Mathematisches Forschungsinstitut Oberwolfach

Report No. 46/2022

DOI: 10.4171/OWR/2022/46

# Mini-Workshop: Mathematical Foundations of Robust and Generalizable Learning

Organized by
Johannes Lederer, Bochum
Po-Ling Loh, Cambridge
Yuting Wei, Philadelphia
Fanny Yang, Zürich

2 October – 8 October 2022

Abstract. Machine learning has become an highly active field of research, but its mathematical underpinnings are still hardly understood. This workshop identified key challenges, and it discussed potential solutions. Bringing together a diverse group of researchers, the workshop established different views on the topic based on notions from statistics, probability theory, and optimization.

*Mathematics Subject Classification (2020):* 62-XX.

## Introduction by the Organizers

Statistical learning and machine learning have achieved remarkable empirical success recently in science and engineering applications, including computer vision, neural language processing, game playing, robotics control, and even protein folding. Despite their enormous success in practice, these learning methods often differ significantly from classical statistical learning, and their generalizability and robustness are poorly understood. The disparity has inspired a recent flurry of theoretical research activity in the intersection of probability, statistics, and optimization, with the aim of exploring statistical interpretations of deep learning and beyond. This workshop identified and discussed two key challenges that permeate machine learning at the cutting edge, but that have been only lightly studied in previous mathematical literature: 1. Robustness, that is, providing a more complete mathematical characterization of the performance of various machine learning algorithms (and possibly newly devised ones) when training and/or test data are

contaminated by adversarial mechanisms. 2. Generalizability and transferability, that is, developing new statistical insights for the generalization performance of over-parameterized models by exploring the prolific interplay between model complexity, sample size, structural (implicit and explicit) biases/regularization, and different kinds of distribution shifts between the training and test data.

The workshop had 13 visiting participants from different countries, and four online participants. The workshop was a blend of researchers with various backgrounds yet a common interest in mathematical research. The heart of the workshop was the discussions of open problems; the following abstracts give an overview of these problems.

## Mini-Workshop: Mathematical Foundations of Robust and Generalizable Learning

### Table of Contents

# Abstracts

## Minimax hypothesis testing

Sivaraman Balakrishnan

The broad goal of the minimax hypothesis testing framework is to characterize the complexity of various hypothesis testing problems by placing them in a framework which parallels that of minimax statistical estimation. Perhaps, the most classical hypothesis testing problem is that of goodness-of-fit testing. Here, we are given samples $X_1, \ldots, X_n \sim P$ and would like to distinguish the hypotheses:

$$H_0 : P = P_0,$$
$$H_1 : P \in \mathcal{P}_1 \text{ and } \rho(P, P_0) \geq \epsilon_n,$$

where $\mathcal{P}_1$ is some family of structured distributions. The goal here is usually to characterize the *critical radius* $\epsilon_n$ – which is roughly the value $\epsilon_n$ for which there exists a non-trivial test to distinguish null from alternate. These are relatively well-studied problems (for common collections of distributions $\mathcal{P}_1$), and what is most surprising is that often the critical radius is much smaller than the minimax rate for estimation, i.e. we can often distinguish null from alternate using many fewer samples than we would need to estimate distributions in these collections.

I'm most interested in understanding better fuzzy versions of this hypothesis testing problem, i.e. suppose now we are interested in distinguishing:

$$H_0 : P \in \mathcal{P}_0 \text{ and } \rho(P, P_0) \leq \nu_n,$$
$$H_1 : P \in \mathcal{P}_1 \text{ and } \rho(P, P_0) \geq \epsilon_n,$$

where $\mathcal{P}_0$ is also a family of structured distributions (potentially, identical to $\mathcal{P}_1$). Now our goal is to characterize sequences $(n, \nu_n, \epsilon_n)$ for which there is a non-trivial test to distinguish these (and of course, to identify such tests).

Here are some more concrete questions that I am interested in (together with some brief motivation):

(1) Suppose we pick $\mathcal{P}_0$ and $\mathcal{P}_1$ to be the family of identity covariance Gaussians. In this case, the problem is relatively well-studied when the distance $\rho$ corresponds to the KL-divergence (i.e. the metric of interest simply measures the $\ell_2$ distance between means).

I believe that the case where $\rho$ instead corresponds to the $\ell_p$ distance (for $p \neq 2$) between means is already unresolved.

One reason why I think this problem is interesting is that tests which are optimal when $\nu_n = 0$ (for instance, the $\chi^2$ test) are provably sub-optimal when $\nu_n > 0$ (which to me is both surprising and interesting).

Another reason why I find this problem fascinating is that lower bound constructions are often much more subtle. In the classical case (when $\nu_n = 0$) lower bounds are obtained from lower bounds for distinguishing the simple null from a mixture under the alternate. In the case when $\nu_n > 0$, we need to construct mixtures under both the null and alternate

which match a large number of moments (and these are inherently much more challenging to design and analyze).

(2) Similarly, the case when $\mathcal{P}_0$ and $\mathcal{P}_1$ are classes of smooth densities (say Hölder with parameter $\beta$) is to my knowledge unresolved.

(3) Another broad motivation for studying problems of this form comes from using fuzzy testing as a bridge to understanding estimation problems.

There are two directions that I think would be interesting to pursue. First, it seems natural that as $\nu_n$ gets large, the rate at which the critical radius will need to scale will begin to look like the functional estimation rate for estimating the functional $\rho(P, P_0)$. It would be interesting to try to characterize (in some generality) how rates transition from testing rates to functional estimation rates (often the latter are much slower than the former).

The other direction which I think might be fruitful to explore is to try to study problems in which the local-minimax rate for functional estimation exhibits some interesting heterogeneity. In more detail, we understand for some problems (testing in multinomial and density models with $\nu_n = 0$) the critical radius exhibits a strong dependence on $P_0$, and this dependence reveals some aspects of the local geometry of the testing problem. On the other hand, it is completely unknown if such effects arise in functional estimation (i.e. does the rate for estimating $\rho(P, P_0)$ depend strongly on $P_0$ or not?). Studying fuzzy testing problems may lead to some answers to this question as well.

## Breaking the sample complexity barrier in reinforcement learning

Yuxin Chen

A central objective of reinforcement learning (RL) is to search for a policy—based on a collection of noisy data samples—that approximately maximizes cumulative rewards, without direct access to a precise description of the underlying environment. Emerging RL applications necessitate the design of sample-efficient solutions in order to accommodate the explosive growth of problem dimensionality. Given that the state space, the action space and the time horizon could all be unprecedentedly enormous, it is often infeasible to request a sample size exceeding the fundamental limit set forth by the ambient dimension in the tabular setting (which enumerates all combinations of state-action pairs). Consequently, how to make the best of use of data samples becomes one of the most pressing issues in contemporary RL, particularly in *sample-starved* applications where data collection is expensive, time-consuming, or even high-stake (e.g., online advertisements, autonomous systems).

Despite a large body of research dedicated to understanding the sample efficiency of RL, however, most of existing results suffered from an enormous sample complexity barrier that prevents one from obtaining a complete trade-off curve between sample complexity and statistical accuracy. For instance, even with an

idealized simulator, the state-of-the-art theory on prior RL methods requires the total sample size to exceed a huge (but unnecessary) threshold, thus restricting their practicality for sample-limited applications. A similar or even higher barrier emerged in prior theory for both online exploratory RL and offline RL. In stark contrast, however, no information-theoretic lower bounds developed thus far preclude us from attaining reasonable learning accuracy when going below the above sample complexity barrier. Such issues are already present in the simplest tabular settings, not to mention more complex scenarios that involve complicated function approximation (e.g., neural networks). Addressing these issues requires substantial expansion of the statistical and algorithmic foundation of RL.

Fortunately, the recent advancement in high-dimensional statistics provides a powerful and versatile toolbox to help accomplish the above goal. In this talk, I use two recent examples to illustrate the utility of high-dimensional statistics in settling the sample complexity of reinforcement learning.

**Breaking the sample size barrier in the presence of a simulator.** Assuming access to an idealized simulator of the unknown environment, our recent work [1] makes progress towards establishing a comprehensive understanding of the fundamental statistical limit as well as how to achieve it efficiently. Consider an infinite-horizon $\gamma$-discounted Markov decision process (MDP) with state space $\mathcal{S}$ and action space $\mathcal{A}$, and one can acquires samples for any state-action pair by querying the simulator. The aim is to achieve the desired sample efficiency with minimal calls to the simulator. However, all prior algorithms (both model-based and model-free) incurred a huge burn-in cost, and hence are not guaranteed to be efficient in sample-starved scenarios. Motivated by the inadequacy of prior theory, we take an important step towards closing the gap between achievability results and information-theoretic lower bounds, where the model-based (a.k.a. plug-in) approach turns out to be unreasonably effective. We propose a randomly perturbed model-based algorithm, and demonstrate its statistical optimality for the full range of accuracy level. To the best of our knowledge, this provides the *first* guarantee in a simulator setting that is optimal for the entire range of sample sizes (beyond which finding a meaningful policy is information theoretically impossible).

**Settling the sample complexity in offline/batch RL.** The next story is concerned with offline or batch RL, which learns using pre-collected data without further exploration. Effective offline RL would be able to accommodate distribution shift and limited data coverage. However, prior algorithms or analyses either suffer from sub-optimal sample complexities or incur high burn-in cost to reach sample optimality, thus posing an impediment to efficient offline RL in sample-starved applications. Our recent work [2] demonstrates that the model-based (or "plug-in") approach achieves minimax-optimal sample complexity without any burn-in cost for tabular Markov decision processes (MDPs). Concretely, consider a $\gamma$-discounted infinite-horizon MDP with $S$ states and effective horizon $\frac{1}{1-\gamma}$, and suppose the distribution shift of data is reflected by some single-policy clipped concentrability coefficient $C^\star$. We prove that the sample complexity scales as $\frac{SC^\star}{(1-\gamma)^3\varepsilon^2}$ in order to yield $\varepsilon$ accuracy, where $\varepsilon$ can take any value between 0 and

$\frac{1}{1-\gamma}$ (thus achieving minimax optimality for the full $\varepsilon$-range). Our algorithms are "pessimistic" variants of value iteration with Bernstein-style penalties, and do not require sophisticated variance reduction. Finally, our analysis is established upon a powerful "leave-one-out" decoupling argument that finds its roots in probability and random matrix theory.

The above two examples only reflect a tip of an iceberg. There are many other scenarios (e.g., online RL, multi-agent RL, partially observed RL) whose sample complexity has yet to be determined. I would like to invite experts from high-dimensional statistics to contribute to the very rich set of problems in reinforcement learning.

## References

[1] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Neural Information Processing Systems (NeurIPS)*, 2020.

[2] Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei, *arXiv preprint arXiv:2204.05275*, 2022.

## Robustness of deep learning-based signal reconstruction: Partial results and open directions

Reinhard Heckel

Deep learning-based methods give state-of-the-art performance for imaging tasks such as denoising and reconstructing an image from few and noisy measurements. Those tasks are traditionally solved with hand-crafted methods such as sparse regularization. Neural networks give higher accuracy and faster reconstruction speed, and are therefore replacing classical sparsity-based methods in imaging applications.

However, there are concerns that those improvements in performance come at the price of robustness. We and others investigated the robustness of a variety of methods empirically, and found no evidence of neural networks being less sensitive to worst-case perturbations and distribution shifts than classical methods [1].

However, it is still an open problem to characterize the robustness of neural networks for signal reconstruction for worst-case and for distribution shifts. In this note, we describe open problems on the worst-case robustness and on characterizing the robustness with regard to distribution shifts.

**Worst-case robustness.** Consider an estimator $f$ that takes as input a measurement $\mathbf{y} \in \mathbb{R}^m$ of an signal (often an image) $\mathbf{x} \in \mathbb{R}^n$. The worst-case risk for such an estimator is defined as

$$R_\epsilon(f) = \mathbb{E}_{(\mathbf{x},\mathbf{y})}\left[\max_{\|\mathbf{e}\|_2 \le \epsilon} \|f(\mathbf{y}+\mathbf{e}) - \mathbf{x}\|_2^2\right],$$

where expectation is over a distribution of image and corresponding measurement $(\mathbf{x}, \mathbf{y})$,and worst-case robustness is measured with respect to a worst-case $\ell_2$ perturbation.

An important open problem is to characterize optimal worst-case estimators for different signal models and for different worst-case noise models. This will enable to quantify accuracy-robustness tradeoffs and give insights into the design of worst-case robust estimators, and how to learn them from data.

Here is what we know for a simple linear setup. Consider a denoising problem where the goal is to estimate a signal $\mathbf{x}$ from a noisy measurement $\mathbf{y} = \mathbf{x}+\mathbf{z}$. Here, the signal $\mathbf{x}$ is drawn uniformly from a $d$-dimensional subspace and $\mathbf{z}$ is additive Gaussian noise with variance $\sigma^2$. It can be shown that an optimal linear worst-case estimator for this setup has the form $f(\mathbf{y}) = \alpha(\sigma, \epsilon)\mathbf{U}\mathbf{U}^T\mathbf{y}$, where $\mathbf{U}\mathbf{U}^T$ is an orthogonal projection onto the signal subspace, and $\alpha(\sigma, \epsilon) \in (0, 1)$ is a shrinkage factor depending on the noise variance and on the worst-case perturbation, as measured by $\epsilon$. This result implies that for this simple linear model, learning via robust optimization is equal to learning via regularization with jittering, i.e., training with additive noise in the input. An interesting question is whether such relations between robust optimization and certain types of regularization also hold in more general setups.

**Distribution shifts.** Machine learning systems are often applied to data that is drawn from a different distribution than the training distribution. Recent work has shown that for a variety of classification and signal reconstruction problems, the out-of-distribution performance is strongly linearly correlated with the in-distribution performance [2]. If this relationship or more generally a monotonic one holds, it has the important consequence that an estimator that is better in distribution than another is also better on out-of-distribution data.

In recent work, we have show that under a simple co-variate shift model, for a large class of estimators based on regularized empirical risk minimization, a monotonic performance relationship holds between in- and out-of-distribution risk, denoted by $R_P$ and $R_Q$, holds [3]. Specifically, $R_P(f) = g(R_Q(f))$ holds for a class of estimators $f \in \mathcal{F}$, and for a fixed monotonic function $g$ that depends on the problem and the distributions $P$ and $Q$.

In contrast, most classical results on distribution shifts focus on deriving bounds that guarantee that the difference of in- and out-of-distribution risk is small, i.e., $|R_P(f) - R_Q(f) \leq \delta$, for some $\delta$ depending on the problem and the distance between the distributions.

An interesting research direction is to derive precise relations between the performance of in- and out-of-distribution performance for classes of distributions and estimators, in order to understand the out-of-distribution generalization performance of estimators better.

## References

[1] M. Z. Darestani, A. Chaudhari, R. Heckel, *Measuring Robustness in Deep Learning Based Compressive Sensing*, International Conference on Machine Learning, 2021.

[2] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, L. Schmidt, *Measuring Robustness to Natural Distribution Shifts in Image Classification*, Advances in Neural Information Processing Systems, 2020.

[3] D. LeJeune, J. Liu, R. Heckel, *Monotonic Risk Relationships under Distribution Shifts for Regularized Risk Minimization*, submitted, 2022.

## Multi-group learning

Daniel Hsu
(joint work with Christopher Tosh)

*Multi-group learning* [3] generalizes the standard framework for statistical learning by asking for a predictor $f\colon \mathcal{X} \to \mathcal{Y}$ with small *conditional risks*

$$R(f \mid g) := E[\ell(f(X), Y) \mid X \in g]$$

for all $g$ from a family of groups $G$. (Here, $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a loss function, like zero-one loss, squared error, etc., and $(X, Y)$ is a random example taking values in $\mathcal{X} \times \mathcal{Y}$.) Groups, i.e., subsets of $\mathcal{X}$, might represent different demographic subpopulations when the input space is composed of features about individuals, so it may be important to pay attention to the average loss per group, rather than just in aggregate over the whole population. We would like a single predictor that has "small" conditional risk for all groups simultaneously. How small the conditional risk is for a group $g$ may depend on how many training data $N_g$ that one has for a group $g$—if we have very little training data for a group, we will not expect to guarantee very small conditional risk for that group—and we also only consider the excess conditional risk relative to what is achievable by some known reference class of predictors $H$.

(If the groups in $G$ were disjoint, we could obtain a separate predictor $h_g \in H$ for each $g \in G$, and then combine them into a single predictor in an obvious way. But we do not want to assume that groups are disjoint.)

The question is whether there is a simple learning strategy for this multi-group generalization of the statistical learning setup (ideally as simple as *empirical risk minimization*) that produces a simple predictor $f$ such that, with high probability over the realization of training data (which is assumed to be an IID sample from the distribution of $(X, Y)$), for all groups $g \in G$,

$$R(f \mid g) \leq \inf_{h \in H} R(h \mid g) + B_{H,G,N_g},$$

where $B_{H,G,N_g}$ is an excess conditional risk bound that may be a function of $H$, $G$, $N_g$. The "best" method we know (from [4]) is based on using an online learning algorithm [1] for the sleeping experts problem [2], combined with online-to-batch post-processing. This gives an excess risk bound for group $g$ of the form $\sqrt{\log(|H||G|)/N_g}$. (The $\log|H|$ can be replaced by VC dimension of $H$, up to another logarithmic factor.) But it produces a complicated predictor: it is an ensemble of $n$ predictors (where $n$ is the total sample size), each of which is already potentially more complicated than any individual reference predictor from $H$.

We know of other methods (also from [4]) that produce simpler predictors—e.g., short decision lists over predictors from the reference class $H$—but they have suboptimal excess conditional risk bounds. Is there an algorithm that produces

simple predictors that are near-optimal? (For example, does a simpler online-to-batch conversion work, like just taking the last iterate? Or, is there a batch version of the online algorithm for sleeping experts?)

## References


[1] Blum, A. and Mansour, Y. From external to internal regret. *Journal of Machine Learning Research*, 8(6), 2007.

[2] Freund, Y., Schapire, R. E., Singer, Y., and Warmuth, M. K. Using and combining predictors that specialize. In *Proceedings of the twenty-ninth annual ACM Symposium on Theory of Computing*, pp. 334–343, 1997.

[3] Rothblum, G. and Yona, G. Multi-group agnostic PAC learnability. In *International Conference on Machine Learning*, pp. 9107–9115, 2021.

[4] Tosh, C. and Hsu, D. Simple and near-optimal algorithms for hidden stratification and multi-group learning. In *Thirty-Ninth International Conference on Machine Learning*, 2022.


## Threshold channels and potential applications

Varun Jog

Let $p$ and $q$ be supported on an alphabet of size $k$ and let $D$ be a small number. Suppose we are interested in channels with input size $k$ and output size $D$. Since $D$ is much smaller $k$, this can be thought as quantizing the original support. Let the output distributions be $\hat{p}$ and $\hat{q}$. We can ask the question in [1]: among all channels (among all quantizers), which is the best one in terms of maximizing the $f$-divergence between the output distributions? (Say KL or Hellinger distance for concreteness.)

This problem is relevant when one is trying to compress; i.e., reduce the support size of distributions while simultaneously preserve some key statistical properties. For example, when doing hypothesis testing, the probability of error or sample complexity are captured by the KL divergence or Hellinger distance, respectively. Maximizing these divergences identifies the optimal way to do hypothesis testing on a reduced support size. One might also be interested from a quantization point of view; see [2] for an information theoretically optimal approach to quantizing.

The data processing inequality from information theory gives that $D_f(p||q) \geq D_f(\hat{p}||\hat{q})$. In my talk, I discussed reverse data processing inequalities; i.e., how much $f$-divergence can be preserved by using the best possible channel. I presented two results in this direction that I shall describe below.

In recent (as yet unpublished) work, we showed that the best channel lies in a family of "threshold channels". These channels are defined by $D-1$ thresholds, $0 < \gamma_1 < \cdots < \gamma_{D-1}$, and the threshold channel is a deterministic channel that maps $\{i : \frac{p_i}{q_i} \in [\gamma_{i-1}, \gamma_i)$ to symbol $i$. Here, we take $\gamma_0 = 0$ and $\gamma_D = +\infty$. that are defined by thresholding the likelihood ratios. This result is very general, in that it holds for all $f$-divergences, and it is even valid in a "post-processing" setting where the output of the best channel is cascaded with an arbitrary fixed channel.

Our second result was to show to analyze the performance of an almost optimal threshold channel and show that the $f$-divergence is preserved (up to logarithmic)

factors with even $D = 2$ as long as the $f$-divergence is symmetric and satisfies some regularity conditions.

The open problems presented in this talk derive from the results as well as the mathematical techniques we developed. One general question was, given the optimality of threshold channels in hypothesis testing and its properties with regards to post-processing, are there other problems in statistics where threshold channels are optimal? I suspect threshold channels have appeared in other contexts in statistics but they have not been recognized as such. For instance, the optimal test for robust hypothesis testing is known to be a clipped likelihood ratio test. This can be thought of as a threshold channel. The two main mathematical techniques we developed were a reverse Markov inequality and a quantized version of Jensen's inequality. Both inequalities appear to be quite fundamental. The reverse Markov inequality may have some application to anti-concentration inequalities that could be worth exploring. The quantized Jensen's inequality provided a novel approach to clustering, where arbitrary convex loss functions on $n$ points may be minimized to distinguish different clusters in the points.

## References

[1] A. Pensia, V. Jog, P-L. Loh, *Communication-constrained hypothesis testing: Optimality, robustness, and reverse data processing inequalities*, arXiv preprint arXiv:2206.02765 (2022).
[2] B. Nazer, O. Or, Y. Polyanskiy. *Information-distilling quantizers*,2017 IEEE International Symposium on Information Theory (ISIT). IEEE, 2017.

## Open problems in data segmentation algorithms

Claudia Kirch

Data segmentation methodology or multiple change point analysis has received considerable attention due to its importance in time series analysis and signal processing, with applications in a variety of fields including natural and social sciences, medicine, engineering and finance. This popularity can partly be explained by the fact that the assumption of piecewise stationarity underlying change point analysis, is one of the simplest forms of departure from stationarity while at the same time, it is found to be reasonable for many applications.

The field combines many interesting and challenging aspects in particular probabilistic aspects with questions of (mini-max) optimality but also computational challenges related to effective optimization techniques and questions of computational complexity.

Traditionally, change point analysis focused on the at-most-one-change situation, where the aim is to develop testing procedures for the null hypothesis of no change point versus the alternative of one change point. Such methodology has been and continues to be developed for all kinds of time series models way beyond univariate changes in the mean such as nonlinear (auto-)regressive models, integer-valued time series or robust statistical methodology (see e.g. the survey articles [2, 4, 5] or Section 4.1 in [3]). In particular, using method-of-moments

approaches allows to derive change point tests for any parametric model or robust change point tests in a very similar fashion as for changes in the mean, even though the mathematical analysis is naturally more involved. It is worth mentioning that said analysis can be conducted without the underlying model assumption to be correct and corresponding tests can be shown to hold their size and have sufficient power in such situations. In this case, the model can be thought of as a proxy to develop a change point test with good statistical properties as long as the best-approximating models differ before and after the change.

As one instance of more complex data, there has been a recent surge of interest in the development of change point methodologies for high-dimensional data (see e.g. Section 4.2 in [3]). One line of works deals with changes in functional data, i.e. data that can be modelled as a (discretisation of an underlying) curve. In this case, such a curve is thought of as an element of a Hilbert (or sometimes Banach) space as the basis of the corresponding mathematical analysis. As such, it is typically rather well behaved in contrast to a panel data setup, which is effectively a multivariate setup, where the number of components is of similar or even greater order than the number of time points - typically with no natural ordering between components. In this case, asymptotic considerations are usually carried out as a double asymptotic letting both the number of time points and the number of components grow to infinity. Most developed change point methodology for panel data works particularly well for sparse changes, i.e. changes that only occur in relatively few of the components (see e.g. Section 3.4 in [1]).

Based on these testing procedures one can typically also obtain an estimator with favorable statistical properties for the unknown time of the change – the change point. As soon as one moves away from the single change to a multiple change situation such estimators can often (depending on the weighting scheme used in testing) still be used for example in combination with binary segmentation procedures, however, the corresponding estimators are often less favorable.

This brings us to data segmentation procedures, which aim at segmenting the data into stretches that are (approximately) stationary translating into a multiple change point problem (the points between stationary stretches). In contrast to the above methodology, testing is no longer the main interest (although it can be used e.g. to give certain guarantees such as controlling the family-wise-error-rate). Indeed, the power behavior of associated tests (if existent at all) is typically less good than for the above tests – even in the presence of more than one change point, while localisation properties for the change point estimators are typically better. The largest body of literature focusing on multiple change point detection from a mathematical perspective deals with the detection and localization of multiple change points in the mean of univariate data: The *canonical segmentation problem*. Existing methodology can broadly be distinguished into (a) methodology based on optimizing a suitable information criterion and (b) methodology making use of change point tests. From a theoretical point of view, it is of particular interest to understand the separation as well as localization rate of a given procedure. Here, the separation rate contains the information how big the signal (magnitude of the

change, distance between change points) needs to be in order to be detectable by this procedure, while the localisation rates entails the information how close the corresponding estimator is to the true change point. In both cases, the information can be given in a multiscale (being more general) or merely in a homogeneous way, depending on whether the procedure allows different changes to have different behavior (in the sense of the relation between the magnitude of the change and the distance to the next neighboring change point). From a computational point of view, the computational complexity of the corresponding algorithm is of particular interest.

A thorough understanding on theoretical and computational performance of different data segmentation methods for the canonical segmentation problem forms the basis for the methodological development in more complex situations - similarly as tests for more complex situations are related to the mean change problem. Furthermore, the challenges arising from the high-dimensionality are orthogonal to those arising from the presence of multiple change points. Consequently, in the coming years, it will be of particular interest to combine different developments from these three areas of change point analysis, i.e. develop and analyse data segmentation algorithms for more complex possibly high-dimensional time-series data.

## References

[1] J. Aston, C. Kirch. *High dimensional efficiency with applications to change point tests*, Electron. J. Statist. 12 (1) 1901 - 1947, 2018.

[2] A. Aue, L. Horváth, *Structural breaks in time series.* Journal of Time Series Analysis, 34: 1-16, 2013.

[3] H. Cho, C. Kirch, *Data segmentation algorithms: Univariate mean change and beyond*, Econometrics and Statistics, 2022+. DOI: 10.1016/j.ecosta.2021.10.008

[4] L. Horváth, G. Rice, *Extensions of some classical methods in change point analysis*, TEST 23, 219–255, 2014.

[5] C. Kirch, J. Tadjuidje Kamgaing, *Detection of change points in discrete-valued time series.* In: Handbook of Discrete-Valued Time Series. Eds. R.A. Davis, S.A. Holand, R.B. Lund, N. Ravishanker, CRC Press, 2016.

## A universal estimation property of the Tukey median

Guillaume Lecué

During the last decade, several estimators of a mean vector or location parameters have been constructed and were proved to satisfy the very same statistical estimation bound as the empirical mean in the ideal i.i.d. Gaussian model even though the data given to these procedures were heavy-tailed and adversarially corrupted. There is a property of the empirical mean which has not receive a lot of attention and that we want to study: the empirical mean achieves the deviation-minimax optimal rate for the mean vector estimation problem with respect to any pseudo-norm. Our aim is to show that the Tukey median also enjoys this property on top of being robust to contamination (a property the empirical mean does not have).

Universal property of the empirical mean. Let us first recall the universal estimation property satisfied by the empirical mean in the Gaussian model.

Let $S$ be a symetric set of $\mathbb{R}^d$ and denote by $\|\cdot\|_S : x \in \mathbb{R}^d \to \sup_{v\in S}\langle v, x\rangle$ the associated pseudo-norm. Let $G_1, \ldots, G_N$ be $N$ i.i.d. $\mathcal{N}(\mu^*, \Sigma)$ Gaussian vectors in $\mathbb{R}^d$ (with mean $\mu$ and covariance matrix $\Sigma$) and denote by $\bar{G}_N = (1/N)\sum_{i=1}^N G_i$ their empirical mean. It follows from the Borell-TIS inequality (Theorem 7.1 in [2] or pages 56-57 in [3]) that for all $0 < \delta < 1$, with probability at least $1-\delta$,

$$\left\|\bar{G}_N - \mu^*\right\|_S = \sup_{v\in S}\left\langle v, \bar{G}_N - \mu^*\right\rangle \leq \mathbb{E}\sup_{v\in S}\left\langle v, \bar{G}_N - \mu^*\right\rangle + \sigma_S\sqrt{2\log(1/\delta)}$$

where $\sigma_S = \sup_{v\in S}\sqrt{\mathbb{E}\left\langle v, \bar{G}_N - \mu^*\right\rangle^2}$ is called the weak variance. It follows that with probability at least $1-\delta$,

$$\left\|\bar{G}_N - \mu^*\right\|_S \leq \frac{\ell^*(\Sigma^{1/2}S)}{\sqrt{N}} + \frac{\sup_{v\in S}\left\|\Sigma^{1/2}v\right\|_2\sqrt{\log(1/\delta)}}{\sqrt{N}} \tag{1}$$

where $\ell^*(\Sigma^{1/2}S) = \sup\left(\left\langle G, x\right\rangle : x \in \Sigma^{1/2}S\right) = \mathbb{E}\left\|\Sigma^{1/2}G\right\|_S$, for $G \sim \mathcal{N}(0, I_d)$, is the Gaussian mean width of the set $\Sigma^{1/2}S$. In particular, in the case where $S = B_2^d$, we recover in (1) the classical subgaussian rate for the mean estimation w.r.t. the $\ell_2^d$-norm

$$\sqrt{\frac{\mathbf{Tr}(\Sigma)}{N}} + \sqrt{\frac{\|\Sigma\|_{op}\log(1/\delta)}{N}}. \tag{2}$$

Moreover, the rate achieved by the empirical mean for the mean estimation problem w.r.t. $\|\cdot\|_S$ is deviation-minimax optimal as proved in the following result

**Theorem 1.** [1] *Let $S$ be a symmetric subset of $\mathbb{R}^d$ such that $\mathrm{span}(S) = \mathbb{R}^d$. If $\hat{\mu} : \mathbb{R}^{Nd} \to \mathbb{R}^d$ is an estimator such that for all $\mu^* \in \mathbb{R}^d$ and all $\delta \in (0, 1/4]$, $\mathbb{P}_{\mu^*}^N\left[\|\hat{\mu} - \mu^*\|_S \leq r^*\right] \geq 1-\delta$ where $\mathbb{P}_{\mu^*}^N$ is the probability distribution of a sample of $N$ i.i.d. Gaussian vectors $\mathcal{N}(\mu^*, \Sigma)$ then*

$$r^* \geq \max\left(\frac{1}{24}\sqrt{\frac{\log 2}{\log(5/4)}}\frac{\ell^*(\Sigma^{1/2}S)}{\sqrt{N}}, \frac{\sup_{v\in S}\left\|\Sigma^{1/2}v\right\|_2}{12}\sqrt{\frac{\log(1/\delta)}{\sqrt{N}}}\right).$$

It follows from the upper bound (1) and the deviation-minimax lower bound from Theorem 1 that the subgaussian rate for the problem of mean estimation in $\mathbb{R}^d$ w.r.t. $\|\cdot\|_S$ is given (up to absolute constant[1]) by

$$r_S^*(\delta) := \max\left(\frac{\ell^*(\Sigma^{1/2}S)}{\sqrt{N}}, \frac{\sup_{v\in S}\left\|\Sigma^{1/2}v\right\|_2\sqrt{\log(1/\delta)}}{\sqrt{N}}\right). \tag{3}$$

Moreover, the empirical mean is an estimator achieving this deviation-minimax optimal rate whatever the norm $\|\cdot\|_S$ is. This is a fundamental property of the empirical mean and one may ask several question related to this property:

[1]In this work, we do not consider the important problem of getting sharp optimal constants.

a) is there any other estimator enjoying this univsersal estimation property?
b) is this property satisfied by the empirical mean is due to the fact that the empirical mean is a sufficient statistics (and that it is somehow better than any other estimator)?
c) is it possible to extend this result beyond the Gaussian case? is it possible to extend this property even when some data are not all Gaussian and may have been corrupted even in the worse corruption model which is the $\epsilon$-adversarial corruption model? (and up to which proportion $\epsilon$ of corruption?)

It is the aim of this work to answer these questions. Our two main tools to solve these questions will be the Tukey depth and the Median-of-Means principle. Expected result for the Tukey median. We would first like to prove the following result.

**Theorem 2.** *We assume that a fraction $\epsilon$ of the $N$ Gaussian vectors $G_1, \dots, G_N$ has been corrupted by an adversary. We denote by $X_1, \dots, X_N$ the resulting adervsarially corrupted dataset given to the Tukey median $\hat{\mu}$ defined as*

$$\hat{\mu} \in \mathbf{argmin}_{\mu \in \mathbb{R}^d} D(\mu) \text{ where } D(\mu) = \sup_{v \in \mathbb{R}^d} \left( \frac{1}{N} \sum_{i=1}^{N} I\left( \langle X_i - \mu, v \rangle \geq 0 \right) \right).$$

*For all $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\|\hat{\mu} - \mu^*\|_S \leq 2 r_S^*(\delta) + \epsilon^2$$

*where $r_S^*(\delta)$ is the deviation-minimax rate of convergence for the $\|\cdot\|_S$ define in (3).*

Then we will consider some Median-of-means versions of the Tukey median to solve the heavy-tailed problem.

## References


[1] Jules Depersin and Guillaume Lecué. Optimal robust mean and location estimation via convex programs with respect to any pseudo-norms 2019.

[2] Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.

[3] Michel Talagrand. *Upper and lower bounds for stochastic processes*, volume 60 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer, Heidelberg, 2014. Modern methods and classical problems.

## Open problems: Covariate shift in non-parametric regression over RKHSs

Nicole Mücke

**Classical non-parametric regression over RKHSs.** The goal of non-parametric regression is to predict a real-valued response $Y$ based on a covariate $X$, being a random variable with values in a measurable space $\mathcal{X}$. It is assumed that the pair $(X, Y)$ comes from an unknown distribution $\mathbb{P}$ and $X$ has marginal $\mathbb{P}_X$. We refer to $\mathbb{P}_X$ as the *source distribution*. For each fixed $x \in \mathcal{X}$, the optimal estimator in a mean-squared sense is given by the regression function $f^*(x) := E[Y|X = x]$, that is,

$$f^* \in \operatorname*{arg\,min}_{f \in L^2(\mathcal{X}, \mathbb{P}_X)} \mathcal{R}(f) \,, \quad \mathcal{R}_{\mathbb{P}}(f) := \mathbb{E}\left[(Y - f(X))^2\right] \,,$$

where the expectation is taken with respect to the distribution $\mathbb{P}$.

To find an estimator $\hat{f}_D$ for $f^*$, we are given an i.i.d. training sample $D = ((X_1, Y_1), ..., (X_n, Y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, i.e. $D \sim \mathbb{P}^n$. the overall aim is to prove optimal minimax bounds for the excess risk

$$\mathcal{R}_{\mathbb{P}}(\hat{f}_D) - \mathcal{R}_{\mathbb{P}}(f^*) = ||\hat{f}_D - f^*||^2_{L^2(\mathbb{P}_X)} \,,$$

with high probability.

A common way to build estimators are by means of regularization and to restrict the search of an estimator to reproducing kernel Hilbert spaces (RKHS) $\mathcal{H}_K$, associated to some bounded kernel $K$. Classical methods arise from the large class of *spectral regularization methods* $\{g_\lambda\}_\lambda$, with $\lambda > 0$ being the regularization parameter. Note that these methods include, among others, (kernel) ridge regression (KRR), gradient descent, PCA, accelerated gradient methods like heavy ball and Nesterov acceleration. Minimax optimal bounds for these class of methods over RKHSs are well understood and have been established in [1], [2]. They crucially rely on structural assumptions, e.g. on an a-priori smoothness of the target $f^*$, expressed in terms of the kernel integral operator $T_{\mathbb{P}} : L^2(\mathcal{X}, \mathbb{P}_X) \to L^2(\mathcal{X}, \mathbb{P}_X)$ and on a proper decay of the eigenvalues of $T_{\mathbb{P}}$, expressed in terms of the effective dimension $tr[T_{\mathbb{P}}(T_{\mathbb{P}} + \lambda)^{-1}]$.

**Covariate shift.** A widely adopted assumption in supervised learning is that the training and test data are sampled from the same distribution. Such a no-distribution-shift assumption, however, is frequently violated in practice.

In the covariate shift version of the above problem, the *target distribution* $\mathbb{Q}_X$ is different from the source distribution $\mathbb{P}_X$. The main goal is to find an estimator $\hat{f}_D$ trained on $D \sim \mathbb{P}^n$ whose excess risk with respect to $\mathbb{Q}_X$ is small, i.e.

$$\mathcal{R}_{\mathbb{Q}}(\hat{f}_D) - \mathcal{R}_{\mathbb{Q}}(f^*) = ||\hat{f}_D - f^*||^2_{L^2(\mathbb{Q}_X)}$$

is small with high probability. To this end, it may be useful to assume that $\mathbb{P}_X$ is in some sense "close to" $\mathbb{Q}_X$.

**State of the art.** Research in RKHSs with covariate shift is rather scarce. Here, we mention the work [3] that analyses KRR with proper reweighting in terms of

the likelihood ratio $\rho(x) = \frac{q(x)}{p(x)}$, $x \in \mathcal{X}$ (existence assumed). Here, it is assumed that both, $\mathbb{P}_X, \mathbb{Q}_X$ are absolutely continuous w.r.t. the Lebesgue measure with densities $p, q$, respectively. Under the assumptions

(1) $\sup_{x\in X} |\rho(x)| \leq B$ , $B > 0$ or
(2) $\mathbb{E}_{\mathbb{P}_X}[\rho(x)^2] \leq \tau^2$, $\tau > 0$

the authors derive near optimal rates of convergence. In particular, it is assumed that $f^* \in \mathcal{H}_K$ and that $K$ has uniformly bounded eigenfunctions. Explicit rates for finite rank kernels and under polynomially decaying eigenvalues are presented.

**Open research.** There are many open and important problems that need to be addressed. Based on the works [1], [2] I aim to generalize weighted KRR to the broader class of spectral regularization kernel methods and to derive best possible error bounds for the excess risk w.r.t. the target distribution. On my way, refined bounds under refined a-priori smoothness assumptions (rather than $f^* \in \mathcal{H}_K$) are to be established. A major drawback of the above mentioned work is that the estimator that is given cannot be calculated in practice as it depends on the unknown likelihood ratio. A key step is to estimate a proper weighting based on additional unlabeled data w.r.t. $\mathbb{Q}_X$ that can simultaneously be used to derive early stopping rules for gradient based methods (or more generally adaptive regularization). A major step will be to bound an adapted notion of the effective dimension, namely $tr[T_{\mathbb{Q}}(T_{\mathbb{P}_X+\lambda})^{-1}]$. Here, it is necessary to control the perturbation of eigenvalues of $T_{\mathbb{Q}}$ for distributions $\mathbb{Q}_X$ in a vicinity of $\mathbb{P}_X$ for an appropriate metric.

## References

[1] G. Blanchard and Nicole Mücke, *Optimal rates for regularization of statistical inverse learning problems*, Foundations of Computational Mathematics **4** (2018), 971–1013.
[2] J. Lin, A. Rudi, L. Rosasco, and V. Cevher, *Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces*, Applied and Computational Harmonic Analysis **3** (2020), 868–90.
[3] C. Ma, R. Pathak, and M. Wainwright, *Optimally tackling covariate shift in RKHS-based nonparametric regression*, arXiv:2205.02986 (2022).

### Early stopping for iterative statistical learning

Markus Reiss

(joint work with Gilles Blanchard, Marc Hoffmann, Laura Hucker)

We discuss the power of early stopping for iterative learning methods as a regularisation method. For the prototypical example of a projection estimator on $m$-dimensional singular spaces of the singular value decomposition of the design or kernel matrix we analyse the bias-variance tradeoff along the subspace dimension $m$, which is sought to be chosen adaptively by a data-driven choice $\hat{m}$. This choice shall be based only on the first iterates (projections on smaller subspaces) and their residual norm. To do so, traditional model selection approaches like AIC, BIC, penalized least-squares or Lepski's method fail.

We prove an oracle inequality for prediction error loss, when the stopping rule is based on the residual norm / the in-sample training error in analogy with the discrepancy principle of numerical analysis. At noise level $\sigma$ and in dimension $D$ we face an additional $\sigma^2\sqrt{D}$-term in the oracle inequality which is due to estimating the (unknown) size $\|\epsilon\|^2$ of the i.i.d. noise variables $(\epsilon_i)_{1\le i\le D}$ by its expected value $\sigma^2 D$. A lower bound shows that this *payment for sequential adaptation* is problem-immanent. In particular, early stopping using a test or validation set cannot overcome this payment.

The extension to estimation (or reconstruction) error requires to rely on minimax results because the oracles in prediction and estimation error may differ too much. A two-stage procedure is proposed where after having stopped (and at least $m_0 = \sqrt{D}$ iterations have been performed) a model selection step on the so far calculated estimators is added. For the prototype problem the stopping rule and in particular the two-stage procedure show excellent finite-sample properties.

The results can be generalized to other linear spectral methods, in particular gradient descent (Landweber method). There are more technicalities involved and the constants in the bounds deteriorate slightly, but the general picture remains. The results presented can be found in [1] and [2].

For nonlinear methods like the popular choices of conjugate gradient or partial least squares (CG/PLS) methods the theory becomes more difficult in particular because no simple bias-variance tradeoff can be formulated. Nevertheless, basing results on an $\omega$-wise stochastic error, monotone upper bounds and a corresponding oracle, it is possible to establish similar results in prediction error for early stopping of the CG/PLS iterations. The highly non-trivial analysis of the estimation error is ongoing work in collaboration with Laura Hucker. A general literature survey on related early stopping results for kernel learning and $L^2$-boosting or matching pursuit is given.

## References

[1] G. Blanchard, M. Hoffmann, M. Reiß, Early stopping for statistical inverse problems via truncated SVD estimation, Electronic Journal of Statistics 12(2), 3204-3231, 2018.

[2] G. Blanchard, M. Hoffmann, M. Reiß, Optimal adaptation for early stopping in statistical inverse problems SIAM/ASA Journal of Uncertainty Quantification 6(3), 1043–1075, 2018.

## Mitigating multiple descents: A general framework for model-agnostic risk monotonization

Alessandro Rinaldo

(joint work with Pratik Patil, Arun Kumar Kuchibothla and Yuting Wei)

Modern machine learning models deploy a large number of parameters relative to the number of observations. Even though such overparameterized models typically have the capacity to (nearly) interpolate noisy training data, they often generalize well on unseen test data, seemingly defying the widely-accepted statistical wisdom

that interpolation will generally lead to over-fitting and poor generalization. See, e.g., the survey papers by [1] and [2] for related references.

A closely related and equally striking feature of overparameterized models is the so-called "double/multiple descent" behavior of the generalization error curve. The non-monotonic behavior of the generalization error suggests the jarring conclusion that, in high dimensions, increasing the sample size might actually yield a worse generalization error. In contrast, it is highly desirable to rely on prediction procedures that are guaranteed to deliver, at least asymptotically, a risk profile that is monotonically increasing in the aspect ratio. The ubiquity of the double and multiple descent phenomenon in over-parameterized settings begs the question:

*Is it possible to modify any given prediction procedure in order to achieve a monotonic risk behavior?*

We answer this question in the affirmative by demonstrating a simple, general-purpose framework that takes as input an arbitrary learning algorithm and returns a modified version whose out-of-sample risk will be asymptotically no larger than the smallest risk achievable beyond the aspect ratio for the problem at hand.

To illustrate the type of guarantees we obtained, we provide an informal version of one of our results. Adopting a standard regression framework, we assume that the data $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ are comprised of $n$ i.i.d. pairs of a $p$-dimensional covariate and a response variable from an unknown distribution. Using $\mathcal{D}_n$, suppose one fits a predictor $\hat{f}$ — a random function that maps $x \in \mathbb{R}^p \mapsto \hat{f}(x) \in \mathbb{R}$. Given a loss function $\ell\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$, we evaluate the performance of $\hat{f}$ by its conditional predictive risk given the data, defined by $R(\hat{f}; \mathcal{D}_n) = \mathbb{E}[\ell(Y_0, \hat{f}(X_0)) \mid \mathcal{D}_n]$, where $(X_0, Y_0)$ is an unseen data point, drawn independently from the data generating distribution.. We are interested in the limiting behavior of the risk under the proportional asymptotic regime in which $n, p \to \infty$ with the aspect ratio $p/n$ converging to a constant $\gamma \in (0, \infty)$. for a wide variety of problems and procedures. We devise a modification of the original procedure $\hat{f}$ that results into a new procedure $\hat{f}^{\text{zerostep}}$, called zero-step procedure whose asymptotic risk profile is provably monotonic in $\gamma$.

**Theorem 1** (Informal monotonization result)**.** *Suppose there exists a deterministic function $R^{\text{det}}(\cdot; \hat{f}) : (0, \infty] \to [0, \infty]$ such that for any $\phi \in (0, \infty]$ for any dataset $\mathcal{D}_n$ consisting of $m$ i.i.d. observations with $p_m$ features, $R(\hat{f}; \mathcal{D}_n) \to R^{\text{deter}}(\cdot; \hat{f})$, whenever $m, p_m \to \infty$ and $p_m/m \to \phi$. Then, under mild assumptions on $R^{\text{det}}$, the loss function $\ell$, and the data generating distribution, the procedure $\hat{f}^{\text{zerostep}}$ satisfies*

$$\left| R(\hat{f}^{\text{zerostep}}; \mathcal{D}_n) - \min_{\zeta \geq \gamma} R^{\text{det}}(\cdot; \hat{f}) \right| \xrightarrow{p} 0$$

*as $n, p \to \infty$ and $p/n \to \gamma \in (0, \infty)$.*

Figure 1 illustrates the above result for the minimum $\ell_2$-norm least squares estimator [3] and the minimum $\ell_1$-norm least squares estimator [4]. The light-blue

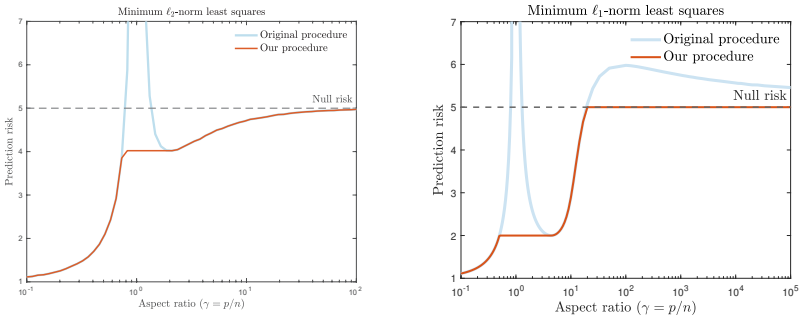


FIGURE 1. Monotonized asymptotic conditional prediction risk of the zero-step procedure for the minimum $\ell_2$-norm and $\ell_1$-norm least squares procedures. The figure in the left panel follows the setup of Figure 2 of [3], and the figure in the right panel follows the setup of Figure 3 of [4] (at sparsity level = 0.01). Both settings assume isotropic features and a linear model with noise variance $\sigma^2 = 1$ and linear coefficients of squared Euclidean norm $\rho^2 = 4$. Note that the risk is lower bounded by $\sigma^2 = 1$ and the risk of the null predictor (null risk) is $\rho^2 + \sigma^2 = 5$.

lines show the asymptotic risk profiles of the two procedures, which are non-monotonic as they diverge to infinity around the interpolation threshold of 1, at which the sample size and the number of features are equal. The red lines depict the risk profiles of the zero-step procedure $\hat{f}^{\text{zerostep}}$, which corresponds to the map

$$\gamma \in (0, \infty) \;\mapsto\; \min_{\zeta \geq \gamma} R^{\text{det}}(\zeta; \hat{f}).$$

The above function is a monotonically non-decreasing function of $\gamma$, regardless of whether $\gamma \mapsto R^{\text{det}}(\gamma; \hat{f})$ is non-monotonic. Furthermore, since

$$\min_{\zeta \geq \gamma} R^{\text{det}}(\zeta; \hat{f}) \leq R^{\text{det}}(\gamma; \hat{f}), \text{ for all } \gamma > 0,$$

the asymptotic risk of $\hat{f}^{\text{zerostep}}$ is no worse than that of $\hat{f}$.

The assumptions required in the theorem are very mild, and apply to a broad range of procedures and settings. The requirements on the loss functions are also mild and can be verified for common loss functions.

An interesting open problem is that of establishing some form of optimality for the monotonized risk.

## REFERENCES

[1] Bartlett, P., Montanari, A. and Rakhlin, A., *Deep learning: a statistical viewpoint,* arXiv preprint arXiv:2103.09177 (2021).

[2] Dar, Y., Muthukumar, V. and Baraniuk, R.G. (2021). A Farewell to the Bias-Variance Tradeoff? An Overview of the Theory of Overparameterized Machine Learning, arXiv preprint arXiv:2109.02355

[3] Hastie, T., Montanari, A., Rosset, S. and Tibshirani, R., *Surprises in high-dimensional ridgeless least squares interpolation,* arXiv preprint arXiv:1903.08560, (2019).
[4] Li, Y. and Wei, Y., *Minimum $\ell_1$-norm interpolators: Precise asymptotics and multiple descent,* arXiv preprint arXiv:2110.09502 (2021).

## Sharp adaptive similarity testing with pathwise stability for ergodic diffusions

Angelika Rohde
(joint work with Johannes Brutsche)

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

$$b_0(x) - \eta \leq b(x) \leq b_0(x) + \eta \quad \forall x \in I.$$

Our main results presented in the talk are the following:

(i) Based on a multiscale statistic and for any significance level $\alpha \in (0,1)$, we construct a threshold level such that the resulting test $\phi_T^\eta$ for the null hypothesis $H_0$ of similarity at tolerance $\eta$ satisfies

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

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

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

The cornerstone of the construction of $Y_\eta$ is the identification of the weak limit of the multiscale test statistic uniformly in $b \in H_0$.

(ii) We prove optimality and adaptivity for the similarity test in the minimax sense. We exemplarily consider the case of alternatives belonging to some Hölder class $\mathcal{H}(\beta, L)$, where deviations are measured in weighted supremum norm which is the equivalent to weighted risk definitions in sharp adaptive drift estimation. Our similarity test is shown to be rate-optimal in the minimax sense, adaptive in both the unknown parameters $\beta$ and $L$, optimal in the constant for the regime $\beta \leq 1$ and here, even sharp adaptive in $L$. The hypotheses construction in the proof of the lower bound involves a delicate fixed point problem as the drift itself appears in the invariant density which pops up in the deviation measure between null and alternative.

(iii) We address the problem of stability for fractional diffusion models where the driving Brownian motion is replaced by a fractional Brownian motion with Hurst index $H \in (0,1)$. Note that $H = 1/2$ corresponds to standard Brownian motion. We prove that the test statistic built from observations in the fractional diffusion model has strong performance properties as the fractional driving noise approaches Brownian motion in the following sense:

- The test is uniformly over the hypothesis of similarity of approximate level $\alpha$
- We prove that the minimax optimality is preserved in a certain sense as the fractional driving process approaches Brownian motion.

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

## Margin maximization with shallow ReLU networks

Matus Telgarsky
(joint work with Ziwei Ji)

This family of open problems is concerned with low-norm deep networks in classification settings where moreover the model is powerful enough to perfectly label the data.

**Background.** In more detail, this perfect labeling or *realizability* assumption in the classical linear regression setting is captured by the ordinary least squares solution

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 \\ \text{s.t.} \quad & Xw = y; \end{aligned}$$

in particular, both this solution (realized by either the pseudoinverse or by gradient descent) not only perfectly labels the data, but moreover has the lowest possible norm. The classification variant is similar, and corresponds to the SVM:

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 \\ \text{s.t.} \quad & y \odot (Xw) \geq 1, \end{aligned}$$

where the notation "$\odot$" refers to element-wise product. This family of open problems will focus on the second version, for classification, but superficial intuition carries over to both.

For deep networks, it is natural to consider a similar problem:

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 \\ \text{s.t.} \quad & y \odot (F(X;w)) \geq 1, \end{aligned} \tag{1}$$

where $F(x;w)$ denotes the output of a neural network architecture with input $x \in \mathbb{R}^d$ and parameters $w \in \mathbb{R}^p$, and $F(X;w)$ overloads this to a batch of data $X \in \mathbb{R}^{n\times d}$. The choice of norm is arbitrary so far, but still it is natural to consider the properties of the solution given by eq. (1). A related question is a smoothed and unconstrained analog given by

$$\min_{w\in\mathbb{R}^p} \frac{\ln \sum_i \exp\left(-y_i F(x_i;w)\right)}{\|w\|^2}. \tag{2}$$

A surprising fact, proved in parts by work of Kaifeng Lyu and Jian Li [1], and also Ziwei Ji and Matus Telgarsky [2], is that gradient descent on standard networks will lead to eq. (2) increasing monotonically and ultimately to a KKT point of eq. (1).

**Problems.** There are many immediate questions related to eqs. (1) and (2), and the behavior of gradient descent.

1. When does gradient descent *globally* maximize eq. (1)?

In certain extreme settings, under many assumptions, this is known to be true, however it is unknown if this has any reflection on practice. Indeed, it seems that practice is in general much more modest.

To build towards a much more modest open problem, then, consider the *Neural Tangent Kernel* or *near initialization* regime, where network weights are only allowed to move a constant distance from initialization as measured by Frobenius norm. With sufficient width, this setting still allows for universal approximation, however the predictors are significantly different than those reached later in training. This leads to the second question.

2. Under which circumstances does gradient descent reach a set of parameters which is in a strong sense better than that reached in the near-initialization regime, when measured with expressions similar to either eq. (1) or eq. (2)?

The final question is a purely technical curiosity relating to the choice of norm in eqs. (1) and (2), and variants which appeared in recent work.

3. Consider a two-layer network with weights $(a, V) \in \mathbb{R}^m \times \mathbb{R}^{m\times d}$. Under which circumstances can the squared norm $\frac{1}{2}\|w\|^2$ be replaced with the tighter quantity $\|a\|\cdot\|V\|$? Furthermore, is there an even tighter expression which can also be used in the same places (e.g., gradient descent and its resulting test error) and can honor the cancellation occurring within large weight matrices?

References

[1] Kaifeng Lyu and Jian Li, *Gradient descent maximizes the margin of homogeneous networks*, ICLR, 2019.

[2] Ziwei Ji and Matus Telgarsky, *Directional convergence and alignment in deep learning*, NeurIPS, 2020.

## Gradient flow, Laplace transforms, and infinitesimal steepest descent: Partial results and open directions

Ryan J. Tibshirani

Implicit regularization is currently a topic of key interest in the machine learning community, as many modern techniques for training neural networks do not use explicit regularization and yet still seem to have the ability to generalize in terms of out-of-sample predictive accuracy. Theory lags far behind current practice, and an important task for the statistics and machine learning community is to explain and understand the precise mechanism of implicit regularization.

One of the simplest and also one of the most widely-used techniques that falls into the general category of implicit regularization is *early stopping*. The foundations of this idea date back at least 30 years in machine learning, where early-stopped gradient descent was found to be effective in training neural networks, and at least 40 years in applied mathematics, where the same idea (here known as early-stopped Landweber iterations) was found to be effective in ill-posed linear inverse problems. Various authors have made connections between $\ell_2$ regularization and the iterates generated by gradient descent (when applied to different loss functions of interest), and several dozens of papers have been written on this topic (including several by attendees of the current workshop) starting in the mid 2000s through to the current day.

In [1], the author and collaborators adopted a continuous-time perspective and considered *gradient flow*, the path traced out by gradient descent iterates as the step size goes to zero. For a differentiable loss function $f$, this is characterized by the differential equation

$$\dot{\theta}(t) = -\nabla f(\theta(t)),$$

subject to the initial condition (say) $\theta(0) = 0$. When $f(\theta) = \frac{1}{2}\|Y - X\beta\|_2^2$, the least squares loss of a response vector $Y$ on a predictor matrix $X$, this differential equation is a linear dynamical system and has an analytical solution. We showed that, under to the calibration $t = 1/\lambda$ between the time in gradient flow and the tuning parameter in ridge regression, these estimates trace out risk curves that are within a universal multiplicative constant of each other (with very little assumptions on the data model). In other words, if one believes that ridge regression "does well" in a particular problem setting—either practically or theoretically—then early-stopped gradient descent should also "do well".

The current talk revisits this connection and develops an even more precise relationship between gradient flow and ridge regression. The following result can be easily verified using elementary arguments.

**Proposition 1.** *For any given response vector $Y$ and feature vector $X$, and loss function $f(\theta) = \frac{1}{2}\|Y - X\beta\|_2^2$, let $\hat{\theta}(\lambda)$ be the corresponding ridge regression solution with parameter $\lambda$, and $\bar{\theta}(t)$ denote the gradient flow solution at time $t$, initialized at $\theta(0) = 0$. Then for any $\lambda > 0$,*

$$\hat{\theta}(\lambda) = \lambda \cdot \mathcal{L}\{\bar{\theta}(t)\}(\lambda),$$

*where $\mathcal{L}\{g(t)\}(s)$ denotes the Laplace transform of a function $t \mapsto g(t)$ evaluated at the input parameter $s$. Equivalently, letting $T_\lambda \sim \mathrm{Exp}(1/\lambda)$, a random variable with an exponential distribution with mean $1/\lambda$, it holds that*

$$\hat{\theta}(\lambda) = \mathbb{E}[\bar{\theta}(T_\lambda)].$$

This relationship is interesting, because averaging the iterates from gradient descent—particularly from *stochastic* gradient descent—is common practice and frequently observed to perform well and provide variance stabilization. This proposition explains that in the simple but fundamental least squares setting, averaging along the gradient flow path with exponentially-decaying weights *exactly* reproduces ridge regression. It also shows where the calibration $t = 1/\lambda$ comes from—if instead of averaging along the gradient flow path, we were to choose just one point to best match ridge regression, then we may as well take the mean $1/\lambda$ of the $\mathrm{Exp}(1/\lambda)$ distribution governing the relationship.

Several open directions are suggested by this relationship. In particular:

(1) Does the Laplace transform reveal an analogous relationship for stochastic gradient descent, which can be described by the infinitesimal dynamics:

$$d\theta(t) = -\nabla f(\theta(t))dt + \Sigma(\theta(t))^{1/2} dW_t,$$

where $W_t$ is a standard Brownian diffusion process, and $\Sigma(\theta(t))$ a particular covariance matrix depending on $\theta(t)$?

(2) The same question, but now for generalized linear models?

Also of great interest is to study the precise relationship between infinitesimal *steepest descent* and explicit regularization. In [2], we studied the statistical performance of iterates from steepest descent on a loss $f$ with respect to a regularizer $g$. In light of the above, it is worth revisiting this from the continuous-time perspective. The infinitesimal dynamics here would be:

$$\dot{\theta}(t) \in \partial g^*\big(-\nabla f(\theta(t))\big),$$

a subdifferential inclusion.

## References

[1] Ali, A., Kolter, J. Z., and Tibshirani, R. J., *A continuous-time view of early stopping for least squares*, International Conference on Artificial Intelligence and Statistics (2019).

[2] Tibshirani, R. J., *A General Framework for Fast Stagewise Algortihms*, Journal of Machine Learning Research, **16** (2015), 2543–2588.

*Reporter: Konstantin Donhauser*

# Participants

**Prof. Dr. Sivaraman Balakrishnan**
Department of Statistics, Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213-3890
UNITED STATES

**Yuansi Chen**
Department of Statistical Science
Duke University
214 Old Chemistry Building
Durham NC 27708-0320
UNITED STATES

**Yuxin Chen**
Department of Statistics and Data Science
The Wharton School
University of Pennsylvania
265 South 37 Street
Philadelphia, PA 19104-6340
UNITED STATES

**Konstantin Donhauser**
Forschungsinstitut für Mathematik
ETH-Zürich
ETH Zentrum
Rämistr. 101
8092 Zürich
SWITZERLAND

**Pegah Golestaneh**
Fakultät für Mathematik
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum
GERMANY

**Dr. Reinhard Heckel**
Technische Universität München
Theresienstr. 90
80333 München
GERMANY

**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. Varun Jog**
Centre for Mathematical Sciences
University of Cambridge
Wilberforce Road
Cambridge CB3 0WB
UNITED KINGDOM

**Prof. Dr. Claudia Kirch**
Fakultät für Mathematik
Otto-von-Guericke-Universität Magdeburg
Postfach 4120
39016 Magdeburg
GERMANY

**Dr. Guillaume Lecué**
École Nationale de la Statistique et de l'Administration Économique
ENSAE
5, Avenue Henry le Châtelier
91120 Palaiseau Cedex
FRANCE

**Prof. Dr. Johannes Lederer**
Fakultät für Mathematik
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum
GERMANY

**Dr. Po-Ling Loh**
Centre for Mathematical Sciences
University of Cambridge
Wilberforce Road
Cambridge CB3 0WB
UNITED KINGDOM

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

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

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

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

**Matus Telgarsky**
University of Illinois, Urbana-Champaign
Urbana 61801
UNITED STATES

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

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

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