characterizes $\hat{f}_\alpha$ as the solution of a linear inverse problem in $\mathcal{H}$ obtained by performing Tikhonov regularization on a stochastic discretization of the integral operator induced by $\mathcal{H}$.

Since its inception, this setting has been refined and generalized in a vast variety of ways ranging from additive models and spectral regularization to kernel PCA and stochastic approximation methods. A common theme in this line of work is the derivation of confidence bounds for the excess risk

$$R(\hat{f}_\alpha) - R(f^\star) = \mathbb{E}\left[(\hat{f}_\alpha(X) - f^\star(X))^2\right] \tag{4}$$

under appropriate regularity assumptions about the smoothness of $f^\star$ and properties of the noise ε over the draw of the sample pairs.

Heavy-tailed noise. In this work, we assume that the real-valued random variable ε has only a finite number of higher conditional absolute moments, i.e., there exists some $q \geq 3$ such that

$$\mathbb{E}[|\varepsilon|^q \mid X] < Q < \infty \quad \text{almost surely.} \tag{5}$$

In such a setting, the family of Fuk–Nagaev inequalities [4] gives sharp nontrivial tail bounds beyond Markov's inequality for sums of heavy-tailed real random variables. In particular, these results show that the tail is dominated by a subgaussian term in a small deviation regime (reflecting the central limit theorem) and a polynomial term in a large deviation regime.

Just as in the light-tailed setting, we show that the optimal excess risk is achieved by balancing the contributions of the approximation error (e.g., the model bias) and the sample error (e.g., the model variance) by choosing a suitable regularization parameter α depending on n and δ.

The key difference to the known results for bounded or subexponential noise is a Fuk–Nagaev term appearing in the sample error, which introduces a regime with an additional polynomial dependence on δ and n.

In the low confidence regime, we can recover excess risk bounds similar to the setting with bounded or subexponential noise, i.e., they exhibit a logarithmic dependence on the confidence parameter δ and a sample size dependence up to $n^{-1/3}$.

The high confidence regime will require significantly stronger regularization than the low confidence setting. The resulting bound exhibits a polynomial worst-case dependence on δ, which is compensated by a better dependence on the sample size in terms of $n^{-(q-1)/q}$, before transitioning to a similar behavior as in the low confidence regime.

References

[1] F. Bauer, S. Pereverzev and L. Rosasco, *On regularization algorithms in learning theory*, Journal of Complexity **23** (2007), 52–72

[2] S. Smale and D.-X. Zhou, *Learning Theory Estimates via Integral Operators and Their Approximations*, Constructive Approximation **26**(2) (2007), 153–172

[3] G. Blanchard and Nicole Mücke, *Optimal Rates for Regularization of Statistical Inverse Learning Problems*, Foundations of Computational Mathematics **18** (2018), 971–1013

[4] D. H. Fuk and S. V. Nagaev, *Probability inequalities for sums of independent random variables*, Theory of Probability and its Applications **16** (1971), 643–660

Gradient optimization methods: large step-sizes and implicit bias

Peter Bartlett

(joint work with Yuhang Cai, Michael Lindsey, Song Mei, Matus Telgarsky, Jingfeng Wu, Bin Yu and Kangjie Zhou)

Optimization in deep learning relies on simple gradient descent algorithms. Although these methods are traditionally viewed as a time discretization of gradient flow, in practice, large step sizes - large enough to cause oscillation of the loss - exhibit performance advantages. We first consider gradient descent (GD) with a constant stepsize applied to logistic regression with linearly separable data, where the constant stepsize η is so large that the loss initially oscillates. We show that GD exits this initial oscillatory phase rapidly – in $O(\eta)$ steps – and subsequently achieves an $O(1/(\eta t))$ convergence rate after t additional steps. Our results imply that, given a budget of T steps, GD can achieve an accelerated loss of $O(1/T^2)$ with an aggressive stepsize $\eta = \Theta(T)$, without any use of momentum or variable stepsize schedulers. Our proof technique also handles general classification loss functions (where exponential tails are needed for the $O(1/T^2)$ acceleration), nonlinear predictors in the neural tangent kernel regime, and online stochastic gradient descent (SGD) with a large stepsize, under suitable separability conditions. Second, we consider this phenomenon in two-layer networks. The typical training of neural networks using large stepsize GD under the logistic loss often also exhibits two distinct phases, where the empirical risk oscillates in the first phase but decreases monotonically in the second phase. We investigate this phenomenon in two-layer networks that satisfy a near-homogeneity condition. We show that the second phase begins once the empirical risk falls below a certain threshold, dependent on the stepsize. Additionally, we show that the normalized margin grows nearly monotonically in the second phase, demonstrating an implicit bias of GD in training non-homogeneous predictors. If the dataset is linearly separable and the derivative of the activation function is bounded away from zero, we show that the average empirical risk decreases, implying that the first phase must stop in finite steps. Finally, we demonstrate that by choosing a suitably large stepsize, GD that undergoes this phase transition is more efficient than GD that monotonically decreases the risk. This analysis applies to networks of any width, beyond the well-known neural tangent kernel and mean-field regimes. Third, we establish the asymptotic implicit bias of GD for generic non-homogeneous deep networks under exponential loss. Specifically, we characterize three key properties of GD iterates starting from a sufficiently small empirical risk, where the threshold is determined by a measure of the network's non-homogeneity. First, we show that a normalized margin induced by the GD iterates increases nearly monotonically. Second, we prove that while the norm of the GD iterates diverges to infinity, the iterates themselves converge in direction. Finally, we establish that this directional limit satisfies the Karush-Kuhn-Tucker conditions of a margin maximization problem. Prior works on implicit bias have focused on homogeneous networks, where scaling inputs by a positive constant factor leads to a polynomial scaling of outputs in that factor. In contrast, our results apply to a broad class of non-homogeneous networks satisfying a mild near-homogeneity condition. In particular, our results apply to networks with many of the typical features of modern machine learning architectures, including residual connections and non-homogeneous activation functions.

To Intrinsic Dimension and Beyond: Efficient Sampling in Diffusion Models

Yuting Wei

(joint work with Zhihan Huang, Gen Li, Yuxin Chen, Yuejie Chi)

Diffusion models have garnered significant attention for their remarkable generative capabilities, producing high-quality samples with enhanced stability. Compared to methods like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), which generate samples in a single forward pass, diffusion models are designed to iteratively denoise samples over hundreds or thousands of steps. A prominent example is the widely used Denoising Diffusion Probabilistic

Models (DDPM) sampler. The current theory suggests the number of denoising steps required for accurate sample generation should scale at least linearly with the data dimension in order to learn the distribution accurately. While various acceleration schemes have been proposed in literature, in practical applications such as high-resolution image synthesis, where the dimensionality of the data can be extremely large, DDPM often requires far fewer steps than predicted by theory while maintaining excellent sample quality.

This gap between theoretical complexity bounds and empirical performance has inspired a strand of recent research, investigating whether diffusion models have implicitly exploited structural properties of real-world data to circumvent worst-case complexity bounds. This talk explores two key scenarios: (1) For a broad class of data distributions with intrinsic dimension k, we prove that the iteration complexity of the DDPM scales nearly linearly with k, which is optimal under the KL divergence metric; (2) For mixtures of Gaussian distributions with k components, we show that DDPM learns the distribution with iteration complexity that grows only logarithmically in k. These results provide theoretical justification for the practical efficiency of diffusion models.

References

[1] G. Li, Y. Wei, Y. Chen, and Y. Chi, *Towards faster non-asymptotic convergence for diffusion-based generative models*, International Conference on Learning Representations, 2024.

[2] G. Li, Y. Wei, Y. Chi, and Y. Chen, *A sharp convergence theory for the probability flow odes of diffusion models*, arXiv:2408.02320, 2024.

[3] Z. Huang, Y. Wei, Y. Chen, *Denoising diffusion probabilistic models are optimally adaptive to unknown low dimensionality*, arXiv:2410.18784, 2024.

[4] G. Li, C. Cai, Y. Wei, *Dimension-free convergence of diffusion models for approximate Gaussian mixtures*, 2025.

Optimal Convex M-Estimation via Score Matching

Min Xu

(joint work with Oliver Feng, Yu-Chun Kao, Richard Samworth)

In the context of linear regression, we construct a data-driven convex loss function with respect to which empirical risk minimisation yields optimal asymptotic variance in the downstream estimation of the regression coefficients. More precisely, given independent observations $\{(X_i, Y_i)\}_{i=1}^n$ with $Y_i = X_i^\top \beta_0 + \varepsilon_i$, X_i taking value on $\mathbb{R}^d$, and ε_i being independent of X_i and having a distribution P_0, we consider estimators of the form

$$\hat{\beta}_\ell = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \sum_{i=1}^n \ell(Y_i - X_i^\top \beta)$$

where $\ell : \mathbb{R} \to \mathbb{R}$ is a convex loss function. Writing $\psi = -\ell'$ and $\hat{\beta}_\psi \equiv \hat{\beta}_\ell$, it is known that under regularity conditions, if $\mathbb{E}\psi(\varepsilon_1) = 0$, then

$$\sqrt{n}(\hat{\beta}_\psi - \beta_0) \xrightarrow{d} N\biggl(0, \frac{\mathbb{E}\psi^2(\varepsilon_1)}{\{\mathbb{E}\psi'(\varepsilon_1)\}^2}\mathbb{E}(XX^\top)^{-1}\biggr).$$

Noting that only the term $V_{P_0}(\psi) := \frac{\mathbb{E}\psi^2(\varepsilon_1)}{\{\mathbb{E}\psi'(\varepsilon_1)\}^2}$ depends on ψ, we study the minimization of $V_{P_0}(\psi)$ over the set of all square integrable non-increasing functions. When P_0 has a uniformly continuous density p_0, we prove that the minimizer ψ_0^* is the best decreasing approximation of the derivative of the log-density of the noise distribution in $L_2(P_0)$ and give an explicit characterization

$$\psi_0^* = \mathsf{LCM}(p_0 \circ F_0^{-1})^{(R)} \circ F_0$$

where F_0 is the distribution function of p_0, F_0^{-1} is the quantile function, $f \mapsto \mathsf{LCM}(f)^{(R)}$ is the right-derivative of the least concave majorant of a function f. Our construction is based on a nonparametric extension of score matching, corresponding to a log-concave projection of the noise distribution with respect to the Fisher divergence.

We develop a finite sample procedure to estimate the optimal ψ_0^* and then to estimate the regression coefficients using the induced convex loss. Our procedure is computationally efficient, and we prove that it attains the minimal asymptotic covariance among all convex M-estimators. As an example of a non-log-concave setting, for Cauchy errors, the optimal convex loss function is Huber-like, and our procedure yields an asymptotic efficiency greater than 0.87 relative to the oracle maximum likelihood estimator of the regression coefficients that uses knowledge of this error distribution; in this sense, we obtain robustness without sacrificing much efficiency. Numerical experiments using our accompanying `R` package `asm` confirm the practical merits of our proposal.

Parallelized Midpoint Randomization for Langevin Monte Carlo

Arnak Dalalyan
(joint work with Lu Yu)

Let the function $f : \mathbb{R}^p \to \mathbb{R}$, referred to as the potential, be such that $\int_{\mathbb{R}^p} e^{-f(\theta)}\, d\theta$ is finite. We call target distribution the probability distribution having the probability density function

$$\pi(\theta) \propto \exp\{-f(\theta)\}, \qquad \theta \in \mathbb{R}^p. \tag{1}$$

The goal of sampling is to devise an algorithm that generates a random vector in $\mathbb{R}^p$ from a distribution which is close to the target one. Throughout the paper, we assume that the potential function f is M-smooth and m-strongly convex for some constants $m, M \in (0, \infty)$ such that $m \leq M$.

Traditional sampling methods often involve sequential processes, which may become computationally burdensome for large datasets or complex models. Parallel computing addresses this challenge by distributing the workload across multiple processors, enabling the concurrent execution of sampling tasks and enhancing computational efficiency, thus accelerating the generation of samples in statistical applications. Building upon the foundations laid by [1], we explore parallel computing for the midpoint randomization method in Langevin Monte Carlo [2, 3]. Our contributions can be summarized as follows.

- We introduce a parallel computing scheme for the randomized midpoint method applied to Langevin Monte Carlo and derive the corresponding convergence guarantees in Wasserstein distance, providing explicit constants and dependence on the initialization and choice of the parameters.
- We also present a comprehensive analysis of the parallel computing for the randomized midpoint method applied to kinetic Langevin Monte Carlo. Compared to previous work, our results offer **a)** small constants and the explicit dependence on the initialization, **b)** does not require the initialization to be at minimizer of the potential, **c)** removes the linear dependence on the sample size, which serves as a crucial step towards extending the method to non-convex potentials.

References

[1] R. Shen, Y. T. Lee, The randomized midpoint method for log-concave sampling, Advances in Neural Information Processing Systems 32 (2019).

[2] Y. He, K. Balasubramanian, M. A. Erdogdu, On the ergodicity, bias and asymptotic normality of randomized midpoint sampling method, Advances in Neural Information Processing Systems 33 (2020) 7366–7376.

[3] L. Yu, A. Karagulyan, A. S. Dalalyan, Langevin monte carlo for strongly log-concave distributions: Randomized midpoint revisited, in: The Twelfth International Conference on Learning Representations.

Statistical Algorithms for Low-Frequency Diffusion Data: A PDE Approach

Sven Wang

(joint work with Matteo Giordano)

This extended abstract summarises the main contributions of our work [1], which develops new computational techniques for statistical inference from low-frequency diffusion data. We consider the problem of statistical inference for multi-dimensional diffusion processes from low-frequency observations. In this setting, traditional likelihood-based methods are notoriously difficult to implement due to the intractability of the transition densities and their gradients. Motivated by these challenges, we develop a novel computational approach that builds on the theory of partial differential equations (PDEs) and leverages spectral techniques for elliptic operators. Our approach is based on the characterisation of the transition densities of the underlying reflected diffusion process as solutions of the associated Fokker–Planck equation with Neumann boundary conditions. Using regularity results from parabolic PDE theory [5], we derive a new representation for the gradient of the likelihood with respect to the unknown diffusivity function. This representation expresses the derivative through a variation-of-constants formula, see also [3], and allows us to avoid the need for costly data augmentation schemes often employed in the analysis of low-frequency diffusion data. Crucially, both the transition densities and their gradients can be approximated via the spectral decomposition of the elliptic generator of the diffusion, a self-adjoint operator in divergence form. This reduces the problem to the numerical solution of standard elliptic eigenvalue problems, for which efficient finite element solvers are available. Our approach thus enables the use of a wide range of statistical algorithms, including gradient-based optimisation methods and gradient-informed Markov chain Monte Carlo (MCMC) samplers. We demonstrate these developments in a nonparametric Bayesian framework using Gaussian process priors [4]. The resulting algorithms allow for the computation of maximum likelihood and maximum a posteriori estimates, posterior means, and quantiles, all without resorting to trajectory simulation or latent variable augmentation. In extensive simulation studies on a two-dimensional domain, our methods show excellent empirical performance, providing accurate reconstruction of the diffusivity function and competitive runtimes even at high sample sizes. Our work opens up several avenues for future research. These include extensions to diffusions with non-divergence form structure, models with noisy observations, and sampling on unbounded domains. Moreover, the PDE-based gradient characterisation may pave the way for a theoretical analysis of the computational complexity of the employed statistical algorithms, such as proving stability bounds and polynomial-time computability [2].

References

[1] Giordano, M. and Wang, S. (2025). Statistical algorithms for low-frequency diffusion data: A PDE approach. *The Annals of Statistics*, to appear.

[2] Nickl, R. and Wang, S. (2024). On polynomial-time computation of high-dimensional posterior measures by Langevin-type algorithms. *Journal of the European Mathematical Society*, **26**, 1031–1112.

[3] Wang, S. (2019). The nonparametric LAN expansion for discretely observed diffusions. *Electronic Journal of Statistics*, **13**, 1329–1358.

[4] Nickl, R. (2024). Consistent inference for diffusions from low frequency measurements. *The Annals of Statistics*, **52**, 519–549.

[5] Lunardi, A. (1995). *Analytic Semigroups and Optimal Regularity for Parabolic Problems.* Birkhäuser.

Rate optimality and phase transition for user-level local differential privacy

Thomas B. Berrett

(joint work with Alexander Kent and Yi Yu)

Most of the existing literature on differential privacy considers the *item-level* setting where each individual in a survey contributes a single data point to a dataset. Here, privacy mechanisms are developed so that, based only on the output of the mechanism, input datasets differing in only a single data point cannot be reliably distinguished. More precisely, writing $d_{\mathrm{H}}(x, x')$ for the Hamming distance between datasets $x, x' \in \mathcal{X}^n$, a privacy mechanism $Q(\cdot|x)$ must satisfy

$$\sup_{A} \sup_{\substack{x,x' \in \mathcal{X}^n \\ d_{\mathrm{H}}(x,x') \leq 1}} \frac{Q(A|x)}{Q(A|x')} \leq e^{\alpha}$$

to be called α-differentially private, where the first supremum is taken over all measurable sets in the output space. However, in many leading implementations of differentially private methodology, such as in mobile devices, each individual holds multiple data points. The naive application of privacy mechanisms satisfying standard, item-level privacy constraints is insufficient to protect the privacy of individuals in these settings, where repeated measurements may lead to *privacy leakage*.

In this work we study the *user-level* variant of differential privacy, where each of n individuals holds a dataset of size T. Writing $d_{\mathrm{H}}^T(y, y')$ for the Hamming distance between datasets $y, y' \in \mathcal{Y}^n$, where $\mathcal{Y} = \mathcal{X}^T$, privacy mechanisms here must satisfy the stronger constraint that

$$\sup_{A} \sup_{\substack{y,y' \in \mathcal{Y}^n \\ d_{\mathrm{H}}^T(y,y') \leq 1}} \frac{Q(A|y)}{Q(A|y')} \leq e^{\alpha},$$

so that a user's entire dataset can be perturbed without significantly affecting the output of the mechanism. We work under *local* privacy constraints, meaning that we do not assume the existence of a trusted central data handler.

We derive a general minimax lower bound, which shows that, for locally private user-level estimation problems, the risk cannot, in general, be made to vanish when n is fixed even when each T grows arbitrarily large. We then derive matching, up to logarithmic factors, lower and upper bounds for univariate and multidimensional mean estimation, sparse mean estimation and non-parametric density estimation. In particular, with other model parameters held fixed, we observe phase transition phenomena in the minimax rates as T varies.

In the case of (non-sparse) mean estimation and density estimation, we see that, for T below a phase transition boundary, the rate is the same as having nT users in the item-level setting. Different behaviour is however observed in the case of s-sparse d-dimensional mean estimation, wherein consistent estimation is impossible when d exceeds the number of observations in the item-level setting, but is possible in the user-level setting when $T \gg s \log(d)$, up to logarithmic factors. This may be of independent interest for applications as an example of a high-dimensional problem that is feasible under local privacy constraints.

Optimal estimation in private distributed functional data analysis

Yi Yu
(joint work with Gengyu Xue and Zhenhua Lin)

We systematically investigate the preservation of differential privacy in functional data analysis, beginning with functional mean estimation and extending to varying coefficient model estimation. Our work introduces a distributed learning framework involving multiple servers, each responsible for collecting several sparsely observed functions. This hierarchical setup introduces a mixed notion of privacy. Within each function, user-level differential privacy is applied to m discrete observations. At the server level, central differential privacy is deployed to account for the centralised nature of data collection. Across servers, only private information is exchanged, adhering to federated differential privacy constraints. To address this complex hierarchy, we employ minimax theory to reveal several fundamental phenomena: from sparse to dense functional data analysis, from user-level to central and federated differential privacy costs, and the intricate interplay between different regimes of functional data analysis and privacy preservation.

To the best of our knowledge, this is the first study to rigorously examine functional data estimation under multiple privacy constraints. Our theoretical findings are complemented by efficient private algorithms and extensive numerical evidence, providing a comprehensive exploration of this challenging problem.

A simple and improved algorithm for noisy, convex, zeroth-order optimisation

Alexandra Carpentier

We consider the setting of convex noisy zeroth-order optimisation. For $d \geq 1$, consider a bounded convex set $\bar{\mathcal{X}} \subset \mathbb{R}^d$ with non-zero volume, and consider a convex function $f : \bar{\mathcal{X}} \to [0,1]$.

We consider a sequential setting with fixed horizon $n \in \mathbb{N} \setminus \{0\}$. At each time $t \leq n$, the learner chooses a point $x_t \in \bar{\mathcal{X}}$ and observes a noisy observation $y_t \in [0,1]$ that is such that $\mathbb{E}[y_t|x_t = x] = f(x)$, and such that y_t knowing x_t is independent of the past observations.

We study the problem of optimising the function f in the sequential game described above, namely after the budget n has been fully used by the learner, she has to predict a point $\hat{x}$ - based on all her observations $(x_t, y_t)_{t \leq n}$ - and her aim will be to estimate the minimum for the function f. Her performance for this task will be measured through the following (simple) *regret*

$$f(\hat{x}) - \inf_{x \in \bar{\mathcal{X}}} f(x),$$

namely the difference between the true infimum of f, and f evaluated at $\hat{x}$.

This setting is known as convex noisy zeroth-order optimisation [3, 2]. In this talk, we discussed the literature in that topic - starting from dimension $d = 1$ - and then considering the challenges in the higher dimensional setting. We then presented shortly the simple algorithm from [1], based on the centre of gravity method, which has a worst-case upper bound on the simple regret of order $\frac{d^2}{\sqrt{n}}$, up to logarithmic terms. While this is not anymore state of the art - see [2] where $\frac{d^{1.5}}{\sqrt{n}}$ is achieved, which has to be compared with the best known lower bound of order $\frac{d}{\sqrt{n}}$ - the main interest of this method is its relative simplicity - also that of its analysis.

References

[1] A. Carpentier, *A simple and improved algorithm for noisy, convex, zeroth-order optimisation*, arXiv:2406.18672, 2024.

[2] T. Lattimore, *Bandit Convex Optimisation*, arXiv:2402.06535, 2024.

[3] A. Nemirovskij and D. Yudin, *Problem complexity and method efficiency in optimization*, Wiley-Interscience, 1983.

Sampling Binary Data by Denoising through Score Functions

Francis Bach
(joint work with Saeed Saremi)

Gaussian smoothing combined with a probabilistic framework for denoising via the empirical Bayes formalism, i.e., the Tweedie-Miyasawa formula (TMF), are the two key ingredients in the success of score-based generative models in Euclidean spaces. Smoothing holds the key for easing the problem of learning and sampling in high dimensions, denoising is needed for recovering the original signal, and TMF ties these together via the score function of noisy data. In this work, we extend this paradigm to the problem of learning and sampling the distribution of binary data on the Boolean hypercube by adopting Bernoulli noise, instead of Gaussian noise, as a smoothing device. We first derive a TMF-like expression for the optimal denoiser for the Hamming loss, where a score function naturally appears. Sampling noisy binary data is then achieved using a Langevin-like sampler which we theoretically analyze for different noise levels. At high Bernoulli noise levels sampling becomes easy, akin to log-concave sampling in Euclidean spaces. In addition, we extend the sequential multi-measurement sampling of Saremi et al. (2024) to the binary setting where we can bring the "effective noise" down by sampling multiple noisy measurements at a fixed noise level, without the need for continuous-time stochastic processes. We validate our formalism and theoretical findings by experiments on synthetic data and binarized images.

References

[1] Francis Bach, Saeed Saremi. Sampling Binary Data by Denoising through Score Functions. Technical report, arXiv:2502.00557, 2025.

[2] Saeed Saremi, Ji-Won Park, Francis Bach. Chain of Log-Concave Markov Chains. Proceedings of the International Conference on Learning Representations (ICLR), 2024.

Estimation beyond missing (completely) at random

Kabir Aladin Verchand
(joint work with Tianyi Ma, Thomas B. Berrett, Tengyao Wang,
and Richard J. Samworth)

We study the effects of missingness on the estimation of population parameters. Moving beyond restrictive missing completely at random (MCAR) assumptions, we formulate two models of departures from missing completely at random. First, we consider a missing data analogue of Huber's ϵ contamination model. For mean estimation with respect to squared error, we show that the minimax quantiles decompose as a sum of the corresponding minimax quantiles under a heterogeneous, MCAR assumption, and a robust error term, depending on ϵ, that reflects the additional error incurred by departure from MCAR:

$$\inf_{\widehat{\theta}\in\widehat{\Theta}} \sup_{\theta\in\mathbb{R}^d} \sup_{P_\theta\in\mathcal{P}_\theta^{\mathrm{Huber}}} \mathrm{Quantile}(1-\delta; P_\theta, \|\widehat{\theta}-\theta\|_2^2)$$
$$\asymp \underbrace{\frac{\mathrm{Tr}(\Sigma^{\mathrm{IPW}})}{n} + \frac{\|\Sigma^{\mathrm{IPW}}\|_{\mathrm{op}}\log(1/\delta)}{n}}_{\text{MCAR term}} + \underbrace{\|\Sigma^{\mathrm{IPW}}\|_{\mathrm{op}}\epsilon}_{\text{MCAR departure}} .$$

In order to achieve this rate, we develop an iterative imputation algorithm which can be layered on top of existing (complete-case) robust mean estimation algorithms.

Unfortunately, Huber's contamination model contains corruptions which may be overly pessimistic, and the estimation rate suffers accordingly. Motivated by this, we next introduce natural classes of realizable ϵ-contamination models, where an MCAR version of a base distribution P is contaminated by an arbitrary missing not at random (MNAR) version of P . These classes are rich enough to capture various notions of biased sampling and sensitivity conditions, yet we show that they enjoy improved minimax performance relative to our earlier arbitrary contamination classes for both parametric and nonparametric classes of base distributions. For instance, with a univariate Gaussian base distribution, consistent mean estimation over realizable ϵ-contamination classes is possible even when ϵ and the proportion of missingness converge (slowly) to 1. In particular, we have:

$$\inf_{\widehat{\theta}\in\widehat{\Theta}} \sup_{\theta\in\mathbb{R}^d} \sup_{P_\theta\in\mathcal{P}_\theta^{\mathrm{Realizable}}} \mathrm{Quantile}(1-\delta; P_\theta, |\widehat{\theta}-\theta|^2)$$
$$\asymp \underbrace{\frac{\sigma^2\log(1/\delta)}{n}}_{\text{MCAR term}} + \underbrace{\frac{\sigma^2\log^2\bigl(1+\frac{\epsilon}{q(1-\epsilon)}\bigr)}{\log\bigl(nq(1-\epsilon)\bigr)}}_{\text{Realizable departure}},$$

except for ϵ contained in an interval of size $o_n(1)$.

Multi-resolution subsampling for linear classification with massive data

Holger Dette
(joint work with Haolin Chen and Jun Yu)

Classification is one of the main tasks in data analysis and numerous classification algorithms have been developed in statistics and machine learning. Often training a classifier on a massive dataset is challenging due large computational costs, even for linear classifiers. Moreover, as pointet out in [Schwartz et al., 2020], the ever-increasing demand for high computing power has negative environmental impacts such as carbon emissions suffer from heavy computer usage.

To tackle the challenges of limiting computing resources, data scientists have to balance statistical accuracy and computational costs and one of the ubiquitous solutions is subsampling. Several authors have demonstrated that subsampling can achieve this goal in many real-world applications. For example, [Wang et al., 2021a] analyze the click-through rate for ByteDance Apps via nonuniform negative subsampling techniques and [Wang et al., 2022] use subsampling to predict customer churn for a security company in China.

We consider subsampling techniques exploiting model information to identify data points in the sample that yield most precise parameter estimates, see for instance, [Ma et al., 2022, Wang et al., 2021b, Wang et al., 2018, Ai et al., 2021] and [Zhang et al., 2024] among many others. We demonstrate that a multi-resolution optimal subsampling method combining summary measures and selected subdata points yields a statistically and computationally efficient linear classifiers which improves the current state of the art in the general linear classification problem substantially. This improvement comes from two aspects. One the one hand, we use the information from the selected subdata and the summary measures to collect the information from the unselected data points. On the other hand, we carefully extricate ourselves from the common point of view that subsampling should reflect the information of the entire data. More specifically, we propose to use sampling techniques for the region we focus on and to use summary measures to collect the information for the rest. As a consequence, we can prove that the resulting estimators become (asymptotically) more efficient and stable compared to other approaches.

References

[Ai et al., 2021] M. Ai, J. Yu, H. Zhang, and H. Wang, *Optimal subsampling algorithms for big data regressions*, Statistica Sinica **31** (2021), 749–772.

[Ma et al., 2015] P. Ma, M. W. Mahoney, and B. Yu, *A statistical perspective on algorithmic leveraging*, Journal of Machine Learning Research **16** (2015), 861–911.

[Ma et al., 2022] P. Ma, Y. Chen, X. Zhang, X. Xing, J. Ma, and M. W. Mahoney, *Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms*, Journal of Machine Learning Research **23** (2022), 1–45.

[Schwartz et al., 2020] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, *Green AI*, Communications of the ACM **63** (2020), pp 54–63.

[Wang et al., 2018] H. Wang, R. Zhu, and P. Ma, *Optimal subsampling for large sample logistic regression*, Journal of the American Statistical Association **113** (2018), 829–844.

[Wang et al., 2021b] L. Wang, J. Elmstedt, W. K. Wong, and H. Xu, *Orthogonal subsampling for big data linear regression*, The Annals of Applied Statistics **15** (2021), 1273–1290.

[Wang et al., 2021a] H. Wang, A. Zhang, and C. Wang, *Nonuniform negative sampling and log odds correction with rare events data*, Advances in Neural Information Processing Systems **34** (2021), 19847–19859.

[Wang et al., 2022] F. Wang, D. Huang, T. Gao, S. Wu, and H. Wang, *Sequential one-step estimator by subsampling for customer churn analysis with massive data sets*, Journal of the Royal Statistical Society Series C: Applied Statistics **71** (2022), 1753–1786.

[Zhang et al., 2024] Y. Zhang, L. Wang, X. Zhang, and H. Wang, *Independence-encouraging subsampling for nonparametric additive models*, Journal of Computational and Graphical Statistics **33**(4) (2024), 1424–1433.

A conversion theorem and minimax optimality for continuum contextual bandits

Alexandre B. Tsybakov
(joint work with Arya Akhavan, Karim Lounici, Massimiliano Pontil)

The contextual bandit problem has been extensively studied in finite action spaces, where algorithms leveraging side information or context have achieved strong performance guarantees, see, e.g., [3]. However, many real-world applications, such as personalized recommendations, control systems, and experimental design, naturally involve continuous action spaces, giving rise to the more general setting of contextual continuum bandits. Extending contextual bandits to continuous domains raises new challenges [5], as the learner must explore an infinite action space and infer the structure of the loss function from limited feedback, often relying on convexity or smoothness assumptions to ensure tractability.

Contextual continuum bandits. Let $\Theta \subseteq \mathbb{R}^d$ be a convex body, and let $f : \mathbb{R}^d \times [0,1]^p \to \mathbb{R}$ be an unknown function. At each round t, a context $\boldsymbol{c}_t \in [0,1]^p$ is revealed by the adversary. Then, based on $\boldsymbol{c}_t$ and the past values $(y_k, \boldsymbol{z}_k, \boldsymbol{c}_k)_{k=1}^{t-1}$ the learner chooses a query point $\boldsymbol{z}_t \in \Theta$ and receives a noisy evaluation:

$$y_t = f(\boldsymbol{z}_t, \boldsymbol{c}_t) + \xi_t, \tag{1}$$

where ξ_t is a scalar noise variable. The learner's objective is to achieve the smallest possible **contextual regret** defined as

$$R_T^{\mathrm{cntx}}(f) := \mathbf{E}\Big[\sum_{t=1}^{T} f(\boldsymbol{z}_t, \boldsymbol{c}_t) - \sum_{t=1}^{T} \min_{\boldsymbol{z}\in\Theta} f(\boldsymbol{z}, \boldsymbol{c}_t)\Big]. \tag{2}$$

Meta-algorithm. We propose a meta-algorithm (Algorithm 1) for the continuum contextual bandit problem.

Let $\mathcal{F}$ be a class of functions $f : \mathbb{R}^d \times [0,1]^p \to \mathbb{R}$. Assume that there exists a possibly randomized online policy $\pi = (\pi_t)_{t=1}^{\infty}$ (the input algorithm), for which we can control its static regret over T runs for any sequence of functions $\{f(\cdot, \boldsymbol{c}_t)\}_{t=1}^{T}$, where $f \in \mathcal{F}$ and $\{\boldsymbol{c}_t\}_{t=1}^{T} \subseteq [0,1]^p$. Namely, if $\{\boldsymbol{z}_t^{\mathrm{input}}\}_{t=1}^{T}$ are updates of $(\pi_t)_{t=1}^{T}$,

Algorithm 1:

```
Input: Randomized policy $\pi = (\pi_t)_{t=1}^{\infty}$, parameter $K \in \mathbb{N}$, partition $\{B_i\}_{i=1}^{K^p}$ of $[0,1]^p$
Initialization: $N_i(0) = 0$, $H(i) = \{0\}$ for all $i = 1, \dots, K^p$, $\boldsymbol{z}_0 = \boldsymbol{0}$
for $t = 1, \dots, T$ do
    if $\boldsymbol{c}_t \in B_i$ then
        $N_i(t) = N_i(t-1) + 1$                                              // Increment count
        Use $\pi_{N_i(t)}$ with $\{y_k, \boldsymbol{z}_k, \boldsymbol{c}_k\}_{k \in H(i)}$ to choose $\boldsymbol{z}_t \in \Theta$     // Select query
        $y_t = f(\boldsymbol{z}_t, \boldsymbol{c}_t) + \xi_t$                                          // Query
        $H(i) \leftarrow H(i) \cup \{t\}$                                          // Update index set
    end
end
```

then there exist $F : [0,\infty) \to [0,\infty)$, $F_1 : [0,\infty) \to [0,\infty)$ such that F is concave, F_1 is non-decreasing, and for all $\{\boldsymbol{c}_t\}_{t=1}^T \subseteq [0,1]^p$ the **static regret** satisfies:

$$\sup_{f \in \mathcal{F}} \mathbf{E}\Big[\sum_{t=1}^{T} f(\boldsymbol{z}_t^{\text{input}}, \boldsymbol{c}_t) - \min_{\boldsymbol{z} \in \Theta} \sum_{t=1}^{T} f(\boldsymbol{z}, \boldsymbol{c}_t)\Big] \le F(T)F_1(T). \tag{3}$$

Let $\{B_i\}_{i=1}^{K^p}$ be the partition of $[0,1]^p$ into K^p equal cubes with edge length $1/K$. We assume that f is Hölder continuous w.r.t. $\boldsymbol{c}$. For $(L,\gamma) \in [0,\infty) \times (0,1]$ we denote by $\mathcal{F}_\gamma(L)$ the class of all functions $f : \mathbb{R}^d \times [0,1]^p \to \mathbb{R}$ such that

$$|f(\boldsymbol{x}, \boldsymbol{c}) - f(\boldsymbol{x}, \boldsymbol{c}')| \le L \left\|\boldsymbol{c} - \boldsymbol{c}'\right\|^{\gamma}, \quad \text{for all} \quad \boldsymbol{x} \in \Theta,\ \boldsymbol{c}, \boldsymbol{c}' \in [0,1]^p, \tag{4}$$

where $\|\cdot\|$ is the Euclidean norm.

Theorem 1 (Static-to-contextual regret conversion)**.** *Let $(L,\gamma) \in [0,\infty) \times (0,1]$. Let π be a randomized policy such that* (3) *holds with a concave function F and a non-decreasing function F_1, and let $\boldsymbol{z}_t$'s be the updates of Algorithm 1. Then,*

$$\sup_{f \in \mathcal{F} \cap \mathcal{F}_\gamma(L)} R_T^{\text{cntx}}(f) \le K^p F\left(\frac{T}{K^p}\right) F_1(T) + 2LT \left(\frac{\sqrt{p}}{K}\right)^{\gamma}. \tag{5}$$

Theorem 1 allows us to control the contextual regret of the output algorithm in terms of the static regret of the input algorithm π. As a corollary, we derive the minimax optimal rates and algorithms in the following three settings.

(a) **Lipschitz contextual bandits.** Lipschitz non-contextual bandits have been extensively studied in the literature, see [4] and the references therein. In [4], the authors propose an algorithm that attains a static regret of order $d^{\frac{1}{2}} T^{\frac{d+1}{d+2}} \log^5(T)$ for $\Theta = [0,1]^d$ and for noise-free observations $\xi_t = 0$ for all t over the class $\mathcal{F}$ of functions such that $|f(\boldsymbol{x}, \boldsymbol{c}) - f(\boldsymbol{x}', \boldsymbol{c})| \le L' \left\|\boldsymbol{x} - \boldsymbol{x}'\right\|$ for all $\boldsymbol{x}, \boldsymbol{x}' \in [0,1]^d$, $\boldsymbol{c} \in [0,1]^p$, where $L' > 0$ is a constant. Taking their method as the input π of Algorithm 1, applying Theorem 1 with optimal choice of K, and assuming for simplicity that $\gamma = 1$, we obtain the following bound on the contextual regret of our output procedure:

$$\sup_{f \in \mathcal{F} \cap \mathcal{F}_1(L)} R_T^{\text{cntx}}(f) \le C\sqrt{pd}\, T^{\frac{p+d+1}{p+d+2}} \log^5(T), \tag{6}$$

where $C > 0$ is a numerical constant. A minimax lower bound on the contextual regret of the order $T^{\frac{p+d+1}{p+d+2}}$ valid on the same class of functions is proved in [5]. Together with (6), it implies that, up to a logarithmic factor, the rate $T^{\frac{p+d+1}{p+d+2}}$ is minimax optimal as function of T and our algorithm attains the minimax rate.

(b) **Convex contextual bandits.** Let $\mathcal{F}$ be the class of functions f such that $f(\cdot, \boldsymbol{c})$ is convex for all $\boldsymbol{c} \in [0,1]^p$, and let ξ_t's be i.i.d. sub-Gaussian for all t. We use the state of the art result on non-contextual convex bandit optimization from [1]. It proposes a polynomial-time algorithm achieving, up to a poly-logarithmic factor, the static regret $d^{3.5}\sqrt{T}$. Taking this algorithm as the input policy π of Algorithm 1 and applying Theorem 1 with optimal choice of K we derive the following bound for the contextual regret of our output procedure:

$$\sup_{f \in \mathcal{F} \cap \mathcal{F}_\gamma(L)} R_T^{\text{cntx}}(f) \le C p^{1/2} d^{7/3} T^{\frac{p+\gamma}{p+2\gamma}}, \tag{7}$$

where $C > 0$ is a factor that depends polynomially on $\log(T)$ and $\log(d)$ and does not depend on p.

(c) **Strongly convex contextual bandits.** Consider the class $\mathcal{F} = \mathcal{F}_{\alpha,\beta}(M)$ of objective functions f such that, for any $\boldsymbol{c} \in [0,1]^p$, the map $\boldsymbol{x} \mapsto f(\boldsymbol{x}, \boldsymbol{c})$ is β-smooth, α-strongly convex and satisfies $\max_{\boldsymbol{x} \in \Theta} |f(\boldsymbol{x}, \boldsymbol{c})| \le M$ for some $\beta, \alpha, M > 0$. Under noisy observations with σ-sub-Gaussian noise, we first propose a non-contextual policy π close to the BCO algorithm of [2] and we prove that it satisfies a static regret bound of the form (3). Combining this policy with Algorithm 1, we obtain for $T \ge d$ the following bound on the contextual regret of the resulting procedure:

$$\sup_{f \in \mathcal{F}_{\alpha,\beta}(M) \cap \mathcal{F}_\gamma(L)} R_T^{\text{cntx}}(f) \le C p^{1/2} d T^{\frac{p+\gamma}{p+2\gamma}} \log(T), \tag{8}$$

where $C > 0$ depends only on $M, \sigma, \beta/\alpha$.

Minimax lower bound. Along with the classes $\mathcal{F}_\gamma(L)$, $\gamma \in (0,1]$, we consider the class $\mathcal{F}_0(L)$, which includes discontinuous functions and is defined as the set of all functions such that $|f(\boldsymbol{x}, \boldsymbol{c}) - f(\boldsymbol{x}, \boldsymbol{c}')| \le L$ for all $\boldsymbol{x} \in \Theta, \boldsymbol{c}, \boldsymbol{c}' \in [0,1]^p$.

Theorem 2. *Let $(\alpha, \beta, M, \gamma, L) \in (0,\infty) \times [3\alpha, \infty) \times [\alpha+1, \infty) \times [0,1] \times [0,\infty)$. Let Θ be the unit Euclidean ball, let $\{\boldsymbol{c}_t\}_{t=1}^T$ be independently distributed according to a suitably defined distribution on $[0,1]^p$, and $\{\xi_t\}_{t=1}^T$ be i.i.d. standard Gaussian random variables. If $\{\boldsymbol{z}_t\}_{t=1}^T$ are outputs of any randomized policy then*

$$\sup_{f \in \mathcal{F}_{\alpha,\beta}(M) \cap \mathcal{F}_\gamma(L)} R_T^{\text{cntx}}(f) \ge \mathtt{A}\left(\min\left(1, L^{\frac{2(p+\gamma)}{p+2\gamma}}\right) T^{\frac{p+\gamma}{p+2\gamma}} + \min\left(T, d\sqrt{T}\right)\right),$$

where $\mathtt{A} > 0$ is a numerical constant.

From (7), (8) and Theorem 2 we deduce that the minimax optimal rate in T for both settings (b) and (c) above is of the order $T^{\frac{p+\gamma}{p+2\gamma}}$ up to a logarithmic factor.

The bound of Theorem 2 with $L > 0$ and $\gamma = 0$ shows that no randomized policy can achieve sub-linear contextual regret on the corresponding class of functions.

Thus, controlling the increments of f with any $L > 0$ in the absence of continuity with respect to $\boldsymbol{c}$ is not sufficient to get a sub-linear contextual regret.

References

[1] H. Fokkema, D. van der Hoeven, T. Lattimore, and J. J. Mayo, *Online Newton method for bandit convex optimisation*, arXiv:2406.06506 (2024).

[2] E. Hazan and K. Levy, *Bandit convex optimization: Towards tight bounds*, Advances in Neural Information Processing Systems **27** (2014).

[3] J. Langford and T. Zhang, *The epoch-greedy algorithm for multi-armed bandits with side information*, Advances in Neural Information Processing Systems **20** (2007).

[4] C. Podimata and A. Slivkins, *Adaptive discretization for adversarial Lipschitz bandits*, In: Conference on Learning Theory (2021), 3788–3805.

[5] A. Slivkins, *Contextual bandits with similarity information*, J. of Machine Learning Research **15** (2014), 2533–2568.

Distributional Regression and Instrumental Variables

Nicolai Meinshausen

(joint work with Xinwei Shen, Anastasiia Holovchak, Sorawit Saengkyongam)

Distributional regression aims to estimate the full conditional distribution of a target variable, given covariates. Popular methods include linear and tree ensemble based quantile regression. We propose a neural network-based distributional regression methodology called 'engression'. An engression model is generative in the sense that we can sample from the fitted conditional distribution and is also suitable for high-dimensional outcomes. Furthermore, we find that modelling the conditional distribution on training data can constrain the fitted function outside of the training support, which offers a new perspective to the challenging extrapolation problem in nonlinear regression. In particular, for 'pre-additive noise' models, where noise is added to the covariates before applying a nonlinear transformation, we show that engression can successfully perform extrapolation under some assumptions such as monotonicity, whereas traditional regression approaches such as least-squares or quantile regression fall short under the same assumptions. Our empirical results, from both simulated and real data, validate the effectiveness of the engression method. In addition to these regression results in [3], we can also show that distributional reconstruction can be useful for dimensionality reduction [2] and instrumental variable regression [1]. The instrumental variable (IV) approach is commonly used to infer causal effects in the presence of unmeasured confounding. Conventional IV models commonly make the additive noise assumption, which is hard to ensure in practice, but also typically lack flexibility if the causal effects are complex. Further, the vast majority of the existing methods aims to estimate the mean causal effects only, a few other methods focus on the quantile effects. Here we aim for estimation of the entire interventional distribution. We propose a novel method called distributional instrumental variables (DIV) We establish identifiability of the interventional distribution under general assumptions and demonstrate an 'under-identified' case where DIV can identify the causal effects while two-step least squares fails to. Our empirical results show that the DIV method performs well for a broad range of simulated data, exhibiting advantages over existing IV approaches in terms of the identifiability and estimation error of the mean or quantile treatment effects.

References

[1] Anastasiia Holovchak, Sorawit Saengkyongam, Nicolai Meinshausen, Xinwei Shen *Distributional Instrumental Variable Method*, arXiv:2502.07641, 2025.

[2] Xinwei Shen, Nicolai Meinshausen *Distributional principal autoencoders*, arXiv:2404.13649, 2024

[3] Xinwei Shen, Nicolai Meinshausen *Engression: Extrapolation through the Lens of Distributional Regression*, Journal of the Royal Statistical Society, Series B (2024), 1–25

Minimax rate for multivariate data under componentwise local differential privacy constraints

Chiara Amorino
(joint work with Arnaud Gloter)

Our research analyses the balance between maintaining privacy and preserving statistical accuracy when dealing with multivariate data that is subject to *componentwise local differential privacy* (CLDP). With CLDP, each component of the private data is made public through a separate privacy channel. This allows for varying levels of privacy protection for different components or for the privatization of each component by different entities, each with their own distinct privacy policies. It also covers the practical situations where it is impossible to privatize jointly all the components of the raw data. We develop general techniques for establishing minimax bounds that shed light on the statistical cost of privacy in this context, as a function of the privacy levels $\alpha_1, \dots, \alpha_d$ of the d components.
We demonstrate the versatility and efficiency of these techniques by presenting various statistical applications. Specifically, we examine nonparametric density and joint moments estimation under CLDP, providing upper and lower bounds that match up to constant factors, as well as an associated data-driven adaptive procedure. Additionally, we conduct a detailed analysis of the effective privacy level, exploring how information about a private characteristic of an individual may be inferred from the publicly visible characteristics of the same individual.

On nonparametric estimation of the interaction function in particle system models

Mark Podolskij
(joint work with Denis Belomestny, Shi-Yuan Zhou)

This talk is dedicated to the study of N-dimensional $\mathbb{R}^d$-valued interacting particle systems described by the equation

$$dX_t^{i,N} = (\varphi \star \mu_t^N)(X_t^{i,N})\, dt + dW_t^i, \quad i = 1, \dots, N,$$

where $t \in [0,T]$, $\varphi\colon \mathbb{R}^d \to \mathbb{R}^d$ is the interaction potential and $(W^i)_{i=1}^N$ are independent d-dimensional Brownian motions. Here, μ_t^N stands for the empirical measure of the particle system at time t, given by

$$\mu_t^N := \frac{1}{N} \sum_{i=1}^N \delta_{X_t^{i,N}},$$

and $\varphi \star \mu(x) := \int \varphi(x-y)\mu(dy) \in \mathbb{R}^d$. We make the assumption that we observe N paths $(X_t^{i,N}, t \in [0,T], i = 1, \dots, N)$ and aim to estimate the unknown interaction function φ as $N \to \infty$ with $T > 0$ being fixed.

We assume the following condition: The interaction function $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ is globally Lipschitz continuous and bounded:

$$\|\varphi(x) - \varphi(y)\| \leq L_\varphi \|x - y\|, \quad \|\varphi(x)\| \leq K_\varphi, \quad x, y \in \mathbb{R}^d$$

for some finite $L_\varphi, K_\varphi > 0$. This condition guarantees the validity of propagation of chaos in the sense that, for all $t \in [0,T]$, μ_t^N converges weakly to μ_t (see, e.g., [2, Theorem 3.1]).

We initiate our exposition by describing the fundamental principles of our estimation methodology. To begin, we consider a sequence of spaces $(S_m)_{m \geq 1}$ of functions valued in $\mathbb{R}^d$, compact with respect to $\|\cdot\|_\infty$. The crucial idea of our approach lies in the following minimization strategy:

$$\min_{f \in S_N} \frac{1}{NT} \sum_{i=1}^N \int_0^T \left\| f \star \mu_t^N(X_t^{i,N}) - \varphi \star \mu_t^N(X_t^{i,N}) \right\|^2 dt.$$

Unfortunately, the above risk function cannot be directly computed from the data since the interaction function φ is unknown. We derive an empirical (noisy) version of the minimization problem by omitting the irrelevant term $\|\varphi \star \mu_t^N(X_t^{i,N})\|^2$ in the integrand and minimizing the resulting quantity:

$$\gamma_N(f) := \frac{1}{NT} \sum_{i=1}^N \Big(\int_0^T \|f \star \mu_t^N(X_t^{i,N})\|^2\, dt - 2 \int_0^T \Big\langle f \star \mu_t^N(X_t^{i,N}), dX_t^{i,N} \Big\rangle \Big)$$

over S_N. For further analysis, we introduce the following bilinear forms:

$$\langle f, g\rangle_N := \frac{1}{NT}\sum_{i=1}^{N}\int_0^T \Big\langle (f \star \mu_t^N)(X_t^{i,N}), (g \star \mu_t^N)(X_t^{i,N})\Big\rangle\, dt,$$

$$\langle f, g\rangle_\star := \frac{1}{T}\int_0^T \int_{\mathbb{R}^d} \Big\langle (f \star \mu_t)(x), (g \star \mu_t)(x)\Big\rangle \mu_t(x)\, dx\, dt.$$

We set $\|f\|_N^2 := \langle f, f\rangle_N$ and $\|f\|_\star^2 := \langle f, f\rangle_\star$. Finally, our estimator φ_N is defined as follows:

$$\varphi_N := \mathrm{argmin}_{f\in S_N}\gamma_N(f).$$

The main result of the talk is the following theorem.

Theorem. Let D_N denote the dimension of the functional space S_N satisfying $D_N^2 N^{-1/2} \to 0$ as $N \to \infty$. Then, for any $q \geq 2$ there exists a constant $C_q > 0$, such that

$$\Big\{\mathbb{E}\Big[\|\varphi_N - \varphi\|_\star^q\Big]\Big\}^{1/q} \leq \inf_{f\in S_N} \|f - \varphi\|_\star + C_q\sqrt{\frac{D_N}{N}}.$$

Due to the contractive properties of the norm $\|\cdot\|_\star$, the approximation error in the above theorem typically decays at an exponential rate. As a result, the convergence is often of parametric order, up to logarithmic factors. In contrast, the situation is more nuanced when considering convergence in the L^2-norm. Depending on the properties of the interaction function φ, the marginal distributions μ_t, and their characteristic functions, one may observe logarithmic or even polynomial L^2-rates of convergence. For a detailed analysis, we refer the reader to [1].

References

[1] D. Belomestny, M. Podolskij and S.-Y. Zhou, *On nonparametric estimation of the interaction function in particle system models*, arXiv:2402.14419, 2024.

[2] L.-P. Chaintron and A. Diez, *Propagation of chaos: a review of models, methods and applications. II. Applications*, Kinetic and Related Models **15**(6) (2022), 1017–1173.

Adaptive density estimation under low-rank constraints

Olga Klopp

(joint work with Julien Chhor and Alexandre B. Tsybakov)

In many applications, one needs to explore relations between two objects that may have a complex structure, yet are linked via a low-dimensional latent space. This situation can be often described by mixture models and low-rank matrix models. For the problem with discrete distributions, one of the important examples is given by the *probabilistic Latent Semantic Indexing* framework for topic models. It assumes that co-occurrences of words and documents are independent given one of K latent topic classes. Then the joint probability matrix of words and documents is a mixture of at most K matrices and its rank does not exceed K, which is typically a small number. Another example of low-rank probability matrix estimation is provided by the Stochastic Block Model. In this case, the problem is to estimate the matrix of connection probabilities of a random graph under the assumption that its nodes fall into K groups with constant connection probabilities within and between each two groups. Such a probability matrix is of rank at most K. Low-rank probability matrix estimation problems also arise in the context of collaborative filtering and matrix completion.

For the problems characterized by continuous probability densities, multi-view models provide a nonparametric analog of classical mixture models. In contrast to these classical models, they do not assume that the components of the mixture depend on finite number of parameters but rather consider them as functions satisfying some general constraints, such as smoothness or just integrability. Densities $f : \mathbb{R}^m \to \mathbb{R}$ satisfying the multi-view model are the form

$$f(x) = \sum_{i=1}^{K} w_i \prod_{j=1}^{m} f_{ij}(x^T e_j) \quad \text{with} \quad \sum_{i=1}^{K} w_i = 1, w_i \geq 0, \tag{1}$$

where e_j's are the canonical basis vectors in $\mathbb{R}^m$ and f_{ij}'s are one-dimensional probability densities. Weights w_i and f_{ij}'s are unknown. In model (1), the resulting function f is the probability density of a random vector $X = (x_1, \dots, x_m) \in [0,1]^m$ with entries $x_1, \dots, x_m$ that are independent conditional on a latent variable that can take K distinct values. In this work, we focus on the setting, where the aim is to explore relations between two variables ($m = 2$) and we explicitly construct polynomial-time estimators achieving the optimal rates for such models.

A relevant question is to check whether the multi-view model holds for a given particular problem in practice. We address this issue by providing estimators that are adaptive to the unknown number of components K varying on a wide scale of values. Very large values of K correspond to the absence of low-rank structure. For such K, our estimator achieves the same rate as the usual nonparametric density estimator of a smooth density (with no additional structure), and we show that this is optimal. In other words, our adaptive estimator achieves the minimax optimal rate regardless of whether the multi-view model holds or not. Thus, adaptation guarantees that checking the low-rank assumption is not necessary in practice.

We prove minimax lower bounds in the total variation distance for general discrete distributions on a set of cardinality D. We generalize [1, 2] in the sense that we derive lower bounds not only in expectation but also in probability and, in contrast to those works, we obtain the lower rate $\sqrt{\frac{D}{n}} \wedge 1$ for all $D, n \geq 1$ with no restriction. Next, under the low-rank matrix structure, we prove lower bounds of the order of $\psi(K, d, n) = \sqrt{\frac{Kd}{n}} \wedge 1$ both in expectation and in probability, with no restriction on K, d, n, where $d = d_1 \vee d_2$. Moreover, we propose a computationally efficient algorithm to estimate a low-rank probability matrix P and show that it attains the same rate $\psi(K, d, n)$ up to a logarithmic factor. Thus, we prove the minimax optimality of this rate and of our algorithm, up to a logarithmic factor. We also propose a method of estimating β-Hölder densities for $\beta \in (0, 1]$ under the generalized multi-view model. Our algorithm achieves the rate of convergence $(K/n)^{\beta/(2\beta+1)} \wedge n^{-\beta/(2\beta+2)}$ up to a logarithmic factor on the class of densities that are (i) β-Hölder over an *unknown* sub-rectangle of $[0,1]^2$ and (ii) represented as a sum of K separable components.

References

[1] Yanjun Han, Jiantao Jiao, and Tsachy Weissman, *Minimax estimation of discrete distributions under l_1 loss*, IEEE Transactions on Information Theory **61**(11) (2015), 6343–6354.

[2] Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh, *On learning distributions from their samples*, Conference on Learning Theory (2015), 1066–1100, PMLR.

Residual permutation test for regression coefficient testing

Tengyao Wang

(joint work with Kaiyue Wen and Yuhao Wang)

We consider the problem of testing whether a single coefficient is equal to zero in linear models when the dimension of covariates p can be up to a constant fraction of sample size n. In this regime, an important topic is to propose tests with finite-sample valid size control without requiring the noise to follow strong distributional assumptions. In this work, we propose a new method, called *residual permutation test* (RPT), which is constructed by projecting the regression residuals onto the space orthogonal to the union of the column spaces of the original and permuted design matrices. RPT can be proved to achieve finite-sample size validity under fixed design with just exchangeable noises, whenever $p < n/2$. Moreover, RPT is shown to be asymptotically powerful for heavy tailed noises with bounded $(1+t)$-th order moment when the true coefficient is at least of order $n^{-t/(1+t)}$ for $t \in [0,1]$. We further proved that this signal size requirement is essentially rate-optimal in the minimax sense. Numerical studies confirm that RPT performs well in a wide range of simulation settings.

The t-test is a supermartingale after all

Wouter M. Koolen

(joint work with Peter Grünwald)

The t-statistic is a widely-used scale-invariant statistic for testing the null hypothesis that the mean is zero. Martingale methods enable sequential testing with the t-statistic at every sample size, while controlling the probability of falsely rejecting the null. For one-sided sequential tests, which reject when the t-statistic is too positive, a natural question is whether they also control false rejection when the true mean is negative. We prove that this is the case using monotone likelihood ratios and sufficient statistics. We develop applications to the scale-invariant t-test, the location-invariant χ^2-test and sequential linear regression with nuisance covariates.

Let us explain here the main question being studied. Consider a data stream $X_1, X_2, \ldots$. We assume throughout that X_i are i.i.d. $\mathcal{N}(\delta\sigma, \sigma^2)$ for some *effect size* δ and *variance* σ^2. Our aim is to disqualify the composite null of no effect

$$\mathcal{H}_0 \;=\; \{\delta = 0, \sigma^2 > 0\}$$

with the help of the composite alternative that the effect size is a given $\delta_+ > 0$

$$\mathcal{H}_+ \;=\; \{\delta = \delta_+, \sigma^2 > 0\}.$$

In both hypotheses, the variance/scale σ^2 is a nuisance parameter. In this particular case, the nuisance is a *group*. This means we can quotient it out, for example by coarsening the data to $Z_i = X_i/|X_1|$. Upon doing that, we end up with a point-vs-point hypothesis test at the coarsened Z^n level, as all elements in either hypothesis agree with each other. So we can look at the likelihood ratio process $(M_n)_{n\geq 0}$

$$M_n \;:=\; \frac{p_{\delta_+}(Z^n)}{p_0(Z^n)}$$

Several expressions for this process can be obtained. Let $S_n = \sum_{i=1}^n X_i$ and $V_n = \sum_{i=1}^n X_i^2$ and $R_n = \frac{S_n}{\sqrt{V_n}}$. We have the *Hypergeometric* form

$$M_n \;=\; \frac{\Gamma\left(\frac{n}{2}\right)\, {}_1F_1\left(\frac{n}{2}; \frac{1}{2}; \frac{\delta_+^2 R_n^2}{2}\right) + \sqrt{2}\delta_+ R_n \Gamma\left(\frac{n+1}{2}\right)\, {}_1F_1\left(\frac{n+1}{2}; \frac{3}{2}; \frac{\delta_+^2 R_n^2}{2}\right)}{\Gamma\left(\frac{n}{2}\right) e^{\frac{n}{2}\delta_+^2}}$$

the *Pochhammer* form

$$M_n \;=\; \frac{1}{\Gamma\left(\frac{n}{2}\right) e^{\frac{n}{2}\delta_+^2}} \sum_{k=0}^{\infty} \frac{\Gamma\left(\frac{k+n}{2}\right)}{k!} \left(\sqrt{2}\delta_+ R_n\right)^k$$

the *Haar* forms

$$M_n \;=\; \frac{\int p_{\mathcal{N}(\delta_+\sigma, \sigma^2)}(X^n)\frac{1}{\sigma} d\sigma}{\int p_{\mathcal{N}(0,\sigma^2)}(X^n)\frac{1}{\sigma} d\sigma} \;=\; \frac{2}{\Gamma\left(\frac{n}{2}\right) e^{\frac{n}{2}\delta_+^2}} \int_0^\infty e^{w\sqrt{2}\delta_+ R_n - w^2} w^{n-1} dw$$

the *non-central Student-t* form

$$M_n \;=\; \frac{P(T_n; n-1, \delta_+\sqrt{n})}{P(T_n; n-1, 0)} \qquad \text{where} \qquad T_n \;=\; R_n\sqrt{\frac{n-1}{n-R_n^2}}$$

Several things are known about M_n.

- M_n is a martingale against the null hypothesis of zero effect $\mathcal{H}_0$.
- M_n is an e-variable against the null hypothesis of negative effect $\mathcal{H}_{\leq 0} := \{\delta \leq 0, \sigma^2 > 0\}$

Yet is it true that M_n is a super-martingale against $\mathcal{H}_{\leq 0}$? The main technical content of the talk is to show that it indeed is. For this we use the monotone likelihood ratio property, which we show does *not* hold for the original data $Z_{n+1}|Z^n$, but it does hold if we replace Z_{n+1} by a sufficient statistic. We show that in general the following:

Theorem 1. *Fix $\delta_0 \leq \delta_+$. Let $(T_n)_{n\in\mathbb{N}}$ be a sequence of sufficient statistics satisfying the monotone likelihood ratio property. Then the process*

$$\left(\prod_{i=1}^{n} \frac{p^{T_i}_{\delta_+}(T_i \mid U^{i-1})}{p^{T_i}_{\delta_0}(T_i \mid U^{i-1})}\right)_{n\in\mathbb{N}}$$

is identical to the likelihood ratio process $\left(\frac{p_{\delta_+}(U^n)}{p_{\delta_0}(U^n)}\right)_{n\in\mathbb{N}}$ and both are "test" (positive, starting at 1*) supermartingales relative to the one-sided null $\mathcal{H}_{\leq 0}$.*

We conclude the talk with an application of this general theorem to linear regression with nuisance covariates.

Semi-supervised classification with non-stationary data

Henry W. J. Reeve

We consider a semi-supervised classification problem with non-stationary label shift. In this scenario, the practitioner observes a labelled dataset followed by a sequence of unlabelled covariate vectors, in which the marginal probabilities of the class labels may change over time. Our objective is to sequentially predict the corresponding class label for each covariate vector without ever observing the ground-truth labels beyond the initial labelled dataset. Previous work has demonstrated the potential of sophisticated variants of online gradient descent to perform competitively with the optimal dynamic strategy [2]. We explore an alternative approach which employ's a varient of Lepski's method. We demonstrate the merits of this alternative methodology by establishing a high-probability regret bound on the test error at a single test time, which adapts automatically to the unknown dynamics of the label probabilities. Furthermore, we give bounds on the average dynamic regret, which match those of the online learning perspective for any given time interval. Our adaptive methodology leverages confidence intervals, whose construction builds upon a recent localised Dvoretzky–Kiefer–Wolfowitz–Massart inequality [3]. This work contains a succinct proof of a conjecture of [1] which holds without any constraints on the failure probability.

For more details, we refer the reader [4].

References

[1] Z. W. Birnbaum, and R. C. McCarty, *A distribution-free upper confidence bound for $\mathbb{P}(Y < X)$, based on independent samples of X and Y*, The Annals of Mathematical Statistics (1958): 558-562.

[2] Bai, Yong and Zhang, Yu-Jie and Zhao, Peng and Sugiyama, Masashi and Zhou, Zhi-Hua, *Adapting to online label shift with provable guarantees*, Advances in Neural Information Processing Systems, 2022.

[3] Henry W J Reeve, *A short proof of the Dvoretzky–Kiefer–Wolfowitz–Massart inequality*, arXiv:2403.16651.

[4] Henry W J Reeve, *An adaptive transfer learning perspective on classification in non-stationary environments*, arXiv:2405.18091.

Uniform confidence bands for centered purely random forests

Mathias Trabs
(joint work with Natalie Neumeyer, Jan Rabe)

We aim for statistical inference for random forests in a classical non-parametric regression setting. To this end, we focus on the theoretically simpler to analyse centered purely random forests where the partition is independent of the observations and the splits are always centered in a randomly chosen direction. Taking into account that each tree in the forests is constructed only based on a subsample, we exploit U-process theory and a Gaussian approximation of the supremum of empirical processes. As a main result we construct uniform confidence bands for centered purely random forests.

Let the observations be given by an i.i.d. sample $(X_i, Y_i) \in [0,1]^p \times \mathbb{R}, i = 1, \dots, n$ with $X_i \sim \mathrm{U}([0,1]^p)$ and regression function $m(x) = \mathbb{E}[Y|X = x]$. For a random partition and a point $x_0 \in [0,1]^p$ we denote the cell which contains x_0 by $A_n(x_0, \omega)$ where ω is a random variable encoding the random mechanism in the construction of the partition. The number of X_i's that fall into $A_n(x_0, \omega)$ is denoted by $\#A_n(x_0, \omega)$. By construction of a centred purely random forest, the volume of all cells $A_n(x_0, \omega)$ after k splits is 2^k. A key quantity in the analysis of the random forest is the kernel function $K_k(x_0, x) = 2^k \mathbb{P}_\omega(x \in A_n(x_0, \omega))$.

If each regression tree is calculated on a random subsample of size r_n, Peng et al. [2] have noticed that the random forest estimator can be written as a generalized (incomplete) U-statistic:

$$U^{(\mathrm{RF})}_{n,r_n,N,\omega}(x_0) := \frac{1}{\hat{N}} \sum_{I \in B_{n,r_n}} \rho_I \sum_{i \in I} Y_i \frac{\mathbb{I}\{X_i \in A_n(x_0; \omega_I)\}}{\#A_n(x_0; \omega_I)},$$

where B_{n,r_n} denotes the set of all subsets of $\{1, \dots, n\}$ of size r_n, ρ_I are i.i.d. $\mathrm{Ber}(N/\binom{n}{r_n})$ random variables and $\hat{N} = \sum_I \rho_I$ is the number of trees in the forest satisfying $\mathbb{E}[\hat{N}] = N$ for some sufficiently large $N \in \mathbb{N}$.

Our confidence band relies on the approximation

$$U^{(\mathrm{RF})}_{n,r_n,N,\omega}(x_0) - m(x_0) \approx \sqrt{\frac{\sigma^2 \Psi_k(x_0)}{n}} \mathbb{G}_n f_{x_0,k} + \mathcal{O}_{\mathbb{P}}(2^{-\alpha k/p}),$$

where $\Psi_k(x_0) := \mathbb{E}[K_k(x_0, X_1)^2]$ and $\mathbb{G}_n$ denotes the empirical process applied to $f_{x_0,k}(X_i, \varepsilon_i) = \sigma^{-1} \Psi_k(x_0)^{-1/2} \varepsilon_i K(x_0, X_i)$. Together with the Gaussian approximation of suprema of empirical processes by Chernozhukov et al. [1], we achieve a uniform asymptotic error bound. To this end, let B_k be a sequence of Gaussian processes with covariance structure

$$\begin{aligned} \mathrm{Cov}(B_k f_{x_1,k}, B_k f_{x_2,k}) &= \mathbb{E}[f_{x_1,k}(X_1, \varepsilon_1) f_{x_2,k}(X_1, \varepsilon_1)] \\ &= \big(\Psi_k(x_1)\Psi_k(x_2)\big)^{-1/2} \mathbb{E}[K_k(x_1, X_1) K_k(x_2, X_1)]. \end{aligned}$$

Theorem. For $k \in \mathbb{N}$ let $\mathbf{S}_k$ be a sequence of random variables such that $\mathbf{S}_k \overset{d}{=} \sup_{x_0 \in [0,1]^p} |B_k f_{x_0,k}|$. For $c_k(\beta) = F_{\mathbf{S}_k}^{-1}(1-\beta)$ and an estimator $\hat{\sigma}$ of $\mathbb{E}[\varepsilon_1^2]$ define

$$\mathcal{C}_{n,N}(x) = \left[U^{(\mathrm{RF})}_{n,r_n,N,\omega}(x) - \hat{\sigma} c_k(\beta) \sqrt{\frac{\Psi_k(x)}{n}}, U^{(\mathrm{RF})}_{n,r_n,N,\omega}(x) + \hat{\sigma} c_k(\beta) \sqrt{\frac{\Psi_k(x)}{n}} \right].$$

Under appropriate assumptions we have

$$\liminf_{n\to\infty} \inf_{m \in \mathcal{H}(\alpha,\Gamma)} \mathbb{P}\big(m(x) \in \mathcal{C}_n(x),\ \forall x \in [0,1]^p\big) \geq 1-\beta,$$

where the infimum is taken over all Hölder regular functions of regularity $\alpha \in (0,1]$ and Hölder norm bounded by $\Gamma > 0$.

References

[1] V. Chernozhukov, D. Chetverikov and K. Kato, *Gaussian approximation of suprema of empirical processes*, The Annals of Statistics **42**(4) (2014), 1564–1597.

[2] W. Peng, T. Coleman and L. Mentch, *Rates of convergence for random forests via generalized U-statistics*, Electronic Journal of Statistics **16**(1) (2022), 232–292.

Training Diagonal Linear Networks with Stochastic Sharpness-Aware Minimization

Gabriel Clara

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

We analyze the landscape and training dynamics of diagonal linear networks in a linear regression task, with the network parameters being perturbed by small isotropic normal noise. The addition of such noise may be interpreted as a stochastic form of sharpness-aware minimization (SAM) and we prove several results that relate its action on the underlying landscape and training dynamics to the sharpness of the loss. In particular, the noise changes the expected gradient to force balancing of the weight matrices at a fast rate along the descent trajectory. In the diagonal linear model, we show that this equates to minimizing the average sharpness, as well as the trace of the Hessian matrix, among all possible factorizations of the same matrix. Further, the noise forces the gradient descent iterates towards a shrinkage-thresholding of the underlying true parameter, with the noise level explicitly regulating both the shrinkage factor and the threshold. For more details, we refer to [1].

References

[1] G. Clara and S. Langer and J. Schmidt-Hieber, *Training Diagonal Linear Networks with Stochastic Sharpness-Aware Minimization*, arXiv:2503.11891, 2025.

Isotonic subgroup selection

Manuel M. Müller
(joint work with Henry W. J. Reeve, Timothy I. Cannings, Richard J. Samworth)

Given a sample of covariate-response pairs, we consider the subgroup selection problem of identifying a subset of the covariate domain where the regression function exceeds a pre-determined threshold. We introduce a computationally-feasible approach for subgroup selection in the context of multivariate isotonic regression based on martingale tests and multiple testing procedures for logically-structured hypotheses. Our proposed procedure satisfies a non-asymptotic, uniform Type I error rate guarantee with power that attains the minimax optimal rate up to polylogarithmic factors. Extensions cover classification, isotonic quantile regression and heterogeneous treatment effect settings. Numerical studies on both simulated and real data confirm the practical effectiveness of our proposal, which is implemented in the `R` package `ISS`.

For more details, we refer the reader to [1].

References

[1] M. M. Müller, H. W. J. Reeve, T. I. Cannings and R. J. Samworth, *Isotonic subgroup selection*, J. Roy. Statist. Soc., Ser. B. **87**(1) (2025), 132–156

Clustered random forests with potential covariate shift

Elliot H. Young
(joint work with Peter Bühlmann)

We introduce Clustered Random Forests, a random forests algorithm for clustered data, arising from independent groups that exhibit within-cluster dependence. The leaf-wise predictions for each decision tree making up clustered random forests takes the form of a weighted least squares estimator, which leverage correlations between observations for improved prediction accuracy. Clustered random forests are shown for certain tree splitting criteria to be minimax rate optimal for pointwise conditional mean estimation, while being computationally competitive with standard random forests. Further, we observe that the optimality of a clustered random forest, with regards to how (population level) optimal weights are chosen within this framework i.e. those that minimise mean squared prediction error, vary under covariate distribution shift. In light of this, we advocate weight estimation to be determined by a user-chosen covariate distribution with respect to which optimal prediction or inference is desired. This highlights a key difference in behaviour, between correlated and independent data, with regards to nonparametric conditional mean estimation under covariate shift. We demonstrate our theoretical findings numerically in a number of simulated and real-world settings, implemented in the `R` package `corrRF`.

References

[1] Elliot Young, Peter Bühlmann, *Clustered random forests with correlated data for optimal estimation and inference under potential covariate shift*, arXiv:2503.12634, 2025.

A Novel Statistical Approach to Analyze Image Classification Problem

Juntong Chen

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

The recent statistical theory of neural networks focuses on nonparametric denoising problems that treat randomness as additive noise. Variability in image classification datasets does, however, not originate from additive noise but from variation of the shape and other characteristics of the same object across different images. To address this problem, we introduce a tractable model for supervised image classification. While from the function estimation point of view, every pixel in an image is a variable, and large images lead to high-dimensional function recovery tasks suffering from the curse of dimensionality, increasing the number of pixels in the proposed image deformation model enhances the image resolution and benefits the object classification problem. In this talk, we focus on an approach based on fitting a convolutional neural network (CNN) to the data. We explicitly characterize the construction of the CNN and establish an approximation result. Under a minimal separation condition, we derive a misclassification error rate that depends on the sample size and the complexity of the deformation class.

For further analysis on image processing and image alignment using a one-nearest neighbor classifier, as well as related simulation studies, we refer the reader to [1].

References

[1] J. Chen, S. Langer and J. Schmidt-Hieber, *A Novel Statistical Approach to Analyze Image Classification Problem.*, arXiv:2206.02151, 2024.

Reflected diffusions as drivers of noise in denoising diffusion models

Asbjørn Holk

(joint work with Claudia Strauch and Lukas Trottner)

In recent years, generative AI has become ubiquitous in all parts of modern life. Many of these models utilise a denoising diffusion model as their mathematical backbone, however there is an inherent mismatch in the unbounded state space of such models and the often bounded support of their target distributions, leading to theoretically unjustified practices such as thresholding. We close this gap by instead considering reflected diffusion models, where thresholding is an integral part of the models. In particular, we show that under certain regularity conditions of the target distribution, a certain class of reflected diffusion models is mini-max optimal in total variation up to a polylogarithmic factor. We achieve this through a spectral decomposition of the transition density which is approximated by a neural network in space and then interpolated in time.

Improving the Convergence Rates of Forward Gradient Descent with Repeated Sampling

Niklas Dexheimer
(joint work with Johannes Schmidt-Hieber)

Forward gradient descent (FGD) has been proposed as a biologically more plausible alternative of gradient descent as it can be computed without backward pass. Considering the linear model with d parameters, previous work has found that the prediction error of FGD is, however, by a factor d slower than the prediction error of stochastic gradient descent (SGD). In this paper we show that by computing ℓ FGD steps based on each training sample, this suboptimality factor becomes $d/(\ell \wedge d)$ and thus the suboptimality of the rate disappears if $\ell \geq d$. We also show that FGD with repeated sampling can adapt to low-dimensional structure in the input distribution. The main mathematical challenge lies in controlling the dependencies arising from the repeated sampling process.

Reporter: Lukas Trottner

Participants

Dr. Chiara Amorino
Departamento de Tecnologia
Universidad Pompeu Fabra
Passeig Circumvallacio, 8
08003 Barcelona
SPAIN

Dr. Francis Bach
Inria
48 rue Barrault
75013 Paris Cedex
FRANCE

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

Dr. Tom Berrett
Department of Statistics
University of Warwick
Coventry CV4 7AL
UNITED KINGDOM

Prof. Dr. Peter Bühlmann
Seminar für Statistik
ETH Zürich
Rämistrasse 101
8092 Zürich
SWITZERLAND

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

Dr. Juntong Chen
Department of Applied Mathematics
University of Twente
Drienerlolaan 5
7522 NB Enschede
NETHERLANDS

Gabriel Clara
Department of Applied Mathematics
University of Twente
P.O. Box 217
7500 AE Enschede
NETHERLANDS

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

Prof. Dr. Holger Dette
Fakultät für Mathematik
Ruhr-Universität Bochum
44780 Bochum
GERMANY

Dr. Niklas Dexheimer
Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente
Drienerlolaan 5
7522 NB Enschede
NETHERLANDS

Chao Gao
Department of Statistics
The University of Chicago
5747 S. Ellis Avenue
Chicago, IL 60637-1514
UNITED STATES

Dr. Helene Halconruy
Télécom SudParis
19 place Marguerie Perey
91129 Palaiseau
FRANCE

Prof. Dr. Marc Hoffmann
CEREMADE
Université Paris Dauphine-PSL
Place du Maréchal de Lattre de Tassigny
75775 Paris Cedex 16
FRANCE

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

Prof. Dr. Wouter Koolen
CWI, Office L132
Science Park 123
1098 XG Amsterdam
NETHERLANDS

Prof. Dr. Sophie Langer
Department of Applied Mathematics
University of Twente
P.O. Box 217
7500 AE Enschede
NETHERLANDS

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

Prof. Dr. Nicolai Meinshausen
Seminar für Statistik
ETH Zürich (HG G 24.2)
Rämistrasse 101
8092 Zürich
SWITZERLAND

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

Prof. Dr. Nicole Mücke
Institut für Mathematische
Stochastik der TU Braunschweig
Postfach 3329
38023 Braunschweig
GERMANY

Manuel Müller
Statistical Laboratory
Centre for Mathematical Sciences
University of Cambridge
Wilberforce Road
Cambridge CB3 0WB
UNITED KINGDOM

Prof. Dr. Jonas Peters
Seminar für Statistik
ETH Zürich
HG G 17
Rämistr. 101
8092 Zürich
SWITZERLAND

Prof. Dr. Mark Podolskij
Université de Luxembourg
Faculté des Sciences, de la Technologie
et de la Communication
162 A, avenue de la Faiencerie
1511 Belvaux
LUXEMBOURG

Dr. Henry Reeve
Department of Mathematics
University of Bristol
University Walk
Bristol BS8 1TW
UNITED KINGDOM

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

Prof. Dr. Patricia Reynaud-Bouret
Laboratoire J.A. Dieudonné
Université Côte d'Azur
Parc Valrose
06108 Nice Cedex
FRANCE

Prof. Dr. Richard Samworth
Statistical Laboratory
Centre for Mathematical Sciences
University of Cambridge
Wilberforce Road
Cambridge CB3 0WB
UNITED KINGDOM

Prof. Dr. Johannes Schmidt-Hieber
Department of Applied Mathematics
University of Twente
Drienerlolaan 5
7522 NB Enschede
NETHERLANDS

Prof. Dr. Rajen Dinesh Shah
Department of Pure Mathematics and
Mathematical Statistics
University of Cambridge
Wilberforce Road
Cambridge CB3 0WB
UNITED KINGDOM

Prof. Dr. Claudia Strauch
Institut für Mathematik
Universität Heidelberg
Im Neuenheimer Feld 205
69120 Heidelberg
GERMANY

Prof. Dr. Taiji Suzuki
Department of Mathematical Informatics
University of Tokyo
Hongo 7-3-1, Bunkyo-ku
Tokyo 113-8656
JAPAN

Asbjørn Holk Thomsen
Matematisk Institut
Aarhus Universitet
8000 Aarhus
DENMARK

Prof. Dr. Mathias Trabs
Institut für Stochastik
Karlsruher Institut f. Technologie (KIT)
76128 Karlsruhe
GERMANY

Dr. Lukas Trottner
School of Mathematics
University of Birmingham
Birmingham B15 2TT
UNITED KINGDOM

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

Prof. Dr. Aad van der Vaart
Faculty of Electrical Engineering,
Mathematics and Computer Science
Delft University of Technology
P. O. Box 356
2628 BL Delft
NETHERLANDS

Dr. Kabir Verchand
School of Industrial and Systems
Engineering
Georgia Institute of Technology
Ferst Drive
Atlanta, GA 30318
UNITED STATES

Dr. Nicolas Verzelen
INRAE, UMR 729, MISTEA, Institut
Agro, Univ. Montpellier
2, Place Viala
34060 Montpellier Cedex 1
FRANCE

Prof. Dr. Martin Wainwright
Massachusetts Institute of Technology
Statistics and Data Science Center
Cambridge, MA 02139
UNITED STATES

Prof. Dr. Sven Wang
Institut für Mathematik
Fachbereich Mathematik
Humboldt-Universität Berlin
10099 Berlin
GERMANY

Dr. Tengyao Wang
Department of Mathematics
London School of Economics
Houghton Street
London WC2A 2AE
UNITED KINGDOM

Dr. Yuting Wei
Department of Statistics
The Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia, PA 19104-6340
UNITED STATES

Dr. Min Xu
Dept. of Statistics
Rutgers University
Hill Center, Busch Campus
New Brunswick, NJ 08903
UNITED STATES

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

Dr. Elliot Young
Statistics Laboratory,
Cambridge University
Wilberforce Road
Cambridge CB3 0WA
UNITED KINGDOM

Prof. Dr. Yi Yu
Department of Statistics
University of Warwick
Room 4.11
Coventry CV4 7AL
UNITED KINGDOM