Mathematisches Forschungsinstitut Oberwolfach

Report No. 41/2023

DOI: 10.4171/OWR/2023/41

Mini-Workshop: Interpolation and Over-parameterization in Statistics and Machine Learning

Organized by
Mikhail Belkin, San Diego
Alexandre Tsybakov, Paris
Fanny Yang, Zurich

17 September – 22 September 2023

Abstract. In recent years it has become clear that, contrary to traditional statistical beliefs, methods that interpolate (fit exactly) the noisy training data, can still be statistically optimal. In particular, this phenomenon of "benign overfitting" or "harmless interpolation" seems to be close to the practical regimes of modern deep learning systems, and, arguably, underlies many of their behaviors. This workshop brought together experts on the emerging theory of interpolation in statistical methods, its theoretical foundations and applications to machine learning and deep learning.

Mathematics Subject Classification (2020): 62-xx.

Introduction by the Organizers

This mini-workshop was attended by a group of researchers who work on several topics related to benign overfitting such as harmless interpolation for linear models and neural networks, implicit bias of first-order algorithms on shallow neural networks and transformers, and other topics. The participants presented novel optimization guarantees and statistical theory for interpolating solutions when training overparameterized models, as well as recent advancements in analyzing the expressivity and learnability of different problem classes by transformer architectures. As a result of the talks, several technical discussions ensued between researchers who had not collaborated before, for example, on extending tight benign overfitting results towards non-Gaussian distributions and proving implicit biases for different architectures and imbalanced data. We expect that many of these discussions will lead to future publications. In total, the workshop program

included 12 talks and a longer discussion session led by Misha Belkin. The lively discussion was centered around the key ingredient behind and the relevance of the existing theory for understanding deep learning. The discussion revealed different opinions, one being that finding the appropriate low-dimensional structure in language would be the key to understanding Large Language Models vs. other alternative concepts not based on linearity and low dimensionality. The debate led to follow-up interactions that are on-going after the workshop. A discussion group that arose from the debate is working on redefining the theory to certify the reliability of neural networks.

Mini-Workshop: Interpolation and Over-parameterization in Statistics and Machine Learning

Table of Contents

Abstracts

A geometrical viewpoint on the benign overfitting phenomenon

Guillaume Lecué
(joint work with Zong Shang)

In the linear regression model, the minimum ℓ_2-norm interpolant estimator has received much attention since it was proved to be consistent even though it fits noisy data perfectly under some condition on the covariance matrix of the input vector and the signal. This phenomenon is now known as *benign overfitting*. Even matching upper and lower bounds for the estimator error of this estimator have been obtained, hence leading to necessary and sufficient conditions for benign overfitting for this estimator.

Motivated by this phenomenon, the study of the generalization property of minimum norm interpolant estimators for other norms have been obtained. They are however limited by some restrictive assumptions on the design and the signal such as the Gaussian assumption on the design or the signal being 1-sparse. Theirs proofs are based on the Convex Gaussian Minmmax Theorem that seems to be difficult to extend beyond the Gaussian case and that lack of some geometrical understanding. There is therefore again a lot to do for the understanding and identification of necessary and sufficient conditions for benign overfitting of minimum norm interpolant estimator for any norm.

In this talk, two geometrical tools are introduced that should be useful to solve this open problem: the Dvoretsky-Milman theorem and isomorphic and restricted isomorphic properties. It is possible to use these tools to get matching upper and lower bounds for the minimum ℓ_2-norm interpolant estimator and it looks like they should play a role for other norms than the ℓ_2 one.

We provide a first analysis of the minimum norm interpolant estimator for some general norm based on these two tools. However the result is not sharp enough to match any lower bound since it does not prove the consistency of the minimum norm interpolant estimator. The main two reasons why this approach is not optimal are: 1) we need to use the independence between the noise and the design; 2) we need to base our analysis on a splitting of the features space into an overfitting part and an estimation part.

Planted regression forests

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

In this talk we discuss a novel interpretable tree based algorithm for prediction in a regression setting, see [1]. Our motivation is to estimate the unknown regression function from a functional decomposition perspective in which the functional components correspond to lower order interaction terms. The idea is to modify

the random forest algorithm by keeping certain leaves after they are split instead of deleting them. This leads to non-binary trees which we refer to as planted trees. An extension to a forest leads to our random planted forest algorithm. Additionally, the maximum number of covariates which can interact within a leaf can be bounded. If we set this interaction bound to one, the resulting estimator is a sum of one-dimensional functions. In the other extreme case, if we do not set a limit, the resulting estimator and corresponding model place no restrictions on the form of the regression function.

In [1] a theory for an idealized version of random planted forests is developed in cases where the interaction bound is low. In [1] we show that if it is smaller than three, the idealized version achieves asymptotically optimal convergence rates up to a logarithmic factor. In the talk we explain this result by looking at the case that the model contains no interaction terms and that thus the model is an additive model. We explain that then the trees of the random planted forests can be interpreted as a modification of the smooth backfitting estimator where the additive components are estimated by iterative updates given by integral transforms. The updates differ from smooth backfitting estimators by replacing two-dimensional smooth kernel density estimators by piecewise constant histogram type estimators. When the tree estimators are averaged to get the forest estimator the discrete nature of the histogram estimators is smoothed out resulting in a forest estimator that is comparable to smooth backfitting estimators in additive models. This fact can be used to explain the near optimal rates of random panted forests in additive models. It also motivates that related results hold in models with higher order interaction terms.

References

[1] Munir Hiabu, Enno Mammen, and Joseph T. Meyer. Random Planted Forest: a directly interpretable tree ensemble. arXiv preprint arXiv:2012.14563 (stat).

Surprising behaviors of sparse min-norm and max-margin interpolators

Konstantin Donhauser

(joint work with Fanny Yang, Guillaume Wang, Michael Aerni, Marco Milanta, Stefan Stojanovic and Nicolo Ruggeri)

Modern machine learning has uncovered an interesting observation: large over parameterized models can achieve good generalization performance despite interpolating noisy training data. In this talk, we study high-dimensional linear models and show how interpolators can achieve fast statistical rates when their structural bias is moderate. More concretely, while minimum-ℓ_2-norm interpolators cannot recover the signal in high dimensions, minimum-ℓ_1-norm interpolators with strong sparsity bias are much more sensitive to noise. In fact, we show that even though they are asymptotically consistent, minimum-ℓ_1-norm interpolators converge with a logarithmic rate much slower than the $O(1/n)$ rate of regularized estimators. In

contrast, minimum-ℓ_p-norm interpolators with $1 \leq p \leq 2$ can trade off these two competing trends to yield polynomial rates close to $O(1/n)$.

References

[1] Guillaume Wang, Konstantin Donhauser, and Fanny Yang. *Tight bounds for minimum ℓ_1-norm interpolation of noisy data.* AISTATS, volume 151, pages 10572-10602, 2022.

[2] Konstantin Donhauser, Nicolo Ruggeri, Stefan Stojanovic, and Fanny Yang. *Fast rates for noisy interpolation require rethinking the effect of inductive bias.* ICML, volume 162, pages 5397-5428, 2022.

[3] Michael Aerni, Marco Milanta, Konstantin Donhauser, and Fanny Yang. *Strong inductive biases provably prevent harmless interpolation.* ICLR, 2023.

[4] Stefan Stojanovic, Konstantin Donhauser, and Fanny Yang. *Tight bounds for maximum ℓ_1-margin classifiers.* arXiv preprint arXiv:2212.03783, 2022.

Interpolation Learning with Short Programs and Shallow Neural Networks

Nathan Srebro

Classical theory, conventional wisdom, and all textbooks, tell us to avoid reaching zero training error and overfitting the noise, and instead balance model fit and complexity. Yet, recent empirical and theoretical results suggest that in many cases overfitting is benign, and even interpolating the training data can lead to good generalization. Can we characterize and understand when overfitting is indeed benign, and when it is catastrophic as classic theory suggests? And can existing theoretical approaches be used to study and explain benign overfitting and the "double descent" curve? I will discuss interpolation learning in linear (and kernel) methods, deep learning, as well as using the universal "minimum description length" or "shortest program" learning rule.

Tempered and benign overfitting in neural networks and kernels

Ohad Shamir

(joint work with Guy Kornowski, Gilad Yehudai, Daniel Barzilai)

Overparameterized neural networks (NNs) are observed to generalize well even when trained to perfectly fit noisy data. This phenomenon motivated a large body of work on "benign overfitting", where interpolating predictors achieve near-optimal performance. Recently, it was conjectured and empirically observed that the behavior of NNs is often better described as "tempered overfitting", where the performance is non-optimal yet also non-trivial, and degrades as a function of the noise level. However, a theoretical justification of this claim for non-linear NNs has been lacking so far. In this talk, we provide several results that aim at bridging these complementing views. We study a simple classification setting with 2-layer ReLU NNs, and prove that under various assumptions, the type of overfitting transitions from tempered in the extreme case of one-dimensional data, to benign in high dimensions. Thus, we show that the input dimension has a crucial role

on the type of overfitting in this setting, which we also validate empirically for intermediate dimensions. In addition, we also discuss some upcoming results on benign and tempered overfitting in kernel regression learning, which is surprisingly not well-understood under realistic assumptions.

Representational strengths and limitations of transformers

Daniel Hsu

(joint work with Clayton Sanford and Matus Telgarsky)

Attention layers, as commonly used in transformers, form the backbone of modern deep learning, yet there is no mathematical description of their benefits and deficiencies as compared with other architectures. This talk presents positive and negative results on the representation power of attention layers, with a focus on relevant complexity parameters such as width, depth, and embedding dimension. The main results establish separations between attention layers and other traditional neural network architectures such as recurrent neural networks, as well as separations between different transformer architectures.

References

[1] Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Representational Strengths and Limitations of Transformers. arXiv preprint arXiv:2306.02896, 2023.

Transformers and Associative Memories

Alberto Bietti

The goal of the work [1] is to provide a simple data model that illustrates how transformer language models can develop basic "in-context reasoning" capabilities during training. In particular, we consider a Markov (bigram) model of discrete sequences that uses both global bigrams/transitions $\pi_b(z'|z)$ as well as local/sequence-specific ones, $p(z'|z = q_k) = \mathbf{1}\{z' = o_k\}$ that override the global transitions on a few specific *trigger* tokens $z = q_k$ to output a given *output* token $z' = o_k$ that is always the same within a given sequence, but is randomly chosen in each different sequence.

Two-layer transformers trained on such a data model develop an "induction head" mechanism [2], whereby the first attention layer attends to previous tokens, while the second layer attends to previous occurrences of the output token. In order to understand how gradient dynamics lead to such a behavior, we view weight matrices as associative memories of the form

$$W = \sum_{(i,j)\in\mathcal{M}} \alpha_{ij} v_j u_i^\top \in \mathbb{R}^{d\times d},$$

where $(u_i)_i$ and $(v_j)_j$ are collections of nearly-orthonormal input and output embedding vectors (e.g., random vectors in high dimension), and $\mathcal{M}$ is a set of relevant pairwise associations. We show the following:

- the induction head mechanism can be implemented with a two-layer transformer with all weights at random initialization, except for three weight matrices (the key-query matrices at both layers, and output-value matrix at the second layer) that have specific associative-memory forms;
- empirically, training these three matrices with SGD recovers these associative memory behaviors
- theoretically, such outer-product associative memory behavior can be recovered with population gradient steps on each layer, in a top-down order.

In the follow-up work [3], we study statistical rates for such associative memories with finite samples and finite dimension, in the presence of heavy-tailed input data. We illustrate the role of the dimension d and of different optimization algorithms for improving the obtained rates.

References

[1] Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. *Birth of a Transformer: A Memory Viewpoint.* NeurIPS, 2023.

[2] N. Elhage, N. Nanda, C. Olsson, et al. *A Mathematical Framework for Transformer Circuits.* Transformer Circuits Thread, 2021.

[3] Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. *Scaling Laws for Associative Memories.* arXiv preprint arXiv:2310.02984, 2023.

Benign Calibration

Matus Telgarsky

The goal of this work was both to study benign overfitting for the logistic loss in as similar a way to the squared loss setting, and and secondly to verify and further investigate the correspondence between benign overfitting and empirical observations. In a bit more detail:

(1) The first task is to study the behavior of gradient descent on the logistic loss with data following standard benign overfitting settings (with appropriate modifications to the label distribution). By contrast with prior work studying benign overfitting in classification settings, the goal here is not to achieve good zero-one loss, but rather to achieve good population logistic risk, with some desire to be closer to an apples-to-applies comparison to the regression setting. Since minimizing the logistic loss for correctly specified models also implies minimization of calibration error (using logistic link), these results also imply good calibration, giving rise to the title.

(2) The second task is to revisit the experimental basis for benign overfitting and see if there are any different phenomena in the logistic loss case, and as much as possible seek out further phenomena, as the present work also considers gradient descent. The two main observations were that (a) there is an early "uniform convergence" phase where the training error is close to the test error, and moreover the latter is optimal, and (b) early stopping is necessary, since the solutions are off at infinity. Preliminary work further

empirically dissecting 2-layer ReLU networks and their correspondence to the benign overfitting setting are ongoing.

The work is ongoing and incomplete. When the data matrix exhibits a clear signal-to-noise ratio, a new set of margin maximization techniques were able to show that the max margin direction is found essentially instantly (whereas all existing analyses have a long burn-in phase), which clarified a few settings but not the "uniform convergence" phase in general. This latter phase is part of ongoing work, all of which will hopefully appear soon in a conference; the author of this section is grateful to the Oberwolfach participants, staff, and general environment for the many educational conversations and overall pleasant setting.

Classification versus regression with ℓ_2-minimizing solutions: A tale of two loss functions

Vidya Muthukumar

(joint work with Mikhail Belkin, Daniel Hsu, Adhyyan Narang, Anant Sahai, Vignesh Subramanian)

In this talk we compare the classification $0-1$ test error of the max-margin support-vector-machine (SVM) and the regression test mean-squared-error of the minimum-ℓ_2-norm interpolator (MNI) under identical models for the data covariance. We show the presence of high-dimensional regimes under which the SVM achieves consistency for classification tasks, but the MNI does not. These results are achieved by providing novel tight upper and lower bounds on the classification test error of the SVM through a two-step proof technique: a) showing a high-probability equivalence between the SVM and MNI under sufficiently (effectively) high-dimensional covariates [1], b) a sharp classification test error analysis of the MNI [2]. Notably, our consistency result for the SVM cannot be obtained through any known data-dependent generalization bound, even with zero label noise.

Formulation. We consider i.i.d. high-dimensional covariates $\{X_i \in \mathbb{R}^d\}_{i=1}^n$ where $d > n$ and binary labels $\{Y_i \in \{-1,+1\}\}_{i=1}^n$. Our focus is the max-margin linear support-vector-machine (SVM) classifier, which takes the form

$$\widehat{\theta}_{\mathsf{SVM}} := \arg\min \|\theta\|_2 \text{ subject to } Y_i\langle X_i, \theta\rangle \geq 1 \text{ for all } i \in [n]. \tag{1}$$

We make a mild assumption of full-rank on the training data matrix, which ensures not only that the separability constraints in (1) are feasible, but also that we can *interpolate* the training data, i.e. we can achieve $\langle X_i, \theta\rangle = Y_i$ for all $i \in [n]$.

First result: Equivalence to interpolation. We wish to sharply analyze the classification test error of the max-margin SVM (1) in high-dimensional settings. A core challenge in doing so is that, unlike minimum-norm-interpolation in linear regression, the SVM does not in general have a closed-form expression. We first show a key structural result: the SVM and the minimum-ℓ_2-norm interpolation of the binary labels $\{Y_i\}_{i=1}^n$, (i.e. $\widehat{\theta}_{\mathsf{MNI}} := \arg\min \|\theta\|_2$ subject to $\langle X_i, \theta\rangle = Y_i$ for all $i \in [n]$) *exactly coincide* with high probability in very high-dimensional settings.

More formally, we assume that the covariates $\{X_i\}_{i=1}^n$ are centered and comprised of independent, 1-sub-Gaussian entries; thus, the covariance matrix $\Sigma = \mathbb{E}[X_i X_i^\top]$ is a diagonal matrix with entries denoted by $\lambda \in \mathbb{R}^d$. Then, we are able to show in [1] that $\widehat{\theta}_{\mathsf{SVM}} = \widehat{\theta}_{\mathsf{MNI}}$ with probability tending to 1 as $n \to \infty$ as long as $d_\infty := \frac{\|\lambda\|_1}{\|\lambda\|_\infty} \gg n \log n$ and $d_2 := \frac{\|\lambda\|_1^2}{\|\lambda\|_2^2} \gg n$. Note that d_∞, d_2 are essentially *effective dimensions*; for isotropic covariance we have $d_\infty = d_2 = d$ and so the equivalence holds w.h.p. as long as $d \gg n \log n$.

We prove this result by analyzing the feasibility of the MNI's dual certificate in conjunction with high-dimensional vector and matrix concentration phenomena that arise when $d_2, d_\infty \gg n$. This result implies an interesting equivalence of training the squared loss or logistic/hinge loss in ultra-high-dimensional regimes of possible independent interest.

Second result: Classification-vs-regression. We next sharply analyzed the classification test error of the MNI $\widehat{\theta}_{\mathsf{MNI}}$, noting from the above result that all conclusions about test error directly carry over to $\widehat{\theta}_{\mathsf{SVM}}$ (whp). We assume that $Y_i = 1$ with probability $g(\langle X_i, \theta^* \rangle)$ where we assume $\theta^* = \hat{e}_t$ to be 1-sparse with $1 \le t \le k \ll n$, and $g(\cdot)$ to be any monotonic link function satisfying $\mathbb{E}[g(X_i)Y_i] \ge c > 0$ for some universal constant c. Notably, even for this very simple signal model the ultra-high-dimensional-regime of interest (i.e. $d_\infty \gg n \log n, d_2 \gg n$) can be shown to prohibit consistency in regression [3]. The core problem is *not* overfitting of noise but attenuation of the signal θ^*; in fact, $\widehat{\theta}_t \to 0$ as $n, d \to \infty$! This implies that the MNI would be inconsistent even on noiseless data; equivalently, even optimally tuned ridge regression would be inconsistent.

Despite signal attenuation we show that the classification task is much more benign (due to being evaluated by the 0-1 loss function) and can be shown to be consistent as $n, d \to \infty$. The main result, contained in [2], is a sharp 0-1 error analysis of $\widehat{\theta}_{\mathsf{MNI}}$ under the generative model assumed above. The analysis shows that classification-consistency is achieved *iff* the *ratio* of a certain contamination term (denoted by $\mathsf{CN} := \sqrt{\sum_{j \ne t} \lambda_j \widehat{\theta}_j^2}$) that measures the energy of components orthogonal to the signal that were recovered) and signal attenuation (denoted by $\mathsf{SU} := \frac{\widehat{\theta}_t}{\theta_t^*}$) tends to 0. Coneretely, we have 0-1 error $\asymp \frac{\mathsf{CN}}{\mathsf{SU}}$, which means that the problematic phenomenon of attenuation ($\mathsf{SU} \to 0$) can be compensated if $\mathsf{CN} \to 0$ at an even faster rate (which turns out to be a byproduct of the benign overfitting phenomenon on *pure noise*). An easily interpretable example is that of bilevel covariance where the first k entries of λ are equal to λ_H and the other $d - k$ entries of λ are equal to $\lambda_L < \lambda_H$. Here, a corollary of our result is that regression consistency holds *iff* $R := \lambda_H / \lambda_L \gg d/n$, while classification consistency holds *iff* $R \gg \sqrt{d/n}$; since $d \gg n$, the latter is clearly a much weaker condition. Our result can also show consistency separations for other covariance models,e.g. the case of polynomially decaying eigenvalues.

Discussion and open problems. We have generalized this pair of results to multiclass classification, kernel methods, and training loss functions beyond the

exponentially-tailed family. Several open problems that I am intrigued by are summarized below:

- An ambitious open problem is showing *the implication of training losses for shallow neural networks*. Some differences may manifest, notably, the presence or absence of neural-tangent-kernel behavior; however, an eventual equivalence to some type of exact interpolation is plausible.
- I am interested in investigating whether *separations between classification and regression tasks exist for nonparametric interpolating methods*, as well as *beyond the case of* 1*-sparse signal.*
- Finally, the *implications of the mentioned signal attenuation for robustness* remain relatively unexplored. Our preprint [4] provides an initial investigation into the (lack of) adversarial robustness in this regime for special Fourier and polynomial feature maps.

References

[1] Daniel Hsu, Vidya Muthukumar and Ji Xu. On the proliferation of support vectors in high dimensions. Journal of Statistical Mechanics: Theory and Experiment 2022, no. 11 (2022): 114011.

[2] Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, Anant Sahai. Classification versus regression in overparameterized regimes: Does the loss function matter? The Journal of Machine Learning Research, 22(1), pp.10104-10172.

[3] Alexander Tsigler, and Peter L. Bartlett. Benign overfitting in ridge regression. J. Mach. Learn. Res. 24 (2023): 123-1.

[4] Adhyyan Narang, Vidya Muthukumar, and Anant Sahai. Classification and Adversarial examples in an Overparameterized Linear Model: A Signal Processing Perspective. arXiv preprint arXiv:2109.13215 (2021)

Implicit Geometries through the Imbalance Lens

Christos Thrampoulidis

(joint work with Ganesh Ramachandra Kini, Tina Behnia, Vala Vakilian)

The talk discusses the following questions: What are the unique structural properties of models learned by deep-net classifiers? Is there an implicit bias towards solutions of a certain geometry and how does this vary across architectures and data? Specifically, how does this implicit geometry change under label imbalances, and is it possible to use this information to design better loss functions for learning with imbalances?

We first discuss the Neural Collapse phenomenon, which refers to the remarkable structural properties characterizing the geometry of class embeddings and classifier weights, found by deep nets when trained beyond zero training error. We remark that this characterization only holds for balanced data; hence, we ask whether it can be made invariant to class imbalances.

We present an affirmative answer. Firstly, we overview a theoretical abstraction of deep-learning training that assumes unconstrained optimization of the last-layer

embeddings and classifiers. For this, we prove that cross-entropy loss with vanishing regularization finds embeddings and classifiers that: (i) always interpolate a simplex-encoded label matrix, and (ii) form a geometry, which we call Simplex-Encoded-Labels Interpolation (SELI) geometry, that is determined by the SVD factors of this same label matrix. Secondly, we present extensive experiments on real imbalanced datasets that confirm convergence to the SELI geometry and thus verify its invariance to the label distribution. We caution that convergence worsens with increasing imbalances and support this finding theoretically by showing that unlike the balanced case, when minorities are present, ridge-regularization plays a critical role in tweaking the geometry. This defines new questions and motivates further investigations into the impact of class imbalances on the rates at which first-order methods converge to their asymptotically preferred solutions.

We then delve into how this newfound structural insight into embeddings' geometry can be harnessed to engineer loss functions for enhanced performance when training with imbalanced data. We review various logit-adjusted parameterizations of cross-entropy (CE) loss, which have been proposed as alternatives to weighted CE for training large models on label-imbalanced data beyond zero training error. These parameterizations are driven by the theory of implicit bias, which has been successful for linear models in inducing bias favoring minority classes. Extending this theory to non-linear models, we characterize the implicit geometry of classifiers and embeddings that are learned by different CE parameterizations. Specifically, we derive closed-form formulas for the angles and norms of classifiers and embeddings as a function of the number of classes, the imbalance and the minority ratios, and the loss hyperparameters. Using these, we show that logit-adjusted parameterizations can be appropriately tuned to learn symmetric geometries irrespective of the imbalance ratio. We present experiments and an empirical study of convergence accuracy in deep-nets to verify our findings.

Some statistical insights into PINNs

Claire Boyer
(joint work with Gérard Biau, Nathan Doumèche)

Physics-informed neural networks (PINNs) combine the expressiveness of neural networks with the interpretability of physical modeling. Their good practical performance has been demonstrated both in the context of solving partial differential equations and in the context of hybrid modeling, which consists of combining an imperfect physical model with noisy observations. As in classical regression analysis, we are interested in estimating an unknown regression function $u^\star$ such that $Y = u^\star(X) + \varepsilon$, for some random noise ε that satisfies $\mathbb{E}(\varepsilon|X) = 0$. What makes the problem original is that the function $u^\star$ is assumed to satisfy (at least approximately) a collection of M PDE-type constraints of order at most K, denoted in a standard form by $\mathcal{F}_k(u^\star, x) \simeq 0$ for $x \in \Omega$ and $1 \leq k \leq M$. Moreover, there exists some subset $E \subseteq \partial\Omega$ and an boundary/initial condition function $h : E \to \mathbb{R}^{d_2}$ such

that, for all $x \in E$, $u^\star(x) \simeq h(x)$. These constraints model some a priori physical information about $u^\star$. However, this knowledge may be incomplete (e.g., the PDE system may be ill-posed and have no or multiple solutions) and/or imperfect (i.e., there is some modeling error, that is, $\mathcal{F}_k(u^\star, x) \neq 0$ and $u^\star|_E \neq h$). This again emphasizes that $u^\star$ is not necessarily a solution of the system of differential equations.

In order to estimate $u^\star$, we assume to have at hand three sets of data:

- (*i*) A collection of i.i.d. random variables $(X_1, Y_1), \dots, (X_n, Y_n)$ distributed as $(X, Y) \in \Omega \times \mathbb{R}^{d_2}$, the distribution of which is *unknown*;
- (*ii*) A collection of i.i.d. random variables $X_1^{(e)}, \dots, X_{n_e}^{(e)}$ distributed according to some *known* distribution μ_E on E;
- (*iii*) A sample of i.i.d. random variables $X_1^{(r)}, \dots, X_{n_r}^{(r)}$ *uniformly distributed* on Ω.

The function $u^\star$ is then estimated by minimizing the empirical risk function

$$R_{n,n_e,n_r}(u_\theta) = \frac{\lambda_d}{n} \sum_{i=1}^{n} \|u_\theta(X_i) - Y_i\|_2^2 + \frac{\lambda_e}{n_e} \sum_{j=1}^{n_e} \|u_\theta(X_j^{(e)}) - h(X_j^{(e)})\|_2^2 + \frac{1}{n_r} \sum_{k=1}^{M} \sum_{\ell=1}^{n_r} \mathcal{F}_k(u_\theta, X_\ell^{(r)})^2 \tag{1}$$

over the class $\mathrm{NN}_H(D)$ of neural networks with H hidden layers of constant width D.

We exhibit that the classical training of PINNs can suffer from systematic overfitting when dealing with polynomial PDE priors: we explicitly construct minimizing sequences of the empirical risk, for which the theoretical risk explodes. To overcome this issue, we suggest to resort to a ridge regularization (implemented in most standard DL libraries), theoretically shown to be sufficient to ensure risk-consistency of empirical risk minimizers.

Then, we discuss how risk-consistency is not enough to ensure a strong convergence of the PINN estimate towards $u^\star$ (in L^2 for instance). To this end, we propose to use an additive Sobolev regularization during training, which is fully compatible with the hybrid modeling paradigm. The resulting doubly-regularized PINN estimate is shown to enjoy a strong convergence property towards $u^\star$ for the class of linear PDEs.

In-context learning linear models with transformers

Peter Bartlett
(joint work with Ruiqi Zhang, Spencer Frei)

Attention-based neural networks such as transformers have demonstrated a remarkable ability to exhibit in-context learning (ICL): Given a short prompt sequence of tokens from an unseen task, they can formulate relevant per-token and next-token predictions without any parameter updates. By embedding a sequence

of labeled training data and unlabeled test data as a prompt, this allows for transformers to behave like supervised learning algorithms. Indeed, recent work has shown that when training transformer architectures over random instances of linear regression problems, these models' predictions mimic those of ordinary least squares. Towards understanding the mechanisms underlying this phenomenon, we investigate the dynamics of ICL in transformers with a single linear self-attention layer trained by gradient flow on linear regression tasks. We show that despite non-convexity, gradient flow with a suitable random initialization finds a global minimum of the objective function. At this global minimum, when given a test prompt of labeled examples from a new prediction task, the transformer achieves prediction error competitive with the best linear predictor over the test prompt distribution. We additionally characterize the robustness of the trained transformer to a variety of distribution shifts and show that although a number of shifts are tolerated, shifts in the covariate distribution of the prompts are not. Motivated by this, we consider a generalized ICL setting where the covariate distributions can vary across prompts. We show that although gradient flow succeeds at finding a global minimum in this setting, the trained transformer is still brittle under mild covariate shifts. We complement this finding with experiments on large, nonlinear transformer architectures, which we show are more robust under covariate shifts.

Reporter: Konstantin Donhauser

Participants

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

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

Dr. Alberto Bietti
Flatiron Institute, Simons Foundation
629 Grand St apt 4A
Brooklyn, NY 11211
UNITED STATES

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

Konstantin Donhauser
Department of Computer Science
ETH Zürich
Universitätsstrasse 6
8092 Zürich
SWITZERLAND

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

Dr. Guillaume Lecué
ESSEC, Business School
3 Av. Bernard Hirsch
95000 Cergy-Pontoise Cedex
FRANCE

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

Dr. Vidya Muthukumar
Coda S1139
Georgia Tech. University
756 W Peachtree St. NW
Atlanta, GA 30332-0430
UNITED STATES

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

Prof. Dr. Nathan Srebro
Toyota Technological Institute
at Chicago
6045 S Kenwood Ave
Chicago, IL 60637
UNITED STATES

Matus Telgarsky
New York University
Courant Institute of Math. Sciences
251 Mercer Street
New York, NY 10012
UNITED STATES

Dr. Christos Thrampoulidis
University of British Columbia
Department of Electrical and Computer Engineering
5500-2332 Main Mall
Vancouver BC V6T 1Z4
CANADA

Prof. Dr. Alexandre B. Tsybakov
CREST - ENSAE, Institut Polytechnique de Paris
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

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