Mathematisches Forschungsinstitut Oberwolfach

Report No. 15/2021

DOI: 10.4171/OWR/2021/15

Mathematical Foundations of Machine Learning (hybrid meeting)

Organized by
Peter Bartlett, Berkeley
Cristina Butucea, Palaiseau
Johannes Schmidt-Hieber, Enschede

21 March – 27 March 2021

Abstract. Machine learning has achieved remarkable successes in various applications, but there is wide agreement that a mathematical theory for deep learning is missing. Recently, some first mathematical results have been derived in different areas such as mathematical statistics and statistical learning. Any mathematical theory of machine learning will have to combine tools from different fields such as nonparametric statistics, high-dimensional statistics, empirical process theory and approximation theory. The main objective of the workshop was to bring together leading researchers contributing to the mathematics of machine learning.

A focus of the workshop was on theory for deep neural networks. Mathematically speaking, neural networks define function classes with a rich mathematical structure that are extremely difficult to analyze because of nonlinearity in the parameters. Until very recently, most existing theoretical results could not cope with many of the distinctive characteristics of deep networks such as multiple hidden layers or the ReLU activation function. Other topics of the workshop are procedures for quantifying the uncertainty of machine learning methods and the mathematics of data privacy.

Mathematics Subject Classification (2010): 62G05, 62G08, 62G20.

Introduction by the Organizers

The workshop *Mathematical Foundations of Machine Learning*, was organized in hybrid format due to the sanitary restrictions at the time. The conference was very well attended by 60 participants: 6 researchers were able to join in person, 54 researchers attended by visio. The participants were from various countries in Europe and America. The schedule totalized 22 talks given by participants.

Virtual rooms were available for discussion and virtual social events were organized on Monday and Thursday.

Machine learning is the umbrella term of a number of data analysis tools for prediction problems that have been mainly developed within computer science. The strength of these methods is the wide applicability and the availability of fast algorithms to process huge datasets. While in the classical statistical framework, parameters have an interpretation (for instance the regression coefficients), the parameters in machine learning are meaningless and the methods are commonly referred to as black box procedures.

To formulate a mathematical framework for such black box procedures is a quickly advancing field at the interface of mathematical statistics and statistical learning.

The workshop was organized on the topics mentioned below.

Neural networks: Most of the recent mathematical contributions in machine learning are on deep neural networks and this was a key topic during the workshop. The concept of a neural network dates back to the fourties and fifties [Rosenblatt, 1958] with a lot of mathematical research carried out during the late eighties and early nineties. Based on their success in image classification, deep neural networks have been popularized only recently. The mathematical analysis of a deep network is much more involved due to the the hierarchical structure and the non-linear dependence of the outcome on the parameters. There is also some difference in terms of the used activation functions. Whereas sigmoidal activation functions have been popular in the nineties, the most prominent activation function for deep neural networks is the so called ReLU (rectified linear unit) activation, which induces many interesting mathematical structures on the network functions making deep ReLU network a mathematically rich object.

During the workshop, David Donoho discussed a new phenomenon, called neural collapse, that occurs for overparametrized neural networks during the terminal phase of deep learning. In a joint talk, Michael Kohler and Sophie Langer presented their recent work on rate-optimal generalization guarantees for learning a shallow network using gradient descent. The work shows that it is possible to combine the analysis of (stochastic) gradient descent, approximation theory and statistical bounds for neural networks into meaningful results. Stefan Richter discussed forecasting time series using deep neural networks. He showed that fast convergence rates can be obtained for pointwise forecasting and estimation of the predictive distribution. Robustness for shallow neural networks was considered in the talk by Sebastien Bubeck. He presented a conjecture stating that interpolation with few network parameters automatically implies a large Lipschitz constant of the network function and also provided some insights and first results why this should be true. Gitta Kutyniok studied an invariance property of graph convolutional networks called transferability. She showed that graph convolutional networks can achieve transferability in two different ways. There were also two talks on theory for optimization. Steffen Dereich derived conditions that guarantee convergence of stochastic gradient descent schemes and applied these results to neural networks. Ohad Shamir looked from a general perspective on non-smooth and non-convex optimization problems and introduced a new generalized notion of stationarity.

High-dimensional statistics: High-dimensional problems have been one of the main focus areas in mathematical statistics during the past 20 years and are closely intertwined with machine learning.

Gilles Blanchard considered estimation of many vector means simultaneously and extended Stein shrinkage to this setup. He argued that the problem becomes in some sense easier as the dimension of the vector space grows. Michael Vogt discussed in his talk a new estimator for the effective noise term that occurs in the analysis of the LASSO. In the talk given by Chao Gao, several estimators for phase synchronization were compared and it was shown that all of these methods achieve the exact minimax estimation risk up to a small additional term. In Jianqing Fan's presentation, a ℓ_p perturbation theory for the hollowed version of the principal component analysis was developed and this was subsequently applied to a community detection problem. Vianney Perchet discussed matching of sparse random graphs in an online setting. To recover the spectrum of the adjacency matrix associated to a graph structure, Tracy Ke considered an approach based on counting short cycles.

Statistical learning tools: Statistical learning deals with the statistical error and the complexity of the generated function classes. Upper and lower bounds on the VC dimension of neural networks were derived and reviewed in the monograph [Anthony and Bartlett, 1990]. For deep ReLU networks; an almost sharp characterization of the VC dimension has been obtained recently in [Bartlett et al., 2019]. The analysis of the VC dimension suggests that deep networks should not perform well, as the VC dimension also depends on the network depth and becomes useless in the case where the number of network parameters is larger than the sample size. Bounds on related notions of complexity based on covering numbers (see, for example, [Anthony and Bartlett, 1990]) depend instead on the scale of network parameters, and there has been a spate of recent results refining these deviation inequalities for networks with ReLU nonlinearities. However, the scaling of these bounds with depth does not match observations of practical deep networks.

Discussing how statistical learning tools can explain the success of deep learning and other machine learning methods was a key focus of the workshop. The success of convolutional neural networks is commonly believed to be due to the underlying invariance structure. In this spirit, Andrea Montanari combined in his talk invariant random features and invariant kernel methods and showed that incorporating invariance results in a reduction of the test error by a factor scaling with some power of the underlying dimension. Daniel Hsu introduced a specific version of self-supervised learning, called contrastive learning, in a setting, where we observe multiple 'views' for each datum. In the context of topic prediction, for instance, two views can be observed if for each document we have access to the abstract and the introduction. It was then argued that linear functions of the learned representations are nearly optimal for contrastive learning. Richard Samworth studied a general scheme for adaptive transfer learning, derived the minimax estimation rate and proposed a minimax rate-optimal estimation procedure.

Zero loss, high gain: One of the surprising phenomena of several machine learning methods is that even if they are trained to have zero loss on the training data, they still perform well on new data. This phenomenon contradicts the existing statistical theory which says that a method interpolating all data points has a huge variance and will do poorly on new data. That overfitting performs well is one of the most intriguing properties of modern machine learning. In [Allen-Zhu et al., 2018, Du et al., 2018], it has been shown that gradient descent with random initialization converges to zero training error in a highly overparametrized setting.

In her talk, Sara van de Geer derived risk bounds for minimum ℓ_1-interpolation in a high-dimensional binary classification model. Interestingly, the risk can still converge to zero in certain regimes and therefore providing a theoretical justification for the zero loss, high gain phenomenon in this setting. In a similar spirit, Flori Bunea studied interpolation estimators for topic models and derived bounds for the estimation risk.

Uncertainty: One of the most pressing problems is to compute the uncertainty of the output of black-box methods. Bayesian approaches come with a built-in notion of uncertainty quantification. The two main problems connected to the Bayesian approach are the computational cost and the frequentist interpretation of Bayesian credible sets.

During the workshop, one session was organized on this topic. Veronika Rockova used generative adversarial networks (GANs) to estimate the likelihood ratio and derived theoretical properties of the Metropolis-Hastings algorithm based on the approximated likelihood. Richard Nickl presented in his talk a Markov chain Monte Carlo method and proved that it converges with polynomial dependence on the dimension of the model. As application computation of the posterior for a PDE model was considered.

References

[Allen-Zhu et al., 2018] Allen-Zhu, Z., Li, Y., and Song, Z. A Convergence Theory for Deep Learning via Over-Parameterization. *arXiv e-prints* (2018), arXiv:1811.03962.

[Anthony and Bartlett, 1990] Anthony, M., and Bartlett, P. L. *Neural network learning: theoretical foundations.* Cambridge University Press, Cambridge (1999).

[Bartlett et al., 2019] Bartlett, P., Harvey, N., Liaw, C., and Mehrabian, A. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research 20* (2019), 1–17.

[Du et al., 2018] Du, S. S., Lee, J. D., Li, H., Wang, L., and Zhai, X. Gradient Descent Finds Global Minima of Deep Neural Networks. *arXiv e-prints* (2018), arXiv:1811.03804.

[Rosenblatt, 1958] Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review 65*, 6 (1958), 386–408.

Workshop (hybrid meeting): Mathematical Foundations of Machine Learning

Table of Contents

Abstracts

Stein effect for estimating many vector means: a “blessing of dimensionality” phenomenon

Gilles Blanchard
(joint work with Hannah Marienwald, Jean-Baptiste Fermanian)

Consider a model with many independent samples from different distributions,

$$\begin{cases} X_\bullet^{(k)} := (X_i^{(k)})_{1\leq i\leq N_k} \overset{i.i.d.}{\sim} \mathbb{P}_k,\ 1\leq k\leq B; \\ \\ (X_\bullet^{(1)},\ldots,X_\bullet^{(B)}) \text{ independent,} \end{cases}$$

where $\mathbb{P}_1,\ldots,\mathbb{P}_B$ are square integrable distributions on $\mathbb{R}^d$ which we will call *tasks.* The goal is the estimation of their means $(\mu_k)_{1\leq k\leq B}$.

To simplify exposition, assume that $\mathbb{P}_i$ is a Gaussian distribution with mean μ_i, variance $\sigma^2 I_d$ and that all samples have the same size N. In the machine learning literature, the problem has been coined as “multiple task averaging” by [Feldman et al., 2014], but can be seen in more traditional statistical/decision theoretical terminology as a “compound decision problem” [Robbins, 1951].

Our motivation for considering this setting is the growing number of large databases taking the above form, where independent bags, corresponding to different but conceptually similar distributions, are available; for example, one can think of k as an index for a large number of individuals, for each of which a number of observations (assumed to be sampled from an individual-specific distribution) have been collected, say medical records, or online activity by some governmental or corporate spying device.

Given estimators $\widehat{\mu}_1,\ldots,\widehat{\mu}_B$ for $\mu_1,\ldots,\mu_B$, we can be interested either in the mean squared error (MSE) for the estimation of each single mean,

$$\mathrm{MSE}(k,\widehat{\mu}_k) := \mathbb{E}\left[\left\|\widehat{\mu}_k-\mu_k\right\|^2\right], k=1,\ldots,B,$$

or in the *compound* MSE, i.e. averaged over all tasks,

$$\overline{\mathrm{MSE}}(\widehat{\mu}_\bullet) := \frac{1}{B}\sum_{k=1}^{B}\mathrm{MSE}(k,\widehat{\mu}_k).$$

The benchmark estimators are the “naive” task-wise empirical means

$$\widehat{\mu}_k^{\mathrm{NE}} := \frac{1}{N}\sum_{i=1}^{N}X_i^{(k)},$$

and it holds

$$\mathrm{MSE}(k,\widehat{\mu}_k^{\mathrm{NE}}) = \mathbb{E}\left[\left\|\widehat{\mu}_k^{\mathrm{NE}}-\mu_k\right\|^2\right] = d\sigma^2 =: \mathcal{E}^{\mathtt{NE}}.$$

A simple idea to improve over the naive estimator for a given task, say the first, is the following. Assume first that an oracle gives us the information that for some $\tau > 0$, it holds

$$\|\mu_1 - \mu_j\|^2 \leq \tau \mathcal{E}^{\mathtt{NE}}, \qquad j = 1, \ldots, V_1, \tag{1}$$

for some $V_1 \leq B$; call this τ-neighbor tasks of task 1 (after reordering indices for convenience). Consider shrinking the naive estimator towards the average mean of neighbor tasks

$$\widetilde{\mu}_1 = \gamma \widehat{\mu}_1^{\mathrm{NE}} + (1-\gamma)\left(\frac{1}{V_1}\sum_{k=1}^{V_1} \widehat{\mu}_k^{\mathrm{NE}}\right).$$

Pick

$$\gamma = \frac{\tau(V_1 - 1)}{(1+\tau)(V_1-1)+1},$$

then by independence of bags and the triangle inequality:

$$\frac{\mathrm{MSE}(1, \widetilde{\mu}_1)}{\mathcal{E}^{\mathtt{NE}}} \leq \frac{\tau}{1+\tau} + \frac{1}{V_1(1+\tau)}. \tag{2}$$

Repeating this over all tasks and summing, it is not difficult to show

$$\frac{\overline{\mathrm{MSE}}(\widetilde{\mu})}{\mathcal{E}^{\mathtt{NE}}} \leq \frac{\tau}{1+\tau} + \frac{\mathcal{N}}{B}\frac{1}{(1+\tau)}, \tag{3}$$

where $\mathcal{N}$ is the covering number of the set of means $\{\mu_1, \ldots, \mu_B\}$ at scale $\sqrt{\tau\mathcal{E}^{\mathtt{NE}}}/2$. (Proof: $\sum_{k=1}^{B} V_i^{-1} \leq \mathcal{N}$, assuming the oracle has given in each case a list of *all* neighbor tasks, i.e. satisfying Eq. (1).)

Thus, assuming the oracle information, in all cases we can improve over naive estimation (task-wise as well as in the compound sense), and the gain can be *substantial* if there exists $\tau \ll 1$ such that the covering number of the true tasks at scale $\sqrt{\tau\mathcal{E}^{\mathtt{NE}}}/2$ is $\ll B$. Whether or not this is the case is context-dependent, but we can easily imagine situations where the set of means has some *structure* resulting in a small covering number (e.g. supported by a low-dimensional manifold; sparse; clustered. . .)

Now to the actually interesting question: what can we do in absence of oracle information? The answer is that we can use *tests* $T_{ij}, (i,j) \in \{1, \ldots, B\}^2$ for

$$(H_{0,ij}) : \|\mu_i - \mu_j\|^2 > \tau\mathcal{E}^{\mathtt{NE}}, \quad \text{against} \quad (H_{1,ij}) : \|\mu_i - \mu_j\|^2 \leq (\tau/2)\mathcal{E}^{\mathtt{NE}}.$$

Assume that we have a good control for both the family-wise type I and II error of these tests, and that they are independent of the data used to control the estimators (for instance, consider splitting each sample in two). Then we can apply the above argument conditionally to the tests, thus getting the controls Eq. (2) and Eq. (3) with some adjustments (τ replaced by $\tau/2$ in the scale of the covering numbers, and an additional factor 2 to account for data splitting with respect to the naive estimator which does not use data splitting).

All of this, though, has potential significance only if we can find such tests for some $\tau \ll 1$, otherwise the potential improvement is almost nonexistent. Now the little miracle or “blessing” of dimensionality is that we can find a family of tests having both controlled Type I and II error controlled provided $\tau \gtrsim d^{-1/4}$ see e.g. [Baraud, 2002]: the testing separation distance is much smaller than the estimation error of the naive estimator in high dimension. (This compensates – somewhat – the usual curse of dimensionality, which is that $\mathcal{E}^{\mathtt{NE}}$ increases linearly with d.) Thus, in high dimensionality, we at least have the potential of an improvement over the naive estimator up to a factor of order $O(d^{-1/4})$, which can be the case if the means have a favorable structure. This improvement can be quantified for the compound estimation error but also for each individual estimation error.

These results have interesting ties to the classical literature on the James-Stein estimator [James and Stein, 1961]; see also [Beran, 1996] and to the compound decision literature (see e.g. [Brown and Greenshtein, 2009], where the multiple mean estimation problem was tackled from the compound decision theory angle, albeit only in dimension 1).

In the papers [Marienwald et al., 2021, Blanchard and Fermanian, 2021] we develop these ideas, in particular we

- give a precise nonasymptotic account of the above phenomenon;
- study in particular precise results for tests based on an unbiased U-statistic for $\|\mu_i - \mu_j\|^2$;
- consider the case of non-isotropic distributions with arbitrary covariance; in this case the role of the ambiant dimension d is replaced by an appropriate notion of *effective dimensionality* (which has to be estimated);
- generalize the Gaussian case to the bounded case and even the case with only moments of order 4 (using the median-of-means methodology);
- apply the above results in conjunction with reproducing kernel methods to improve *kernel mean embedding* (see [Muandet et al., 2017]) estimation;
- illustrate the performance of the approach on simulated and real data.

References

[Baraud, 2002] Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606.

[Beran, 1996] Beran, R. (1996). Stein estimation in high dimensions: a retrospective. *Madan Puri Festschrift*, E. Denker and M. Brunner eds., 91-110.

[Blanchard and Fermanian, 2021] Blanchard, G. and Fermanian, J.-B. (2021). Nonasymptotic signal detection and two-sample tests in high dimension (Submitted)

[Brown and Greenshtein, 2009] Brown, L. and Greenshtein, E. (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. Ann. Statist. 37(4): 1685-1704

[Feldman et al., 2014] Feldman, S., Gupta, M. R., and Frigyik, B. A. (2014). Revisiting Stein's paradox: multi-task averaging. *Journal of Machine Learning Research*, 15(106):3621–3662.

[James and Stein, 1961] James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pages 361–379.

[Marienwald et al., 2021] Marienwald, H, Fermanian, J.-B. and Blanchard, G. (2021). High-Dimensional Multi-Task Averaging and Application to Kernel Mean Embedding. Proc. Conference on Artificial Intelligence and Statistics (AISTATS 2021).

[Muandet et al., 2017] Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2017). Kernel mean embedding of distributions: a review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141.

[Robbins, 1951] Robbins, H. (1951). Asymptotically subminimax solutions of compound decision problems. Proc. Second Berkeley Symp. Math. Statist. Probab. 1 131-148.

[Zhang, 2003] Zhang, C.-H. Compound decision theory and empirical Bayes methods. Ann. Statist. 31(2): 379-390.

A law of robustness for two-layers neural networks

Sébastian Bubeck

(joint work with Yuanzhi Li, Dheeraj Nagaraj)

I will present a mathematical conjecture potentially establishing overparametrization as a law of robustness for neural networks. I will tell you some of the things that we already know about this conjecture. Time-permitting I will include a discussion of how to think about various quantities for higher order tensors (their rank, the relation between spectral norm and nuclear norm, and concentration for random tensors).

References

[1] S. Bubeck, Y. Li, D. Nagaraj *A law of robustness for two-layers neural networks*, arXiv:2009.14444, (2020).

Prediction under latent factor regression: adaptive PCR, interpolating predictors and beyond

Florentina Bunea

(joint work with Xin Bing, Seth Strimas-Mackey, Marten Wegkamp)

This work is devoted to the derivation and analysis of finite sample prediction risk bounds for a class of linear predictors of a random response $Y \in \mathbb{R}$ from a high-dimensional, and possibly highly correlated random vector $X \in \mathbb{R}^p$, when the vector (X, Y) follows a latent factor regression model, generated by a latent vector of dimension lower than p. We assume that there exist a random, unobservable, latent vector $Z \in \mathbb{R}^K$, a deterministic matrix $A \in \mathbb{R}^{p\times K}$, and a coefficient vector $\beta \in \mathbb{R}^K$ such that

$$\begin{aligned} Y &= Z^\top \beta + \varepsilon, \\ X &= AZ + W, \end{aligned} \tag{1}$$

with some unknown $K < p$. The random noise $\varepsilon \in \mathbb{R}$ and $W \in \mathbb{R}^p$ have mean zero and second moments $\sigma^2 =: \mathbb{E}[\varepsilon^2]$ and $\Gamma =: \mathbb{E}[WW^\top]$, respectively. The random variable ε and random vectors W and Z are mutually independent. Throughout the paper, both $\Sigma_Z := \mathbb{E}[ZZ^\top]$ and A have rank equal to K.

Independently of this model formulation, but based on the belief that Y depends chiefly on a lower-dimensional approximation of X, prediction of Y via principal components (PCR) is perhaps the most utilized scheme, with a history dating back many decades.

Given the data $\mathbf{X} = (\mathbf{X_1}, \dots, \mathbf{X_n})^\top$ and $\mathbf{Y} = (\mathbf{Y_1}, \dots, \mathbf{Y_n})$ consisting of n independent copies of $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$, PCR-$k$ predicts $Y_* \in \mathbb{R}$ after observing a new data point $X_* \in \mathbb{R}^p$ by

$$\begin{aligned} \widehat{Y}^*_{\mathbf{U_k}} &= X_*^\top \mathbf{U_k} \left[\mathbf{U_k^\top X^\top X U_k}\right]^+ \mathbf{U_k^\top X^\top Y} \\ &= X_*^\top \mathbf{U_k} \left[\mathbf{X U_k}\right]^+ \mathbf{Y}, \end{aligned} \tag{2}$$

where $\mathbf{U_k}$ is the $p \times k$ matrix of the top eigenvectors of the sample covariance matrix $\mathbf{X^\top X/n}$, relative to the largest k eigenvalues, where k is ideally determined in a data-dependent fashion and M^+ denotes the Moore-Penrose inverse of a matrix M.

Model (1) provides a natural context for the theoretical analysis of PCR-k prediction. It is perhaps surprising that its theoretical study so far is limited to asymptotic analyses of the out-of-sample prediction risk for PCR-K as $p, n \to \infty$. To the best of our knowledge, finite sample prediction risk bounds for $\widehat{Y}^*_{\mathbf{U_k}}$, corresponding to data-dependent choices of k, are lacking in the literature, and their study under factor models of unknown K, possibly varying with n, provides motivation for this work.

To obtain risk bounds for PCR, we prove a master theorem, Theorem 1.1, that establishes a finite sample prediction risk bound for linear predictors of the general form

$$\widehat{Y}^*_{\widehat{B}} = X_*^\top \widehat{B} \left(\widehat{B}^\top \mathbf{X}^\top \mathbf{X} \widehat{\mathbf{B}}\right)^+ \widehat{B}^\top \mathbf{X}^\top \mathbf{Y}, \tag{3}$$

where $\widehat{B} \in \mathbb{R}^{p \times q}$ is an appropriate matrix that may be deterministic or depend on the data $\mathbf{X}$, with dimension q allowed to be random.

This approach has the benefit of not only covering the special case of PCR, corresponding to choice $\widehat{B} = \mathbf{U_k}$, but of offering a unifying analysis of other prediction schemes of the form (3). One important example corresponds to $\widehat{B} = \mathbf{I_p}$, which leads to another model agnostic predictor, the generalized least squares estimator (also known as the minimum norm interpolating predictor), which has enjoyed revamped popularity in the last two years.

Using the full data matrix $\mathbf{X}$ for prediction – instead of just the first k principal components as in PCR – leads to additional bias compared to PCR prediction. However, in the high-dimensional regime $p \gg n$, this bias can become small and choosing $\widehat{B} = \mathbf{I_p}$ can become a viable alternative to PCR that requires no tuning parameters.

In addition to these two model-agnostic prediction methods, Theorem 1.1 can be used to analyze predictors directly tailored to model (1), which are shown formally to be of type (3) in Section 4.2 of [1]. We give a particular expression of $\widehat{B}$, as well as the corresponding prediction analysis, under further modelling restrictions that render parameters K, A and β identifiable. The model specifications given in the aforementioned Section 4.2. allow us to view A as a cluster membership matrix, making it possible to address a third, understudied, class of examples pertaining to prediction from low-dimensional feature representation, that of prediction of

Y via latent cluster centers, for features that exhibit an overlapping clustering structure corresponding to A.

1. Main results

We write $(X, Y) \sim$ sG-FRM(θ), with $\theta =: (K, A, \beta, \Sigma_Z, \Gamma, \sigma^2)$, if (X, Y) satisfy model 1, and ε, Z, W are sub-Gaussian, with respective sub-Gaussian constants γ_ε, γ_z and γ_w. Define

$$\delta_W := \delta_W(\theta) = c \left[\|\Gamma(\theta)\|_{\mathrm{op}} + \frac{\mathrm{tr}(\Gamma(\theta))}{n} \right], \tag{4}$$

with $c = c(\gamma_w)$ being some positive constant.

Theorem 1.1. *Let $\widehat{B} = \widehat{B}(\mathbf{X}) \in \mathbb{R}^{\mathbf{p} \times \mathbf{q}}$ for some $q \geq 1$, and set*

$$\widehat{r} := rank\left(\mathbf{X}\mathbf{P}_{\widehat{\mathbf{B}}}\right), \qquad \widehat{\eta} := \frac{1}{n}\sigma^2_{\widehat{r}}\left(\mathbf{X}\mathbf{P}_{\widehat{\mathbf{B}}}\right), \qquad \widehat{\psi} := \frac{1}{n}\sigma^2_1\left(\mathbf{X}\mathbf{P}^{\perp}_{\widehat{\mathbf{B}}}\right). \tag{5}$$

For any $\theta = (K, A, \beta, \Sigma_Z, \Gamma, \sigma^2)$ with $K \leq Cn/\log n$ for some positive constant $C = C(\gamma_z)$ such that $(X, Y) \sim$ sG-FRM(θ), there exists some absolute constant $c > 0$ such that

$$\begin{aligned} \mathbb{P}_\theta\Bigg\{ \mathcal{R}(\widehat{B}) - \sigma^2 \lesssim \left[\frac{\|\Gamma\|_{\mathrm{op}}}{\widehat{\eta}} \widehat{r} + \left(1 + \frac{\delta_W}{\widehat{\eta}}\right)(K \wedge \widehat{r} + \log n) \right] \frac{\sigma^2}{n} \\ + \left[\left(1 + \frac{\|\Gamma\|_{\mathrm{op}}}{\widehat{\eta}}\right)\delta_W + \left(1 + \frac{\delta_W}{\widehat{\eta}}\right)\widehat{\psi} \right] \beta^\top (A^\top A)^{-1} \beta \Bigg\} \geq 1 - c/n. \end{aligned} \tag{6}$$

Here the symbol $\lesssim$ means the inequality holds up to a multiplicative constant possibly depending on the sub-Gaussian constants γ_ε, γ_z and γ_w.

The interpretation of this bound, as well as detailed derivations of the excess risk bounds corresponding to our three main examples are given in [1].

References

[1] X. Bing, F. Bunea, S. Strimas-Mackey, M. Wegkamp *Prediction under latent factor regression: adaptive PCR, interpolating predictors and beyond*, arXiv https://arxiv.org/abs/2007.10050, 2021.

Several structured thresholding bandit problem

Alexandra Carpentier

(joint work with James Cheshire, Pierre Menard, Andrea Locatelli, Maurilio Gutzeit)

In this talk we will discuss the thresholding bandit problem, i.e. a sequential learning setting where the learner samples sequentially K unknown distributions for T times, and aims at outputting at the end the set of distributions whose means μ_k are above a threshold τ. We will study this problem under four structural assumptions, i.e. shape constraints: that the sequence of means is monotone, unimodal, concave, or unstructured (vanilla case). We will provide in each case minimax results on the performance of any strategies, as well as matching algorithms. This will highlight the fact that even more than in batch learning, structural assumptions have a huge impact in sequential learning.

Convergence of stochastic gradient descent schemes for Łojasiewicz-landscapes

Steffen Dereich

(joint work with Sebastian Kassing)

In this talk we discuss stochastic gradient descent (SGD) schemes. We first fix the notation. We let $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\in\mathbb{N}_0}, \mathbb{P})$ be a filtered probability space, $F : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function and let $(X_n)_{n\in\mathbb{N}_0}$ be an adapted sequence of $\mathbb{R}^d$-valued random variables such that for every $n \in \mathbb{N}$

$$X_n = X_{n-1} - \gamma_n(\nabla F(X_{n-1}) + D_n),$$

where

- $(\gamma_n)_{n\in\mathbb{N}}$ is a sequence of strictly positive reals, the *step-sizes*,
- $(D_n)_{n\in\mathbb{N}}$ is an $(\mathcal{F}_n)_{n\in\mathbb{N}}$-adapted sequence of martingale differences, the *perturbation*,
- X_0 is an $\mathcal{F}_0$-measurable random variable, the *initial value*.

We discuss convergence of

$$(i)\ \ (F(X_n)),\ \ (ii)\ \ (\nabla F(X_n))\quad \text{and}\quad (iii)\ \ (X_n)$$

under weak assumptions. Whereas convergence of (i) and (ii) are considered in various articles convergence of (iii) is rather subtle without imposing restrictive assumptions. We restrict attention to two events: we let

$$\mathbb{L} = \Big\{\limsup_{n\to\infty} |X_n| < \infty\Big\}$$

and for $p \geq 1$ and a sequence $(\sigma_n)_{n\in\mathbb{N}}$ of strictly positive reals,

$$\mathbb{M}^p_\sigma = \Big\{\limsup_{n\to\infty} \sigma_n^{-1}\mathbb{E}[|D_n|^p|\mathcal{F}_{n-1}]^{1/p} < \infty\Big\}.$$

The first result concerns convergence of (i) and (ii).

Theorem: (see [1]) Let $p \in (1,2]$ and suppose that ∇F is locally Lipschitz continuous. If

$$\sum_{n=1}^{\infty} (\gamma_n \sigma_n)^p < \infty \quad \text{and} \quad \sum_{n=1}^{\infty} \gamma_n = \infty,$$

then, on $\mathbb{L} \cap \mathbb{M}_\sigma^p$, almost surely,

$$\text{the limit } (F(X_n))_{n \in \mathbb{N}_0} \text{ exists and } \lim_{n\to\infty} \nabla F(X_n) = 0.$$

In the case where F does not posses a continuum of critical points[1], this result entails that also on $\mathbb{L} \cap \mathbb{M}_\sigma^p$, almost surely, (X_n) converges.

In the case where F possesses a continuum of critical points there exist examples for which the solution (x_t) to the ordinary differential equation

$$\dot{x}_t = -\nabla F(x_t)$$

stays local and (x_t) does not converge. One needs additional assumptions.

Definition: A C^1-function $F : \mathbb{R}^d \to \mathbb{R}$ is said to be a *Łojasiewicz-function* if for every critical point x of F there exists $\beta \in [\frac{1}{2}, 1)$, $\text{Ł} > 0$ and a neighbourhood U_x of x such that for all $y \in U_x$

$$|\nabla F(y)| \geq \text{Ł} |F(y) - F(x)|^\beta.$$

The relevance of the previous definition stems from two properties

- every real analytic function is a Łojasiewicz-function, see [2, 3], and
- if a solution to the ODE $\dot{x}_t = -\nabla F(x_t)$ stays local for a Łojasiewicz-function F, then (x_t) converges.

We provide the following analogue of the latter result for SGD:

Theorem: (see [1]) Let F be a Łojasiewicz-function with locally Lipschitz continuous derivative and $p \geq 2$. Suppose that for $n \in \mathbb{N}$

$$\gamma_n = C_\gamma n^{-\gamma} \quad \text{and} \quad \sigma_n = n^\sigma,$$

where $C_\gamma > 0$, $\gamma \in (\frac{1}{2}, 1]$ and $\sigma \in \mathbb{R}$. If

$$\frac{2}{3}(\sigma + 1) < \gamma \quad \text{and} \quad \frac{1}{2\gamma - \sigma - 1} < p,$$

then, on $\mathbb{L} \cap \mathbb{M}_\sigma^p$, the process (X_n) converges, almost surely, to a critical point of F (possibly a saddle point or a local maximum).

Moreover, as proved in [1], particular machine learning problems involving deep learning networks with analytic activation functions are related to objective functions F that are real analytic and thus Łojasiewicz-functions.

[1] F is said to posses a continuum of critical points, if there exists an injective mapping $\varphi : [0,1] \to \{\text{set of critical points of } F\}$.

References

[1] S. Dereich and S. Kassing, *Convergence of stochastic gradient descent schemes for Łojasiewicz-landscapes*, arXiv:2102.09385 [cs.LG].
[2] S. Łojasiewicz, *Sur le probléme de la division.*, Studia Math. **18** (1959), 87–136.
[3] S. Łojasiewicz, *Une propriété topologique des sous-ensembles analytiques réels*, Les Équations aux Dérivées Partielles (1963), 87–89.

Prevalence of Neural Collapse during the terminal phase deep learning training

David Donoho
(joint work with Vardan Papyan, XY Han)

Modern deep neural networks for image classification have achieved super-human performance. Yet, the complex details of trained networks have forced most practitioners and researchers to regard them as blackboxes with little that could be understood. This paper considers in detail a now-standard training methodology: driving the cross-entropy loss to zero, continuing long after the classification error is already zero. Applying this methodology to an authoritative collection of standard deepnets and datasets, we observe the emergence of a simple and highly symmetric geometry of the deepnet features and of the deepnet classifier; and we document important benefits that the geometry conveys – thereby helping us understand an important component of the modern deep learning training paradigm.

This is joint work with Vardan Papyan, U Toronto and XY Han, Cornell. It covers a paper which appeared in September 2021 in Proc Natl Acad Sci.We will also discuss several papers by theory researchers which appeared in response. We will also discuss our view of the current relationship between theory and practice in this field.

An ℓ_p theory of PCA and spectral clustering

Jianqing Fan
(joint work with Emmanuel Abbe, Kaizheng Wang)

Principal Component Analysis (PCA) is a fundamental tool in statistics and machine learning. Its applications range from factor analysis and tensor decomposition to blind deconvolution and manifold learning. The computational efficiency and statistical accuracy make PCA a top choice for analyzing massive data. While existing study of PCA focuses on the recovery of principal components and their associated eigenvalues, there are few precise characterizations of individual principal component scores that yield low-dimensional embedding of samples. Since all the downstream tasks account on the quality of embedding, the lack of investigation hinders the analysis of various spectral methods for community detection, clustering, ranking, synchronization and so on.

To analyze the performance of spectral methods, one often relies on the uniform (L_∞) control of errors across individual principal component scores. However, uniform control over all entries often leads to vacuum bounds if the sample size is too small or the signal is too weak. In that case, one can only hope to establish bounds for a reasonably large proportion of the entries based on more refined analysis. In this talk, we first develop an L_p perturbation theory for a hollowed version of PCA in reproducing kernel Hilbert spaces which provably improves upon the vanilla PCA in the presence of heteroscedastic noises. Through a novel L_p analysis of eigenvectors, we investigate entrywise behaviors of principal component score vectors and show that they can be approximated by linear functionals of the Gram matrix in L_p norm, which includes L_2 and L_∞ as two special cases. The entrywise analysis is formalized via the powerful leave-one-out decoupling technique.

We illustrate herewith the merits of the ℓ_p analysis using spectral clustering for a mixture of two Gaussians. Let $\boldsymbol{y} \in \{\pm 1\}^n$ be a label vector with i.i.d. Rademacher entries and $\boldsymbol{\mu} \in \mathbb{R}^d$ be a deterministic mean vector, both of which are unknown. Consider the model

$$\boldsymbol{x}_i = y_i \boldsymbol{\mu} + \boldsymbol{z}_i, \qquad i \in [n], \tag{1}$$

where $\{\boldsymbol{z}_i\}_{i=1}^n$ are i.i.d. $N(\boldsymbol{0}, \boldsymbol{I}_d)$ vectors. The goal is to estimate $\boldsymbol{y}$ from $\{\boldsymbol{x}_i\}_{i=1}^n$. Since $\mathbb{P}(y_i = 1) = \mathbb{P}(y_i = -1) = 1/2$, $\{\boldsymbol{x}_i\}_{i=1}^n$ are i.i.d. samples from a mixture of two Gaussians $\frac{1}{2}N(\boldsymbol{\mu}, \boldsymbol{I}_d) + \frac{1}{2}N(-\boldsymbol{\mu}, \boldsymbol{I}_d)$.

By construction, $\bar{\boldsymbol{X}} = (\bar{\boldsymbol{x}}_1, \cdots, \bar{\boldsymbol{x}}_n)^\top = \boldsymbol{y}\boldsymbol{\mu}^\top$ and $\bar{\boldsymbol{G}} = \|\boldsymbol{\mu}\|_2^2 \boldsymbol{y}\boldsymbol{y}^\top$ with $\bar{\boldsymbol{u}}_1 = \boldsymbol{y}/\sqrt{n}$ and $\bar{\lambda}_1 = n\|\boldsymbol{\mu}\|_2^2$. Hence, $\operatorname{sgn}(\boldsymbol{u}_1)$ becomes a natural estimator for $\boldsymbol{y}$, where $\operatorname{sgn}(\cdot)$ is the entrywise sign function. A fundamental question is whether the empirical eigenvector $\boldsymbol{u}_1$ is informative enough to accurately recover the labels in competitive regimes. To formalize the discussion, we denote by

$$\mathrm{SNR} = \frac{\|\boldsymbol{\mu}\|_2^4}{\|\boldsymbol{\mu}\|_2^2 + d/n} \tag{2}$$

the signal-to-noise ratio of model (1). Consider the challenging asymptotic regime where $n \to \infty$ and $1 \ll \mathrm{SNR} \lesssim \log n$[1]. The dimension d may or may not diverge. According to Theorem 3.2 in [1] , the spectral estimator $\operatorname{sgn}(\boldsymbol{u}_1)$ achieves the minimax optimal misclassification rate

$$e^{-\frac{1}{2}\mathrm{SNR}(1+o(1))}. \tag{3}$$

In order to get this result, we start from an ℓ_p analysis of $\boldsymbol{u}_1$. Theorem 3.3 in [1] shows that

$$\mathbb{P}(\min_{s=\pm 1} \|s\boldsymbol{u}_1 - \boldsymbol{G}\bar{\boldsymbol{u}}_1/\bar{\lambda}_1\|_p < \varepsilon_n \|\bar{\boldsymbol{u}}_1\|_p) > 1 - Ce^{-p} \tag{4}$$

[1] In Theorem 3.2 in [1], we derive results for the exact recovery of the spectral estimator, i.e. $\mathbb{P}(\operatorname{sgn}(\boldsymbol{u}_1) = \pm\boldsymbol{y}) \to 1$, when $\mathrm{SNR} \gg \log n$. Here we omit that case and discuss error rates.

for $p = \mathrm{SNR}$, some constant $C > 0$ and some deterministic sequence $\{\varepsilon_n\}_{n=1}^{\infty}$ tending to zero. On the event $\|s\boldsymbol{u}_1 - \boldsymbol{G}\bar{\boldsymbol{u}}_1/\bar{\lambda}_1\|_p < \varepsilon_n \|\bar{\boldsymbol{u}}_1\|_p$, we apply a Markov-type inequality to the entries of $(s\boldsymbol{u}_1 - \boldsymbol{G}\bar{\boldsymbol{u}}_1/\bar{\lambda}_1)$:

$$\frac{1}{n}|\{i: \ |(s\boldsymbol{u}_1 - \boldsymbol{G}\bar{\boldsymbol{u}}_1/\bar{\lambda}_1)_i| > \sqrt{\varepsilon_n/n}\}| \le \frac{\frac{1}{n}\sum_{i=1}^n |(s\boldsymbol{u}_1 - \boldsymbol{G}\bar{\boldsymbol{u}}_1/\bar{\lambda}_1)_i|^p}{(\sqrt{\varepsilon_n/n})^p}$$

$$\overset{\text{(i)}}{=} \left(\frac{\|s\boldsymbol{u}_1 - \boldsymbol{G}\bar{\boldsymbol{u}}_1/\bar{\lambda}_1\|_p}{\sqrt{\varepsilon_n}\|\bar{\boldsymbol{u}}_1\|_p}\right)^p \le \varepsilon_n^{p/2}, \tag{5}$$

where (i) follows from $\bar{\boldsymbol{u}}_1 = \boldsymbol{y}/\sqrt{n}$ and $\|\bar{\boldsymbol{u}}_1\|_p^p = n(1/\sqrt{n})^p$. Hence all but an $\varepsilon_n^{\mathrm{SNR}/2}$ fraction of $\boldsymbol{u}_1$'s entries are well-approximated by those of $\boldsymbol{G}\bar{\boldsymbol{u}}_1/\bar{\lambda}_1$. On the other hand, since the misclassification error is always bounded by 1, the exceptional event in (4) may at most contribute an $Ce^{-\mathrm{SNR}}$ amount to the final error. Both $\varepsilon_n^{\mathrm{SNR}/2}$ and $Ce^{-\mathrm{SNR}}$ are negligible compared to the optimal rate $e^{-\mathrm{SNR}/2}$ in (3). This helps us show that the ℓ_p bound (4) ensures sufficient proximity between $\boldsymbol{u}_1$ and $\boldsymbol{G}\bar{\boldsymbol{u}}_1/\bar{\lambda}_1$, and the analysis boils down to the latter term.

We now explain why $\boldsymbol{G}\bar{\boldsymbol{u}}_1/\bar{\lambda}_1$ is a good target to aim at. Observe that

$$(\boldsymbol{G}\bar{\boldsymbol{u}}_1)_i = [\mathcal{H}(\boldsymbol{X}\boldsymbol{X}^\top)\bar{\boldsymbol{u}}_1]_i = \sum_{j\neq i}\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle y_j/\sqrt{n} \propto \langle \boldsymbol{x}_i, \hat{\boldsymbol{\mu}}^{(-i)}\rangle, \tag{6}$$

where $\hat{\boldsymbol{\mu}}^{(-i)} = \frac{1}{n-1}\sum_{j\neq i}\boldsymbol{x}_j y_j$ is the leave-one-out sample mean. Consequently, the (unsupervised) spectral estimator $\mathrm{sgn}[(\boldsymbol{u}_1)_i]$ for y_i is approximated by $\mathrm{sgn}(\langle \boldsymbol{x}_i, \hat{\boldsymbol{\mu}}^{(-i)}\rangle)$, which coincides with the (supervised) linear discriminant analysis given additional labels $\{y_j\}_{j\neq i}$. This oracle estimator turns out to capture the difficulty of label recovery. That is, $\mathrm{sgn}(\boldsymbol{G}\bar{\boldsymbol{u}}_1/\bar{\lambda}_1)$ achieves the optimal misclassification rate in (3).

Above we provide high-level ideas about why the spectral estimator $\mathrm{sgn}(\boldsymbol{u}_1)$ is optimal. Inequality (4) ties $\boldsymbol{u}_1$ and its linearization $\boldsymbol{G}\bar{\boldsymbol{u}}_1/\bar{\lambda}_1$ together. The latter is connected to the genie-aided estimator through (6). As a side remark, the relation (6) hinges on the fact that $\boldsymbol{G}$ is hollowed. Otherwise there would be a square term $\langle \boldsymbol{x}_i, \boldsymbol{x}_i\rangle$ making things entangled.

We apply the newly developed perturbation theory to sub-Gaussian mixture models for clustering analysis and contextual stochastic block models for community detection. Intuitively, stronger signal allows for larger p in the L_p analysis and makes tighter error control possible. For the sub-Gaussian mixture model, our choice of p depends on the signal-to-noise ratio characterized by the separation between components, the sample size and the dimension. This adaptive choice yields optimality guarantees for spectral clustering. The misclassification rate is explicitly expressed as a simple exponential function of the signal-to-noise ratio, which implies exact recovery as a specific example. Perhaps surprisingly, the L_p analysis reveals intimate connections between the fully unsupervised spectral estimator and Fisher's linear discriminant analysis, which is a supervised classification procedure. Our results significantly improve upon prior arts which mostly focus on more complicated algorithms such as semidefinite programs or impose extra restrictions on the dimension and the signal strength.

In the contextual community detection problem, one observes both the network connections of nodes and their attributes. The network connections are modeled through a stochastic block model and the node attributes are modeled through a Gaussian mixture model that is independent of the network given the communities. The L_p theory and linearization of eigenvectors lead to a tuning-free aggregated spectral estimator that is conceptually simple and computationally efficient. Remarkably, it adaptively integrates the two sources of information based on their relative signal strengths. The estimator achieves the information threshold for exact recovery and has an optimal misclassification rate below that threshold. Moreover, our results readily imply optimal spectral clustering for the stochastic block model and Gaussian mixture model separately. Simulation experiments lend further support to our theoretical findings.

References

[1] E. Abbe, J. Fan, and K. Wang (2020). An ℓ_p theory of PCA and spectral clustering. *https://arxiv.org/abs/2006.14062.*

Exact Minimax Estimation for Phase Synchronization

Chao Gao
(joint work with Anderson Y. Zhang)

The phase synchronization problem is to estimate n unknown angles $\theta_1^*, \cdots, \theta_n^*$ from noisy measurements of $(\theta_j^* - \theta_k^*)$ mod 2π. In this paper, we consider the following additive model:

$$Y_{jk} = z_j^* \bar{z}_k^* + \sigma W_{jk} \in \mathbb{C}, \tag{1}$$

for all $1 \leq j < k \leq n$, where we use the notation $\bar{x}$ for the complex conjugate of x. We assume that each $z_j^* \in \mathbb{C}_1 = \{x \in \mathbb{C} : |x| = 1\}$ and we can thus write it as $z_j^* = e^{i\theta_j^*}$. The additive noise W_{jk} in (1) is assumed to be i.i.d. standard complex Gaussian.[1] Our goal in this paper is to study minimax optimal estimation of the vector $z^* \in \mathbb{C}_1^n$ under the loss function

$$\ell(\widehat{z}, z^*) = \min_{a \in \mathbb{C}_1} \sum_{j=1}^{n} |\widehat{z}_j a - z_j^*|^2. \tag{2}$$

We remark that the minimization over a global phase in the definition of (2) is necessary. This is because the global phase is not identifiable from the pairwise observations (1).

Various estimation procedures have been considered and studied in the literature. For example, the maximum likelihood estimator (MLE) is defined as the global maximizer of the following constrained optimization problem

$$\max_{z \in \mathbb{C}_1^n} z^{\mathrm{H}} Y z, \tag{3}$$

[1] For $W_{jk} \sim \mathcal{CN}(0,1)$, we have $\mathrm{Re}(W_{jk}) \sim \mathcal{N}\left(0, \frac{1}{2}\right)$ and $\mathrm{Im}(W_{jk}) \sim \mathcal{N}\left(0, \frac{1}{2}\right)$ independently.

where Y is Hermitian with $Y_{jk} = \bar{Y}_{kj}$ for all $1 \leq k < j \leq n$ and $Y_{jj} = 0$ for all $j \in [n]$. Note that (3) can be shown to be equivalent to $\min_{z \in \mathbb{C}_1^n} \sum_{1 \leq j < k \leq n} |Y_{jk} - z_j \bar{z}_k|^2$. It was shown in [1] that the MLE satisfies $\ell(\widehat{z}, z^*) \leq C\sigma^2$ with high probability for some constant $C > 0$. However, the optimization (3) is nonconvex and computationally infeasible in general. To address this problem, generalized power method (GPM) and semi-definite programming (SDP) have been considered in the literature to approximate the solution of (3). The generalized power method is defined through the iteration,[2]

$$z_j^{(t)} = \frac{\sum_{k \in [n] \setminus \{j\}} Y_{jk} z_k^{(t-1)}}{\left| \sum_{k \in [n] \setminus \{j\}} Y_{jk} z_k^{(t-1)} \right|}. \tag{4}$$

In other words, one repeatedly computes the product $Yz^{(t-1)}$ and projects this vector to $\mathbb{C}_1^n$ through entrywise normalization. When the iteration (4) is initialized by the eigenvector method, [2] shows that $z^{(t)}$ converges to the global maximizer of (3) at a linear rate under the noise level condition $\sigma^2 = O\left(\frac{n}{\log n}\right)$. The semidefinite programming is a convex relaxation of (3). It refers to the following optimization problem,

$$\max_{Z = Z^{\mathrm{H}} \in \mathbb{R}^{n \times n}} \mathsf{Tr}(YZ) \quad \text{subject to } \operatorname{diag}(Z) = I_n \text{ and } Z \succeq 0. \tag{5}$$

In general, the solution of (5) is an $n \times n$ matrix and needs to be rounded. When $\sigma^2 = O(n^{1/2})$, it was proved by [1] that the solution to (5) is a rank-one matrix $\widehat{Z} = \widehat{z}\widehat{z}^{\mathrm{H}}$, with $\widehat{z}$ being a global maximizer of (3). This result was recently proved by [2] to hold under a weaker condition $\sigma^2 = O\left(\frac{n}{\log n}\right)$. Given the fact that SDP solves (3), we know that it also achieves the same high-probability error bound $\ell(\widehat{z}, z^*) \leq C\sigma^2$ as that of the MLE under the additional condition $\sigma^2 = O\left(\frac{n}{\log n}\right)$.

Despite these estimation procedures studied in the literature, it remains an open problem what the optimal error under the loss (2) is. In this paper, we establish a minimax lower bound for phase synchronization. We show that

$$\inf_{\widehat{z} \in \mathbb{C}_1^n} \sup_{z \in \mathbb{C}_1^n} \mathbb{E}_z \ell(\widehat{z}, z) \geq (1 - \delta)\frac{\sigma^2}{2}, \tag{6}$$

for some $\delta = o(1)$ under the condition that $\sigma^2 = o(n)$. This provides a stronger characterization of the fundamental limits of the phase synchronization problem than the Cramér-Rao lower bound, which only holds for unbiased estimators. Instead, the lower bound in (6) holds for both unbiased and biased estimators. Moreover, in this paper, we prove the MLE, the GPM and the SDP all achieve the error bound

$$\ell(\widehat{z}, z^*) \leq (1 + \delta)\frac{\sigma^2}{2}, \tag{7}$$

[2] When the denominator of (4) is zero, take $z_j^{(t)}$ to be an arbitrary value in $\mathbb{C}_1$.

for some $\delta = o(1)$ with high probability under the same condition $\sigma^2 = o(n)$. In other words, these three estimators are not only rate-optimal, but are also exactly asymptotically minimax by achieving the correct leading constant in front of the optimal rate.

To formally state our main result, we introduce a more general statistical estimation setting that allows the possibility of missing entries. Instead of observing Y_{jk} for all $1 \leq j < k \leq n$, we assume each Y_{jk} is observed with probability p. In other words, consider a random graph $A_{jk} \sim \text{Bernoulli}(p)$ independently for all $1 \leq j < k \leq n$, and we only observe Y_{jk} that follows (1) when $A_{jk} = 1$. Define $A_{jk} = A_{kj}$ for $1 \leq k < j \leq n$ and $A_{jj} = 0$ for $j \in [n]$. The full observations can be organized into two Hermitian matrices A and $A \circ Y$, where $\circ$ denotes the matrix Hadamard product. The MLE, the GPM and the SDP can be extended by replacing Y_{jk} in (3), (4) and (5) with $A_{jk}Y_{jk}$.

Theorem 1.1. *Assume* $\sigma^2 = o(np)$ *and* $\frac{np}{\log n} \to \infty$. *Then, there exists some* $\delta = o(1)$ *such that*

$$\inf_{\widehat{z} \in \mathbb{C}_1^n} \sup_{z \in \mathbb{C}_1^n} \mathbb{E}_z \ell(\widehat{z}, z) \geq (1-\delta)\frac{\sigma^2}{2p}. \tag{8}$$

Moreover, MLE, GPM and SDP (the normalized leading eigenvector of the SDP solution) all achieve the error bound

$$\ell(\widehat{z}, z^*) \leq (1+\delta)\frac{\sigma^2}{2p}, \tag{9}$$

with probability at least $1 - n^{-1} - \exp\left(-\left(\frac{np}{\sigma^2}\right)^{1/4}\right)$.

Theorem 1.1 immediately implies (6) and (7) as a special case of $p = 1$, and is the first statistical analysis of phase synchronization for $p < 1$. We remark that both conditions $\sigma^2 = o(np)$ and $\frac{np}{\log n} \to \infty$ are essential for the results of the above theorem to hold. Since the minimax risk of the problem is $\frac{\sigma^2}{2p}$, the condition $\sigma^2 = o(np)$, which is equivalent to $\frac{\sigma^2}{2p} = o(n)$, guarantees that the minimax risk is of smaller order than the trivial one. The order n is trivial, since $\ell(z, z^*) \leq 4n$ for any $z, z^* \in \mathbb{C}_1^n$. When $p = 1$, the necessity of $\sigma^2 = o(n)$ for a nontrivial recovery is understood in the literature. The condition $\frac{np}{\log n} \to \infty$ guarantees that the random graph A is connected with high probability. It is known that when $p \leq c\frac{\log n}{n}$ for some sufficiently small constant $c > 0$, the random graph has several disjoint components, which makes the recovery of z^* up to a global phase impossible.

References

[1] Afonso S Bandeira, Nicolas Boumal and Amit Singer, *Tightness of the maximum likelihood semidefinite relaxation for angular synchronization*, Mathematical Programming **163** (2017), 145–167.

[2] Yiqiao Zhong and Nicolas Boumal, *Near-optimal bounds for phase synchronization*, SIAM Journal on Optimization **28** (2018), 989–1016.

Contrastive learning, multi-view redundancy, and linear models

Daniel Hsu
(joint work with Akshay Krishnamurthy, Christopher Tosh)

Self-supervised learning is an empirically successful approach to unsupervised learning based on creating artificial supervised learning problems. A popular self-supervised approach to representation learning is contrastive learning, which leverages naturally occurring pairs of similar and dissimilar data points, or multiple views of the same data. This work provides a theoretical analysis of contrastive learning in the multi-view setting, where two views of each datum are available. We first prove that linear functions of the learned representations are nearly optimal on downstream prediction tasks whenever the two views provide redundant information about the label. We also prove that, in the context of topic models (and other multi-view mixture models), the learned representation can be interpreted as a linear transformation of the posterior moments of the hidden topics given the words observed in a document.

References

[1] C. Tosh, A. Krishnamurthy, and D. Hsu. Contrastive estimation reveals topic posterior information to linear models. *arXiv:2003.02234*, 2020.

[2] C. Tosh, A. Krishnamurthy, and D. Hsu. Contrastive learning, multi-view redundancy, and linear models. In *International Conference on Algorithmic Learning Theory*, 2021.

Counting Cycles in Networks

Tracy Ke
(joint work with Jiashun Jin, Shengming Luo, Minzhe Wang, Wanjie Wang)

In many network models, the quantity of interest (community structure, mixed-memberships) is a low-rank signal matrix, masked by noise. The spectrum of the signal matrix plays a fundamental role in network analysis and is of major interest. We propose to recover the spectrum by counting short cycles in the adjacency matrix. The cycle counts provide a good estimate for the moments of the spectrum, which can thus be used to estimate the spectrum. Compared to empirical spectrum, the proposed estimators are more accurate in a wide range of parameter settings.

The idea can also be adapted to solve many other problems. One of such problems is global testing, where the goal is to test whether the network only has one community or multiple communities. We find that counting cycles with a centered adjacency matrix gives rise to an easy-to-use testing statistic that is asymptotically $N(0,1)$ under null and achieves the optimal phase transition. The test is competitive in a wide range of network settings, where we allow severe degree heterogeneity and mixed-memberships.

References

[1] J. Jin, Z. Ke, S. Luo, *Optimal adaptivity of signed-polygon statistics for network testing*, Annals of Statistics (to appear) (2021).

The Smoking Gun: Statistical Theory Improves Neural Network Estimates

Michael Kohler, Sophie Langer
(joint work with Alina Braun and Harro Walk)

Deep neural networks have achieved impressive results in various applications, e.g., in image classification (Krizhevsky, Sutskever and Hinton (2012)), text classification (Kim (2014)), machine translation (Wu et al. (2016)) and mastering of games (Silver et al. (2017)). Unfortunately, those results have been achieved without derivation of any mathematical or statistical theory of the estimates. Recently, quite a few papers were published dealing with the theoretical results behind deep learning. Approximation properties were analyzed, e.g., in Yarotsky (2017), Yarotsky and Zhevnerchuck (2020) and Lu et al. (2020)). Bauer and Kohler (2019), Schmidt-Hieber (2020) and Kohler and Langer (2021) considered deep neural network least squares estimates in a statistical setting and derived rate of convergence results. While those results partly explain the success of neural networks, they did not take into account all three aspects, namely approximation, generalization and optimization, simultaneously and could therefore not improve neural network estimates in applications. But should it not be the purpose of statistical theory to improve estimates in practice? In our talk we analyze neural networks with one hidden learned by gradient descent. This analysis considers all three aspects, namely approximation, generalization and optimization of deep learning theory, simultaneously and we are able to improve the performance of our estimates in practice. In particular, we analyze the L_2 error of neural network regression estimates with one hidden layer. Under the assumption that the Fourier transform of the regression function decays suitably fast, we show that an estimate, where all initial weights are chosen according to proper uniform distributions and where the weights are learned by gradient descent, achieves a rate of convergence of $1/\sqrt{n}$ (up to a logarithmic factor). Our statistical analysis implies that the key aspect behind this result is the proper choice of the initial inner weights and the adjustment of the outer weights via gradient descent. This indicates that we can also simply use linear least squares to choose the outer weights. We prove a corresponding theoretical result and compare our new linear least squares neural network estimate with standard neural network estimates via simulated data. Our simulations show that our theoretical considerations lead to an estimate with an improved performance. Hence the development of statistical theory can indeed improve neural network estimates. That is why we consider this result as the smoking gun of neural network theory.

References

[1] B. Bauer and M. Kohler *On deep learning as a remedy for the curse of dimensionality in nonparametric regression*, Annals of Statistics **47**, 2261–2285.

[2] Y. Kim, *Convolutional Neural Networks for Sentence Classification* arXiv: 1408.5882 (2014).

[3] M. Kohler and S. Langer *On the rate of convergence of fully connected very deep neural network regression estimates using ReLU activation functions*, arXiv: 1908.11133, To appear in Annals of Statistics.

[4] A. Krizhevsky, I. Sutskever and G.E. Hinton, *ImageNet classification with deep convolutional neural networks* In F. Pereira et al. (Eds.), *Advances In Neural Information Processing Systems* **25** (2012), 1097–1105. Red Hook, NY: Curran.

[5] J. Lu, Z. Shen, H. Yang and S. Zhang, *Deep network approximation for smooth functions* arXiv: 2001.03040 (2020).

[6] J. Schmidt-Hieber, *Nonparametric regression using deep neural networks with ReLU activation function (with discussion)*, Annals of Statistics **48** (2020), 1875–1897.

[7] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Huber et al. *Mastering the game of go without human knowledge.* Nature **550** (2017), 354–359.

[8] Y. Wu, M. Schuster, Z. Chen, Q. Le, M. Norouzi, W. Macherey, M. Krikum et al. *Google's neural machine translation system: Bridging the gap between human and machine translation*, arXiv: 1609.08144.

[9] D. Yarotsky, *Optimal approximation of continuous functions by very deep ReLU networks*, COLT **75** (2018), 639–649.

[10] D. Yarotsky and A. Zhevnerchuk, *The phase diagram of approximation rates of deep neural networks*, arXiv: 1906.09477 (2020).

Transferability of spectral graph convolutional neural networks

Gitta Kutyniok

(joint work with Michael Bronstein and Ron Levie)

In many applications in data science the data is represented by graphs. Graph convolutional networks (CNNs), which are extensions of standard CNNs to graph structured data, have achieved resounding success in the past few years. In a standard CNN, the network receives a signal defined over a Euclidean rectangle, and at each layer applies a set of convolutions/filters, an activation function, and, optionally, pooling. A graph CNN has the same architecture, with the only difference that signals are defined over the vertices of graph domains. In a machine learning setting, the general architecture of the CNN is fixed, but the specific filters to use in each layer are free parameters. In training, the filter coefficients are optimized to minimize some loss function. In some situations, the data consists of many different graphs, and many different signals on these graphs (multi-graph setting). In these situations, if two graphs represent the same underlying phenomenon, and the two signals given on the two graphs are similar in some sense, the output of the CNN on both signals should be similar as well. This property is typically termed *transferability*, and is an essential requirement if we wish the CNN to generalize well on the test set in multi-graph settings. In fact, transferability can be seen as a special type of generalization capability. Analyzing and proving transferability of spectral graph CNNs is the focus of this talk.

Graph CNNs can achieve transferability in different ways, and we consider two categories of such ways. In *concept-based transferability*, when a graph CNN is exposed to a multi-graph training set, it can learn "concepts" that promote transferability. On the other hand, *principle transferability* is the built-in capability of graph CNNs to generalize between graphs that represent the same phenomenon, independently of training and of specific filters. The latter approach is the focus of this talk, which is based on [1].

Convolution operators on graphs. The definition of spectral convolution on graphs is inspired by the convolution theorem in Euclidean domains, that states that convolution in the spatial domain is equivalent to pointwise multiplication in the frequency domain. To define the frequency domain of a graph, we consider the (self-adjoint) graph Laplacian $\mathbf{\Delta}$, and use its eigenvalues as frequencies and its eigenvectors as the corresponding Fourier modes. Graph filters $\mathbf{F}$ are defined via a *functional calculus* implementation, where the frequency responses are parameterized by a function $f : \mathbb{R} \to \mathbb{C}$. Namely, given a graph signal $\mathbf{s}$,

$$\mathbf{F}\mathbf{s} = f(\mathbf{\Delta})\mathbf{s} := \sum_{n=1}^{N} f(\lambda_n)(\boldsymbol{\psi}_n^* \cdot \mathbf{c})\boldsymbol{\psi}_n \tag{1}$$

where $\{\boldsymbol{\psi}_n\}_{n=1}^N$ are the eigenvectors of $\mathbf{\Delta}$, λ_n are the eigenvalues, and $\boldsymbol{\psi}_n^*$ is the conjugate transpose of $v\boldsymbol{\psi}_n$. Here, the scalars $\{f(\lambda_n)\}_{n=1}^N$ are the frequency responses of the filter. Functional calculus filters are computationally efficient, linearly stable with respect to perturbations in the graph [2].

The majority of researchers from the graph CNN community currently focus on developing spatial methods. One typical motivation for favoring spatial methods is the claim that spectral methods are not transferable, and thus do not generalize well on graphs unseen in the training set. The goal in this talk is to debunk this misconception, and to show that state-of-the-art spectral graph filtering methods are transferable. Interestingly, [3] showed in an extensive study that spectral graph CNNs obtain state-of-the-art results in well known multi-graph benchmarks.

Principle transferability of spectral graph CNNs. We present a framework of transferability, allowing to compare graphs of incompatible sizes and topologies. To accommodate the comparison of incompatible graphs, our approach resorts to non-graph theoretical considerations, assuming that graphs are observed from some underlying non-graph spaces. In our approach, graphs are regarded as discretizations of underlying corresponding "continuous" Borel spaces. This makes sense, since a weighted graph can be interpreted as a set of points (vertices) and a decreasing function of their distances (edge weights). As a basic assumption, two graphs are comparable, or represent the same phenomenon, if both discretize the same space. This approach allows us to prove transferability under small perturbations of the adjacency matrix, but more generally, allows us to prove transferability between graphs with incompatible structures. In the following we present a simplified adaptation of our results, where graphs are discretized from metric measure spaces via sampling.

The way to compare two graphs is to consider their embeddings to the metric space they both discretize. For intuition, consider the special case where the metric space is a manifold. Any manifold can be discretized to a graph/polygon-mesh in many different ways, resulting in different graph topologies. A filter designed/learned on one polygon-mesh should have approximately the same repercussion on a different polygon-mesh discretizing the same manifold. To compare the filter on the two graphs, we consider a generic signal defined on the continuous space, and sampled to both graphs. After applying the graph filter on the sampled signal on both graphs, we interpolate both results back to two continuous signals. In our analysis we show that these two interpolated continuous signals are approximately equal. To this end, we develop a digital signal processing (DSP) framework akin to the classical Nyquist–Shannon approach, where now analog domains are metric-measure spaces, and digital domains are graphs.

Consider a metric space $\mathcal{M}$ with a Borel measure μ, and take the space of signals of $\mathcal{M}$ as $L_2(\mathcal{M})$. Consider a self-adjoint operator $\mathcal{L}$ in $L_2(\mathcal{M})$ that we call the *metric Laplacian*. We suppose that $\mathcal{L}$ has a discrete spectrum, with eigenvalues $\lambda_0 < \lambda_1 < \dots$ and corresponding eigenfunctions $\phi_n : \mathcal{M} \to \mathbb{C}$. The metric-Laplacian models the geometry in $\mathcal{M}$. We define band-limited spaces in $L_2(\mathcal{M})$ (Paley-Wiener spaces) by $PW(\lambda_M) = \mathrm{span}\{\phi_m\}_{m=0}^{M}$. Denote by $P(\lambda_M)$ the orthogonal projection upon $PW(\lambda_M)$.

Graphs are sampled from metric spaces by sampling nodes as points in $\mathcal{M}$. We consider a set of graphs $\{G_n\}_n$ with N_n nodes each. Given N_n sample points $G^n = \{x_k^n\}_{k=1}^{N_n} \subset \mathcal{M}$, the sampling operator $S_n : C(\mathcal{M}) \to L_2(G^n)$ is defined by $S_n s = \{s(x_K^n)\}_{k=1}^{N_n}$ for any continuous metric space signal $s \in C(\mathcal{M})$. We define the interpolation of between $L_2(G^n)$ and $PW(\lambda_M)$ as the adjoint operator of the operator $S_n P(\lambda_M)$. Namely, $I_{n;\lambda_M} = \big(S_n P(\lambda_M)\big)^*$. Note that the term *interpolation* is adopted here from the classical Nyquist–Shannon DSP theory. However, $I_{n;\lambda_M}$ only approximates the values at the nodes, and does not interpolate accurately.

Now, consider two graphs G_1 and G_2, with corresponding graph Laplacians $\mathbf{\Delta}_1$ and $\mathbf{\Delta}_2$, that represent the same phenomenon. Adopting our basic assumption, we thus suppose that both graphs approximate the metric space $\mathcal{M}$ in the following sense. For some fixed Paley-Wiener space $PW(\lambda_M)$, and for each $n = 1, 2$ and any metric space signal $s \in PW(\lambda_M)$, we have $\|\mathcal{L}s - I_{n;\lambda_M}\mathbf{\Delta}_n S_n s\| \approx 0$. By the triangle inequality, we can also show

$$\|I_{1;\lambda_M}\mathbf{\Delta}_1 S_1 s - I_{2;\lambda_M}\mathbf{\Delta}_2 S_2 s\| \approx 0. \tag{2}$$

The following theorem proves in this situation that any Lipschitz continuous functional calculus filter f is linearly stable in the perturbation error (2).

Theorem 1. *Consider the above construction, and let $\lambda_M > 0$ be a band with $\|I_{n;\lambda_M}\| < C$ for $n = 1, 2$. Let $f : \mathbb{R} \to \mathbb{C}$ be a Lipschitz continuous function, with*

Lipschitz constant D, and denote $\|f\|_{\mathcal{L},M} = \max_{0\leq m\leq M}\{|f(\lambda_m)|\}$. Then

$$\begin{aligned}\|f(\mathcal{L})P(\lambda_M) - I_{n;\lambda_M} f(\mathbf{\Delta}_n)S_nP(\lambda_M)\| \leq & DCM\,\|S_n\mathcal{L}P(\lambda_M) - \mathbf{\Delta}_nS_nP(\lambda_M)\| \\ & + \|f\|_{\mathcal{L},M}\left\|P(\lambda_M) - I_{n;\lambda_M}S_n^{\lambda_M}P(\lambda_M)\right\|.\end{aligned} \tag{3}$$

As a result of (3) and by the triangle inequality, we have that $\|I_{1;\lambda_M}f(\mathbf{\Delta}_1)S_1P(\lambda_M) - I_{2;\lambda_M}f(\mathbf{\Delta}_2)S_2P(\lambda_M)\|$ is linearly stable with respect to $\|I_{1;\lambda_M}\mathbf{\Delta}_1S_1P(\lambda_M) - I_{2;\lambda_M}\mathbf{\Delta}_2S_2P(\lambda_M)\|$ and $\max_{n=1,2}\left\|P(\lambda_M) - I_{n;\lambda_M}S_n^{\lambda_M}P(\lambda_M)\right\|$. Last, we can extend the transferability of filters property to a transferability of sprectral graph CNNs property.

To show that filters are transferable via (3), one must first show that the Laplacians are transferable. For this, we prove that graph Laplacians which are randomly sampled from metric space Laplacian are transferable in high probability.

References

[1] R. Levie, W. Hang, L. Bucci, M. M. Bronstein and G. Kutyniok, *Transferability of Spectral Graph Convolutional Neural Networks*, arXiv:1907.12972 [cs.LG], 2019.

[2] R. Levie, E. Isufi and G. Kutyniok, *On the Transferability of Spectral Graph Filters*, in 2019 13th International conference on Sampling Theory and Applications (SampTA), 2019.

[3] A. Nilsson and X. Bresson, *An Experimental Study of the Transferability of Spectral Graph Networks*, arXiv:2012.10258 [cs.LG], 2020.

On polynomial-time computation of high-dimensional posterior measures by Langevin-type algorithms

Richard Nickl
(joint work with Sven Wang)

The problem of generating random samples of high-dimensional posterior distributions is considered. The main results consist of non-asymptotic computational guarantees for Langevin-type MCMC algorithms which scale polynomially in key quantities such as the dimension of the model, the desired precision level, and the number of available statistical measurements. As a direct consequence, it is shown that posterior mean vectors as well as optimisation based maximum a posteriori (MAP) estimates are computable in polynomial time, with high probability under the distribution of the data. These results are complemented by statistical guarantees for recovery of the ground truth parameter generating the data.

Our results are derived in a general high-dimensional non-linear regression setting (with Gaussian process priors) where posterior measures are not necessarily log-concave, employing a set of local 'geometric' assumptions on the parameter space, and assuming that a good initialiser of the algorithm is available. The theory is applied to a representative non-linear example from PDEs involving a steady-state Schrödinger equation.

References

[1] Richard Nickl, Sven Wang, *On polynomial-time computation of high-dimensional posterior measures by Langevin-type algorithms*, arXiv:2009.05298

Online Matching in Sparse Random Graphs

Vianney Perchet

(joint work with Nathan Noiry, Flore Sentenac)

Motivated by sequential budgeted allocation problems, we investigate online matching problems where connections between vertices are not i.i.d., but they have fixed degree distributions – the so-called configuration model. We estimate the competitive ratio of the simplest algorithm, "greedy", by approximating some relevant stochastic discrete processes by their continuous counterparts, that are solutions of an explicit system of partial differential equations. This technique gives precise bounds on the estimation errors, with arbitrarily high probability as the problem size increases.

More precisely, we assume that the degree distribution on one side (the $\mathcal{U}$-side) has a generating function $\phi_{\mathcal{U}}$ and of expectation $\mu_{\mathcal{U}}$ and, on the other side (the $\mathcal{V}$-side), has a generating function $\phi_{\mathcal{V}}$ and of expectation $\mu_{\mathcal{V}}$.

The main result is that given $N \geq 1$ and $T = \frac{\mu_{\mathcal{U}}}{\mu_{\mathcal{V}}} N$, let M_T be the matching built by "Greedy" on the configuration model induced by the above degree distributions; then the following convergence holds in probability:

$$\frac{|\mathrm{M}_T|}{N} \xrightarrow[N \to +\infty]{\mathbf{P}} 1 - \phi_{\mathcal{V}}(1 - G(1)).$$

where G is the unique solution of the following ordinary differential equation:

$$G'(s) = \frac{1 - \phi_{\mathcal{U}}\left(1 - \frac{1}{\mu_{\mathcal{U}}}\phi'_{\mathcal{U}}\left(1 - G(s)\right)\right)}{\frac{\mu_{\mathcal{V}}}{\mu_{\mathcal{U}}}\phi'_{\mathcal{U}}(1 - G(s))}; \quad G(0) = 0.$$

Moreover, for any $s \in [0, 1]$, if $M_T(s)$ is the matching obtained by "greedy" after seeing a proportion s of vertices of $\mathcal{V}$, then

$$\frac{|\mathrm{M}_T(s)|}{N} \xrightarrow[N \to +\infty]{\mathbf{P}} 1 - \phi_{\mathcal{U}}(1 - G(s)).$$

Convergence rates are explicit; with probability exponentially large, at least $1 - \zeta N \exp(-\xi N^{c/2})$,

$$\sup_{s \in [0,1]} \left| \frac{|\mathrm{M}_T(s)|}{N} - \left(1 - \phi_{\mathcal{U}}(1 - G(s))\right) \right| \leq \kappa N^{-c},$$

where ζ, ξ, κ depend only on the (first two) moments of both $\pi_{\mathcal{V}}$ and $\pi_{\mathcal{U}}$, and c is some universal constant (set arbitrarily as 1/20 in the proof).

Forecasting time series with neural networks

Stefan Richter

(joint work with Nathawut Phandoidaen, Moritz Haas)

Given is a high-dimensional stationary time series $X_1, ..., X_n \in \mathbb{R}^d$. The goal we aim to investigate is the prediction of X_{n+1} or low-dimensional statistics $T(X_{n+1})$ given the past lags $X_n, ..., X_{n-r}$, where $r \in \mathbb{N}$. We investigated two basic approaches:

(1) Find f such that $X_{n+1} \approx f(X_n, ..., X_{n-r+1})$ (which leads to point forecasts)

(2) Find distribution F such that $F \overset{d}{\approx} \mathbb{P}^{X_{n+1}|X_n,...,X_{n-r+1}}$ (which leads to distributional forecasts)

For both inference of f and F we consider approaches with neural networks and provide statistical guarantees under conditions on the class of neural networks.

1. Point forecasting

We consider the simple model

$$X_t = f^*(X_{t-1}) + \varepsilon_t, \quad t = 1, ..., n,$$

where $f^* : \mathbb{R}^d \to \mathbb{R}^d$ and ε_t is i.i.d. Gaussian noise with $\mathbb{E}\varepsilon_t = 0$. Estimation of f^* is performed with neural networks $\hat{f}$, and quality assessment via the prediction error $\mathbb{E}R(\hat{f})$, where $|\cdot|$ is the Euclidean norm and

$$R(f) := \frac{1}{d}\mathbb{E}[|X_1 - f(X_0)|_2^2].$$

We pose the following encoder-decoder assumption on f^* which mimics the idea that the time series evolution takes place via a compression of the information of the state before and is afterwards 'spread out' again to all components.

Assumption 1. *$f^* = f^*_{dec} \circ f^*_{enc}$, where $f^*_{enc} = g_2 \circ g_1$,*

- *$g_1 : \mathbb{R}^d \to \mathbb{R}^D$, and any component of g_1 only depends on $\tilde{d} \ll d$ components and is in C^β*
- *$g_2 : \mathbb{R}^D \to \mathbb{R}^{\tilde{d}}$ with $\tilde{d} \ll d$ is C^∞,*
- *$f^*_{dec} : \mathbb{R}^{\tilde{d}} \to \mathbb{R}^d$ is in C^β.*

For estimation of f^*, we consider neural networks which are defined as follows (cf. [4]) Let $\sigma : \mathbb{R} \to \mathbb{R}$ be some activation function, e.g. $\sigma(x) = \max\{x, 0\}$.

Definition 1.

$$\mathcal{F}(L,p,s) := \{g : \mathbb{R}^d \to \mathbb{R}^d \,|\, g \text{ is a network with } L \text{ layers, width vector } p \text{ and sparsity level } s\},$$

where $g \in \mathcal{F}(L,p,s)$ has the form

$$g(x) = W^{(L+1)} \cdot \sigma(v^{(L)} + W^{(L)} \cdot \sigma(...W^{(2)} \cdot \sigma(v^{(1)} + W^{(1)}x)...)),$$

$W^{(l)} \in \mathbb{R}^{p_l \times p_{l-1}}$, $v^{(l)} \in \mathbb{R}^{p_l}$, and $\sum_{l=1}^{L+1} \left\{\|W^{(l)}\|_0 + \|v^{(l)}\|_0\right\} \leq s$.

To adopt for the encoder-decoder structure of f^*, we ask $\mathcal{F}(L,p,s)$ to have a layer with only $\tilde{d}$ dimensions as follows.

Definition 2. $\mathcal{F}(L,p,s,J,\tilde{d}) := \{f \in \mathcal{F}(L,p,s) \,|\, \textit{Layer } J \textit{ has } p_J = \tilde{d} \textit{ dimensions}\}$.

The empirical risk minimizer of $R(f)$ connected to this class reads now

$$\hat{f}_n :\in \operatorname{argmin}_{f\in\mathcal{F}(L,p,s,J,\tilde{d})}\hat{R}_n(f), \qquad \hat{R}_n(f) := \frac{1}{nd}\sum_{i=2}^{n}|X_i - f(X_{i-1})|_2^2.$$

It should be noted that in practice, $\hat{f}_n$ is approximated by a stochastic gradient descent method optimizer $\hat{f}_n^{\approx}$, where the quality of $\hat{f}_n^{\approx} \approx \hat{f}_n$ is current research. Therefore the theoretical results derived in this report are not directly applicable in practice, but they allow for a rough idea how the structure of the network has to be chosen. We obtained the following result (cf. [2]), which is based on oracle inequalities and empirical process theory based on [3] and the use of approximation results from [4].

Theorem 1.1. *Suppose that the β-mixing or functional dependence coefficients $\delta(j)$ of X_i satisfy $\delta(j) \le Cj^{-\alpha}$ $(\alpha, C > 1)$. Let*

$$\phi_n = n^{-\frac{2\beta}{2\beta+\tilde{d}}}.$$

Suppose that

- *Number of layers:* $\log_2(4\max\{\tilde{d},\beta\})\log_2(n) \le J \le L \lesssim \log(n)$,
- *Layer size:* $n\phi_n \lesssim \min_{l\in\{2,\dots,L-1\}\setminus\{J\}} p_l$,
- *Number of nonzero weights:* $s \asymp n\phi_n \log(n)$.

Then

$$\mathbb{E}R(\hat{f}) - R(f^*) \lesssim \log(n)^3\phi_n^{\frac{\alpha}{\alpha+1}} = \log(n)^3 n^{-\frac{\alpha}{\alpha+1}\frac{2\beta}{2\beta+\tilde{d}}}.$$

The important result is that the exponent of the nonparametric rate does not depend on the dimension d of the time series, but only on the compression dimension $\tilde{d}$. The strength of the polynomial dependence comes into play with an additional factor $\frac{\alpha}{\alpha+1}$ which lies between $\frac{1}{2}$ and 1. It is not clear up to now if this rate is optimal.

2. Distributional forecasting

The original idea is based on WGANs from machine learning. Given is a latent space $\mathbb{R}^{d_Z}$ and user-generated variables $Z_1, ..., Z_n \sim \mathbb{P}^Z$ with a chosen distribution $\mathbb{P}^Z$ independent of $X_1, ..., X_n$. For simplicity, the aim is to forecast the distribution of some statistics $T(X_1) \in \mathbb{R}^{d_T}$ given X_0. This is done by defining an estimator $\hat{g} : \mathbb{R}^{d_Z} \times \mathbb{R}^d \to \mathbb{R}^{d_T}$ which minimizes the 1-Wasserstein distance

$$W(\mathbb{P}^{T(X_1),X_0}, \quad \mathbb{P}^{g(Z,X_0),X_0})$$

over a certain class of functions g. If such an $\hat{g}$ is found, then $\{\hat{g}(Z_i, x) : i = 1, ..., N\}$ with user-generated variables $Z_1, ..., Z_N$ mimics the conditional distribution $T(X_1)|X_0 = x$, which enables us to provide distributional forecasts of $T(X_1)$

given $X_0 = x$ (*). We define $\hat{g}$ as follows: First, the Kantorovich formulation of the 1-Wasserstein distance is used:

$$W(\mathbb{P}_1, \mathbb{P}_2) = \sup_{f:\mathbb{R}^d \to \mathbb{R}, \|f\|_L \le 1} \Big\{ \int f d\mathbb{P}_1 - \int f d\mathbb{P}_2 \Big\},$$

where $\|f\|_L$ denotes the Lipschitz constant of a function f. This distance is approximated by a supremum over a class of neural networks ('critic networks') $\mathcal{F}(L_f, p_f, s_f)$,

$$W_n(g) = \sup_{f \in \mathcal{F}(L_f, p_f, s_f), \|f\|_L \le 1} \{\mathbb{E}f(T(X_1), X_0) - \mathbb{E}f(g(Z, X_0), X_0)\}.$$

The estimator is obtained via minimization of the corresponding empirical version over a class of 'generator networks' $\mathcal{F}(L_g, p_g, s_g)$,

$$\begin{aligned} \hat{g}_n \quad &:\in \quad \mathrm{argmin}_{g \in \mathcal{F}(L_g, p_g, s_g)} \hat{W}_n(g), \\ \hat{W}_n(g) \quad &:= \quad \sup_{f \in \mathcal{F}(L_f, p_f, s_f), \|f\|_L \le 1} \Big\{ \frac{1}{n} \sum_{i=2}^{n} f(T(X_i), X_{i-1}) - \frac{1}{n} \sum_{i=2}^{n} f(g(Z_i, X_{i-1}), X_{i-1}) \Big\} \end{aligned}$$

Again, $\hat{g}_n$ is an empirical risk minimizer which is not available in practice but is approximated by min-max-gradient descent methods. Therefore, our results should only be viewed as a first step towards a full theory. If there exists $g^* : \mathbb{R}^{d_Z} \times \mathbb{R}^d \to \mathbb{R}^{d_T}$ such that $\mathbb{P}^{(g^*(Z,X_0),X_0)} = \mathbb{P}^{(T(X_1),X_0)}$, fast convergence rates for $\hat{g}_n$ can be obtained by posing encoder-decoder assumptions on g^* as follows:

Assumption 2. *$g^* = g_{dec} \circ g_{enc}$, where $g_{enc} = g_{enc,1} \circ g_{enc,0}$, where*

- *$g_{enc,0} : \mathbb{R}^{d+d_Z} \to \mathbb{R}^D$, and any component of g_1 only depends on $d_g \ll d$ components and is in C^β*
- *$g_{enc,1} : \mathbb{R}^D \to \mathbb{R}^{d_g}$ is in C^∞ and $d_g \ll d + d_Z$,*
- *$g_{dec} : \mathbb{R}^{d_g} \to \mathbb{R}^{d_T}$ is in C^β.*

In [1], we proved the following result:

Theorem 2.1. *Let $\phi_n = n^{-\frac{2\beta}{2\beta+d_g}}$. Suppose that*

(i) $L_g \asymp \log(n)$,
(ii) $\min_{l=1,\ldots,L_g} p_{g,l} \gtrsim n\phi_n$,
(iii) $s_g \asymp n\phi_n \log(n)$
(iv) $L_f \le L_g$, $s_f \le s_g$.

Suppose for the β-mixing coefficients of X_i that $\beta(k) \le \kappa \cdot k^{-\alpha}$ ($\kappa, \alpha > 1$). Then

$$\mathbb{E}W_n(\hat{g}_n) = \mathbb{E}W_n(\hat{g}_n) - W_n(g^*) \lesssim \Big(\frac{s_f L_f \log(s_f L_f)}{n}\Big)^{1/2} + \phi_n^{1/2} \log(n)^{3/2},$$

If now $\mathcal{F}(L_f, p_f, s_f)$ is chosen 'large enough', then the above result implies the weak convergence $(\hat{g}_n(X_0), X_0) \xrightarrow{d} (T(X_1), X_0)$ which in turn justifies (*) above (cf. [1]).

References

[1] M. Haas and S. Richter. Statistical analysis of Wasserstein GANs with applications to time series forecasting, 2020, arXiv 2011.03074

[2] N. Phandoidaen and S. Richter. Forecasting time series with encoder-decoder neural networks, 2020, arXiv 2009.08848

[3] N. Phandoidaen and S. Richter. Empirical process theory for locally stationary processes, 2021, arXiv 2007.05737

[4] J. Schmidt-Hieber "Nonparametric regression using deep neural networks with ReLU activation function," The Annals of Statistics, Ann. Statist. 48(4), 1875-1897, (August 2020)

Metropolis-Hastings via Classification

Veronika Rockova
(joint work with Tetsuya Kaji)

This paper develops a Bayesian computational platform at the interface between posterior sampling and optimization in models whose marginal likelihoods are difficult to evaluate. Inspired by adversarial optimization, namely Generative Adversarial Networks (GAN), we reframe the likelihood function estimation problem as a classification problem. Pitting a Generator, who simulates fake data, against a Classifier, who tries to distinguish them from the real data, one obtains likelihood (ratio) estimators which can be plugged into the Metropolis-Hastings algorithm. The resulting Markov chains generate, at a steady state, samples from an approximate posterior whose asymptotic properties we characterize. Drawing upon connections with empirical Bayes and Bayesian mis-specification, we quantify the convergence rate in terms of the contraction speed of the actual posterior and the convergence rate of the Classifier. Asymptotic normality results are also provided which justify inferential potential of our approach. We illustrate the usefulness of our approach on simulated data.

Adaptive transfer learning

Richard J. Samworth
(joint work with Henry W. J. Reeve and Timothy I. Cannings)

In transfer learning, we wish to make inference about a target population when we have access to data both from the distribution itself, and from a different but related source distribution. We introduce a flexible framework for transfer learning in the context of binary classification, allowing for covariate-dependent relationships between the source and target distributions that are not required to preserve the Bayes decision boundary. Our main contributions are to derive the minimax optimal rates of convergence (up to poly-logarithmic factors) in this problem, and show that the optimal rate can be achieved by an algorithm that adapts to key aspects of the unknown transfer relationship, as well as the smoothness and tail parameters of our distributional classes. This optimal rate turns out to have several regimes, depending on the interplay between the relative sample sizes and the strength of the transfer relationship, and our algorithm achieves optimality by careful, decision tree-based calibration of local nearest-neighbour procedures.

Elephant in the Room: Non-Smooth Non-Convex Optimization

Ohad Shamir

It is well-known that finding global minima of non-convex optimization problems is computationally hard in general. However, the problem of finding stationary-like points (at least in terms of making the gradient small) is tractable even with simple gradient-based methods, and received much attention in recent years (e.g., Nesterov [5], Jin et al. [2], Carmon et al. [1]). The resulting literature has been largely motivated by the rising importance of non-convex problems such as deep learning, but in fact, does not quite address them: Nearly all positive results in this area require the objective function to be either smooth or have other structural properties which are seldom satisfied in deep learning problems. This highlights the importance of understanding what we can do efficiently on non-convex, non-smooth optimization problems.

In the talk, we described some results, challenges, and possible approaches to tackle this fundamental question. We began by revisiting the recent paper of Zhang et al. [6], which pointed out that minimizing the gradient norm is not possible in the non-smooth setting, and proposed an alternative notion of (δ, ϵ)-stationarity[1], along with computationally efficient methods which provably find such points. However, this notion can also lead to counter-intuitive behavior, at least in some cases: There are functions and points which are stationary-like under this definition, but do not resemble stationary points, and with all gradients in a δ-neighborhood being large.

We then proceeded to examine two alternative approaches, with other trade-offs in terms of computational efficiency and performance:

- First, we studied the notion of getting δ-close to points whose gradient norm is less than ϵ. Although intuitive, we showed a strong impossibility result in a standard oracle complexity framework [4], implying that under mild conditions, any algorithm with non-trivial guarantees will have oracle complexity exponential in the dimension.
- Second, we considered the approach of reduction to the smooth case: Namely, given a function f, find a smooth approximation $\tilde{f}$, which ϵ-approximates f (uniformly over $\mathbb{R}^d$) and has Lipschitz gradients, and find approximately stationary points with respect to $\tilde{f}$. Interestingly, for non-convex functions, there appears to be a trade-off between performance and computational tractability in computing such smooth approximations: On the one hand, there are very simple and computationally efficient methods (such as convolution with a smooth distribution function), that lead to the gradient Lipschitz parameter scaling with the dimension. On the other

[1] Namely, points where the convex hull of gradients in a δ-neighborhood contains vectors whose norm is less than ϵ

hand, there are essentially optimal methods with dimension-free guarantees (in particular, Lasry-Lions regularization [3]), that seem computationally intractable. In upcoming work, we prove that this trade-off is necessary, again in an oracle complexity framework: Under mild assumptions, to get any dimension-dependence better than standard convolutions, the oracle complexity must be exponential in the dimension – hence ruling out computational tractability.

Overall, we argue that theoretically understanding nonsmooth nonconvex optimization is an intersting and still relatively unexplored area, with different criteria leading to different trade-offs in terms of computational efficiency, performance and plausability. Besides the general question of which criterion will prove most suitable, there are also quite a few specific open questions, such as more precisely characterizing the oracle complexity for each of the settings we considered.

References

[1] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, pages 1–50, 2019.

[2] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1724–1732. JMLR. org, 2017.

[3] Jean-Michel Lasry and Pierre-Louis Lions. A remark on regularization in hilbert spaces. *Israel Journal of Mathematics*, 55(3):257–266, 1986.

[4] Arkadii Semenovich Nemirovsky and David Borisovich Yudin. *Problem complexity and method efficiency in optimization.* Wiley, 1983.

[5] Yurii Nesterov. How to make the gradients small. *Optima. Mathematical Optimization Society Newsletter*, (88):10–11, 2012.

[6] Jingzhao Zhang, Hongzhou Lin, Suvrit Sra, and Ali Jadbabaie. On complexity of finding stationary points of nonsmooth nonconvex functions. *arXiv preprint arXiv:2002.04130*, 2020.

On minimum ℓ_1-norm interpolation

Sara van de Geer

(joint work with Geoffrey Chinot, Felix Kuchelmeister, Matthias Löffler)

We consider the classification problem, where one observes an input matrix $X \in \mathbb{R}^{n\times p}$ and a binary response $Y \in \{\pm 1\}^n$ given by $Y = \text{sign}(X\beta^* + \xi)$. The unknown vector $\beta^* \in \mathbb{R}^p$ is normalized to have ℓ_2-norm $\|\beta^*\|_2 = 1$ and $\xi \in \mathbb{R}^n$ is an unobservable noise vector. Aim is to estimate β^* and build from this a classification rule for predicting the label of an unlabelled observation. We study the max-margin classifier, which is a value of $b \in \mathbb{R}^p$ solving the maximal margin problem

$$\max_{b\neq 0} \min_{1\leq i\leq n} \frac{Y_i(Xb)_i}{\|b\|_1} =: \hat{\gamma},$$

where $\|\cdot\|_1$ denotes the ℓ_1-norm. As is shown in for example the papers [3], [5], and [4], the max-margin classifier is closely related to the ada-boost algorithm developed by [2]. The max-margin estimator is proportional to the minimum ℓ_1-norm estimator

$$\hat{\beta} := \arg\min\left\{ \|b\|_1 : \min_{1\le i\le n} Y_i(Xb)_i \ge 1 \right\}.$$

Note that the estimator interpolates the data in the sense that $\text{sign}(X\hat{\beta}) = Y$.

Theorem 1.1. *Suppose that the n rows of X are i.i.d. copies of a standard Gaussian random row vector $\mathbf{x} \in \mathbb{R}^{1\times p}$ and that ξ is independent of X with i.i.d Gaussian entries with mean zero and variance σ^2. Let $0 < \delta < 1$ be arbitrary. There exists a constants $\{c_1, c_2, c_3, c_4, c_5, c_6\}$ such that for $n \le p^\delta / c_1$ and $\log p \le n/c_2$, with probability at least $1 - n^{-1/c_3}$*

$$\hat{\gamma} \ge \frac{1}{c_4}\left(\frac{\log p}{n} \frac{1}{\|\beta^*\|_1 + \sigma\sqrt{n/\log p}} \right)^{\frac{1}{3}},$$

$$\frac{\|\hat{\beta}\|_1}{\|\hat{\beta}\|_2} \le c_5 \left(\|\beta^*\|_1 + \sigma\sqrt{n/\log p} \right) \log p,$$

and

$$\left\| \frac{\hat{\beta}}{\|\beta\|_2} - \beta^* \right\|_2 \le c_5 \left(\frac{\log p}{n} \|\beta^*\|_1^2 + \sigma^2 \right)^{\frac{1}{4}} \log^{\frac{1}{2}} p.$$

The theorem can be extended to the case of adversarial noise at the cost of an additional log-factor.

The first result in Theorem 1.1 for the margin is derived using bounds obtained in [1] and is in fact optimal. The other two results rely on the first result but may be sub-optimal. Note that a rate of convergence for the misclassification error follows immediately from the ℓ_2-rate of convergence by Grothendieck's identity, which says that for a standard Gaussian random vector $\mathbf{x} \in \mathbb{R}^{1\times p}$, and for all $b \in \mathbb{R}^p$ with $\|b\|_2 = 1$, one has

$$\begin{aligned} \mathbb{P}(\text{sign}(\mathbf{x}\beta^*) \ne \text{sign}(\mathbf{x}b)) &= \frac{1}{\pi}\arccos(\beta^{*T} b) \\ &= \frac{1}{\pi} d_{\text{G}}(\beta^*, b), \end{aligned}$$

where $d_{\text{G}}(\beta^*, b)$ is the Geodesic distance between the vectors β^* and b. Thus, in the context of Theorem 1.1, the Bayes error is of order σ for σ small, whereas our bound has a term of order $\sqrt{\sigma}$ which dominates σ for σ small. On the other hand, if σ is small, the error due to the noise may be of smaller order than the error for the noiseless problem.

References

[1] G. Chinot, M. Löffler, M. and S. van de Geer, *On the robustness of minimum-norm interpolators*, arXiv:2012.00807 (2020).

[2] Y. Freund and R.M. Schapire, *A decision-theoretic generalization of on-line learning and an application to boosting*, **55** (1997), 119–139.

[3] S. Rosset, and J. Zhu and T. Hastie, *Boosting as a regularized path to a maximum margin classifier*, Journal of Machine Learning Research **5** (2004), 120–140.

[4] M. Telgarsky, *Margins, shrinkage, and boosting*, International Conference on Machine Learning (2013), 307–315.

[5] T. Zhang, and B. Yu, *Boosting with early stopping: convergence and consistency*, Annals of Statistics **33** (2004),1538–1579.

Estimating the lasso's effective noise

Michael Vogt
(joint work with Johannes Lederer)

Consider the high-dimensional linear model $Y = \boldsymbol{X}\beta^* + \varepsilon$ with response vector $Y \in \mathbb{R}^n$, design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, target vector $\beta^* \in \mathbb{R}^p$, and random noise $\varepsilon \in \mathbb{R}^n$. We allow for a dimension p that is of the same order or even much larger than the sample size n, and we assume a target vector β^* that is sparse. A popular estimator of β^* in this framework is the lasso [14]

$$(1) \qquad \hat{\beta}_\lambda \in \underset{\beta\in\mathbb{R}^p}{\arg\min} \Big\{ \frac{1}{n}\|Y - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_1 \Big\},$$

where $\lambda \in [0,\infty)$ is a tuning parameter. The lasso estimator satisfies the well-known prediction bound

$$(2) \qquad \lambda \geq \frac{2\|\boldsymbol{X}^\top\varepsilon\|_\infty}{n} \quad \Longrightarrow \quad \frac{1}{n}\|\boldsymbol{X}(\beta^* - \hat{\beta}_\lambda)\|_2^2 \leq 2\lambda\|\beta^*\|_1,$$

which is a direct consequence of the basic inequality for the lasso [2, Lemma 6.1] and Hölder's inequality. This simple bound highlights that a crucial quantity in the analysis of the lasso estimator is $2\|\boldsymbol{X}^\top\varepsilon\|_\infty/n$. We call this quantity henceforth the *effective noise*.

The effective noise does not only play a central role in the stated prediction bound but rather in almost all known finite-sample bounds for the lasso. Such bounds, called oracle inequalities, are generally of the form [2, 7, 9]

$$(3) \qquad \lambda \geq (1+\delta)\,\frac{2\|\boldsymbol{X}^\top\varepsilon\|_\infty}{n} \quad \Longrightarrow \quad \|\beta^* - \hat{\beta}_\lambda\| \leq \kappa\lambda$$

with some constant $\delta \in [0,\infty)$, a factor $\kappa = \kappa(\beta^*)$ that may depend on β^*, and a (pseudo-)norm $\|\cdot\|$. Oracle inequalities of the form Eq. (3) are closely related to tuning parameter calibration for the lasso: they suggest to control the loss $L(\beta^*, \hat{\beta}_\lambda) = \|\beta^* - \hat{\beta}_\lambda\|$ of the lasso estimator $\hat{\beta}_\lambda$ by taking the smallest tuning parameter λ for which the bound $\|\beta^* - \hat{\beta}_\lambda\| \leq \kappa\lambda$ holds with given probability $1-\alpha$. Denoting the $(1-\alpha)$-quantile of the effective noise $2\|\boldsymbol{X}^\top\varepsilon\|_\infty/n$ by λ^*_α, we immediately derive from the oracle inequality Eq. (3) that

$$(4) \qquad \mathbb{P}\Big(\|\beta^* - \hat{\beta}_{(1+\delta)\lambda}\| \leq \kappa(1+\delta)\lambda\Big) \geq 1-\alpha$$

for $\lambda \geq \lambda^*_\alpha$. Stated differently, $\lambda = (1+\delta)\lambda^*_\alpha$ is the smallest tuning parameter for which the oracle inequality Eq. (3) yields the finite-sample bound $\|\beta^* - \hat{\beta}_\lambda\| \leq \kappa\lambda$ with probability at least $1-\alpha$. Importantly, the tuning parameter choice $\lambda = (1+\delta)\lambda^*_\alpha$ is not feasible in practice, since the quantile λ^*_α of the effective noise is not observed. An immediate question is, therefore, whether the quantile λ^*_α can be estimated.

The effective noise is also closely related to high-dimensional inference. To give an example, we consider testing the null hypothesis $H_0 : \beta^* = 0$ against the alternative $H_1 : \beta^* \neq 0$. Testing this hypothesis corresponds to an important question in practice: do the regressors in the model $Y = \boldsymbol{X}\beta^* + \varepsilon$ have any effect on the response at all? A test statistic for the hypothesis H_0 is given by $T = 2\|\boldsymbol{X}^\top Y\|_\infty/n$. Under H_0, it holds that $T = 2\|\boldsymbol{X}^\top \varepsilon\|_\infty/n$, that is, T is the effective noise. A test based on the statistic T can thus be defined as follows: reject H_0 at the significance level α if $T > \lambda^*_\alpha$. Since the quantile λ^*_α is not observed, this test is not feasible in practice, which brings us back to the question of whether the quantile λ^*_α can be estimated.

We devise an estimator of the quantile λ^*_α of the effective noise based on bootstrap. Besides the level $\alpha \in (0, 1)$, it does not depend on any free parameters, which means that it is fully data-driven. The estimator can be used to approach a number of statistical problems in the context of the lasso. Here, we focus on two such problems: (i) tuning parameter calibration for the lasso and (ii) inference on the parameter vector β^*.

(i) Tuning parameter calibration for the lasso. Our estimator $\hat{\lambda}_\alpha$ of the quantile λ^*_α can be used to calibrate the lasso with essentially optimal finite-sample guarantees. Specifically, we derive finite-sample statements of the form

$$\mathbb{P}\Big(\|\beta^* - \hat{\beta}_{(1+\delta)\hat{\lambda}_\alpha}\| \leq \kappa(1+\delta)\lambda^*_{\alpha-\nu_n}\Big) \geq 1 - \alpha - \eta_n, \tag{5}$$

where $0 < \nu_n \leq Cn^{-K}$ and $0 < \eta_n \leq Cn^{-K}$ for some positive constants C and K. Statement Eq. (5) shows that calibrating the lasso with the estimator $\hat{\lambda}_\alpha$ yields almost the same finite-sample bound on the loss $L(\beta^*, \beta) = \|\beta^* - \beta\|$ as calibrating it with the oracle parameter λ^*_α. In particular, Eq. (5) is almost as sharp as the oracle bound $\mathbb{P}(\|\beta^* - \hat{\beta}_{(1+\delta)\lambda^*_\alpha}\| \leq \kappa(1+\delta)\lambda^*_\alpha) \geq 1 - \alpha$, which is obtained by plugging $\lambda = \lambda^*_\alpha$ into Eq. (4).

Finite-sample guarantees for the practical calibration of the lasso's tuning parameter are scarce. Exceptions include finite-sample bounds for Adaptive Validation [5] and Cross-Validation [4]. One advantage of our approach via the effective noise is that it yields finite-sample guarantees not only for a specific loss but for any loss for which an oracle inequality of the type Eq. (3) is available. Another advantage is that it does not depend on secondary tuning parameters that are difficult to choose in practice; the only parameter it depends on is the level $1-\alpha$, which plays a similar role as the significance level of a test and, therefore, can be chosen in the same vein in practice.

(ii) Inference on the parameter vector β^*. Our estimator $\hat{\lambda}_\alpha$ of the quantile λ^*_α can also be used to test hypotheses on the parameter vector β^* in the model $Y = \boldsymbol{X}\beta^* + \varepsilon$. Consider again the problem of testing $H_0 : \beta^* = 0$ against $H_1 : \beta^* \neq 0$. Our approach motivates the following test: reject H_0 at the significance level α if $T > \hat{\lambda}_\alpha$. We prove under mild regularity conditions that this test has

the correct level α under H_0 and is consistent against alternatives that are not too close to H_0. Moreover, we show that the test can be generalized readily to more complex hypotheses.

High-dimensional inference based on the lasso has turned out to be a very difficult problem. Some of the few advances that have been made in recent years include tests for the significance of small, fixed groups of parameters [1, 16, 6, 10, 8], tests for the significance of parameters entering the lasso path [12], rates for confidence balls for the entire parameter vector (and infeasibility thereof) [13, 3], and methods for inference after model selection [11, 15]. In stark contrast to most other methods for high-dimensional inference, our tests are completely free of tuning parameters and, therefore, dispense with any fine-tuning.

References

[1] A. Belloni, V. Chernozhukov & C. Hansen, *Inference on treatment effects after selection among high-dimensional controls*, The Review of Economic Studies **81** (2013), 608–650.

[2] P. Bühlmann & S. van de Geer, *Statistics for high-dimensional data: methods, theory and applications*, Springer (2011).

[3] T. Cai & Z. Guo, *Accuracy assessment for high-dimensional linear regression*, Annals of Statistics **46** (2018), 1807–1836.

[4] D. Chetverikov, Z. Liao & V. Chernozhukov, *On cross-validated Lasso*, arXiv preprint (2016).

[5] M. Chichignoud, J. Lederer & M. Wainwright, *A practical scheme and fast algorithm to tune the lasso with optimality guarantees*, Journal of Machine Learning Research **17** (2016), 1–20.

[6] S. van de Geer, P. Bühlmann, Y. Ritov & R. Dezeure, *On asymptotically optimal confidence regions and tests for high-dimensional models*, Annals of Statistics **42** (2014), 1166–1202.

[7] C. Giraud, *Introduction to high-dimensional statistics*, CRC Press (2014).

[8] D. Gold, J. Lederer and J. Tao, *Inference for high-dimensional instrumental variables regression*, Journal of Econometrics **217** (2020), 79–111.

[9] T. Hastie, R. Tibshirani & M. Wainwright, *Statistical learning with sparsity: the lasso and generalizations*, CRC Press (2015).

[10] A. Javanmard & A. Montanari, *Confidence intervals and hypothesis testing for high-dimensional regression*, Journal of Machine Learning Research **15** (2014), 2869–2909.

[11] J. Lee, D. Sun, Y. Sun, J. Taylor, *Exact post-selection inference, with application to the lasso*, Annals of Statistics **44** (2016), 907–927.

[12] R. Lockhart, J. Taylor, R. Tibshirani & R. Tibshirani, *A significance test for the lasso*, Annals of Statistics **42** (2014), 413–468.

[13] R. Nickl & S. van de Geer, *Confidence sets in sparse regression*, Annals of Statistics **41** (2013), 2852–2876.

[14] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society: Series B **58** (1996), 267–288.

[15] R. Tibshirani, J. Taylor, R. Lockhart & R. Tibshirani, *Exact post-selection inference for sequential regression procedures*, Journal of the American Statistical Association **111** (2016), 600–620.

[16] C.-H. Zhang & S. Zhang, *Confidence intervals for low dimensional parameters in high dimensional linear models*, Journal of the Royal Statistical Society: Series B **76** (2014), 217–242.

Reporter: Yann Issartel

Participants

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

Dr. Annika Betken
Department of Statistics
University of Twente
P.O. Box 217
7500 AE Enschede
NETHERLANDS

Prof. Gilles Blanchard
Laboratoire de Mathématiques
Université Paris-Saclay
F-91405 Orsay
FRANCE

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

Prof. Dr. Sébastien Bubeck
Microsoft Research
One Microsoft Way
Redmond, WA 98052-6399
UNITED STATES

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

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

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

Prof. Dr. Alexandra Carpentier
Fakultät für Mathematik
Otto-von-Guericke-Universität
Magdeburg
Postfach 4120
39016 Magdeburg
GERMANY

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. Steffen Dereich
Institut für Mathematische Statistik
Universität Münster
Einsteinstrasse 62
48149 Münster
GERMANY

Prof. Dr. David L. Donoho
Department of Statistics
Stanford University
Sequoia Hall
Stanford, CA 94305-4065
UNITED STATES

Prof. Dr. John Duchi
Department of Statistics and Electrical Engineering
Stanford University
Sequoia Hall
390 Jane Stanford Way
Stanford CA 94305-4065
UNITED STATES

Prof. Dr. Jianqing Fan
Department of Operations Research and Financial Engineering
Princeton University
Princeton NJ 08544
UNITED STATES

Gianluca Finocchio
Faculty EEMCS
University of Twente
P.O. Box 217
7500AE Enschede
NETHERLANDS

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

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

Dr. Suriya Gunasekar
Microsoft Research
One Microsoft Way
Redmond, WA 98052-6399
UNITED STATES

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

Prof. Dr. Matthias Hein
Universität Tübingen
Dept. of Computer Science
Maria-von-Linden-Strasse 6
72076 Tübingen
GERMANY

Prof. Dr. Daniel J. Hsu
Department of Computer Science
Data Science Institute
Columbia University
500 West 120 Street
P.O. Box MC0401
New York, NY 10027
UNITED STATES

Yann Issartel
École Nationale de la Statistique
e de l'Adm. Economique
ENSAE
3, avenue Pierre-Larousse
92245 Malakoff
FRANCE

Sebastian Kassing
Institut für Mathematische Statistik
Universität Münster
Einsteinstrasse 62
48149 Münster
GERMANY

Prof. Dr. Zheng T. Ke
Department of Statistics
Harvard University
Science Center
One Oxford Street
Cambridge MA 02138-2901
UNITED STATES

Prof. Dr. Yongdai Kim
Department of Statistics
Seoul National University
Seoul 151-747
KOREA, REPUBLIC OF

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

Prof. Dr. Michael Kohler
Fachbereich Mathematik - Stochastik
TU Darmstadt
Schlossgartenstr. 7
64289 Darmstadt
GERMANY

Solt Kovács
Seminar für Statistik
ETH Zürich (HG G 18)
Rämistrasse 101
8092 Zürich
SWITZERLAND

Prof. Dr. Gitta Kutyniok
Mathematisches Institut
Ludwig-Maximilians-Universität
München
Theresienstraße 39
80333 München
GERMANY

Dr. Sophie Langer
Fachbereich Mathematik
Technische Universität Darmstadt
Schloßgartenstrasse 7
64289 Darmstadt
GERMANY

Prof. Gabor Lugosi
Department of Economics
Pompeu Fabra University
Ramon Trias Fargas 25-27
08005 Barcelona, Catalonia
SPAIN

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

Joseph Theo Meyer
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 205
69120 Heidelberg
GERMANY

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

Dr. Nicole Mücke
Fachbereich Mathematik, Sekr.MA 8-5
Technische Universität Berlin
Straße des 17. Juni 136
10623 Berlin
GERMANY

Prof. Dr. Axel Munk
Institut für Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstrasse 7
37077 Göttingen
GERMANY

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

Prof. Vianney Perchet
Crest, ENSAE
Criteo AI Lab
5 Avenue Le Chatelier
91120 Palaiseau 91120
FRANCE

Prof. Dr. Massimiliano Pontil
Istituto Italiano di Tecnologia and
University College London
Via Morego, 30,
16163 Genova
ITALY

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

Dr. Stefan Richter
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 205
69120 Heidelberg
GERMANY

Prof. Dr. Veronika Rockova
Chicago Booth
369 Charles M. Harper Center
5807 South Woodlawn Avenue
Chicago, IL 60637
UNITED STATES

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

Prof. 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
P.O.Box 217
7500 AE Enschede
NETHERLANDS

Prof. Ohad Shamir
Department of Computer Science and
Applied Mathematics
Weizmann Institute of Science
Rehovot 7610001
ISRAEL

Thomas Staudt
Mathematisches Institut
Georg-August-Universität Göttingen
Bunsenstr. 3-5
37073 Göttingen
GERMANY

Dr. Lukas Steinberger
Institut für Statistik und Operations
Research
Fakultät für Wirtschaftswissenschaften
Universität Wien
Oskar-Morgenstern-Platz 1
1090 Wien
AUSTRIA

Prof. Dr. Nike Sun
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139-4307
UNITED STATES

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

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

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

Alexander Tsigler
Department of Statistics
University of California, Berkeley
367 Evans Hall
Berkeley CA 94720-3860
UNITED STATES

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

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

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

Prof. Dr. Michael Vogt
Institut für Statistik
Universität Ulm
Helmholtzstrasse 20
89081 Ulm
GERMANY

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

Petr Zamolodtchikov
Department of Applied Mathematics
University of Twente
P.O.Box 217
7500 AE Enschede
NETHERLANDS

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