Fanghui Liu homepage: https://www.lfhsgre.org/ fanghui.liu@epfl.ch

EPFL

Research Statement

The role of over-parameterization in machine learning
- a function space theory view

Statistical machine learning (ML) has enjoyed immense practical success in recent years, especially deep learning in the era of big data. Rich mathematical theory explaining a flurry empirical results can help drive further advances but current theoretical understanding on ML models that can represent the data with the desired accuracy is still incomplete. For example, modern neural networks (NNs) work in an *over-parameterized* regime, where the number of parameters is much larger than the number of samples. They can perfectly fit the (noisy) training data but still generalize well, which goes against the conventional wisdom in classical learning theory.

My research aims to fill in the gap in the ML theory community from the perspective of *function space theory*, which focuses on which hypothesis space is suitable for learning at first, and then study generalization properties of machine learning models from supervised learning to reinforcement learning (RL).

The commonly used function space in learning theory is the reproducing kernel Hilbert space (RKHS), which provides the ability to approximate functions by nonparametric functional representations. The starting point of my research is based on the fact that RKHS is not large enough. For instance, to approximate a single ReLU neuron with an ε-approximation error, kernel methods in RKHS, e.g., neural tangent kernel (NTK) [JGH18] in deep learning theory, require a number of samples $\Omega(\varepsilon^{-d})$, exponential in feature dimensionality d [Bac17], *a.k.a.*, curse of dimensionality (CoD). Accordingly, my research aims to theoretically understand generalization guarantees of learning in a series of more and more general function spaces as below.

$$\begin{array}{ccccccc} \text{RKHS} & \xrightarrow{\text{bivariate form}} & \text{hyper-RKHS} & \xrightarrow{\text{variational form}} & \text{Barron space} & \xrightarrow{\text{smoothness}} & \text{Besov space} \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ \text{kernel methods} & \longrightarrow & \text{hyper-kernel methods} & \longrightarrow & \text{two-layer NNs} & \longrightarrow & \text{deep NNs} \end{array}$$

It covers several topics in my research *from kernel methods to neural networks*, centering around theoretical understanding generalization guarantees of machine learning models, especially *over-parameterized* models, provably and efficiently. My research from current achievements to future plan includes kernel learning, kernel approximation, double descent of over-parameterized models, and reinforcement learning theory in function approximation, see the unifying framework in Fig. 1.

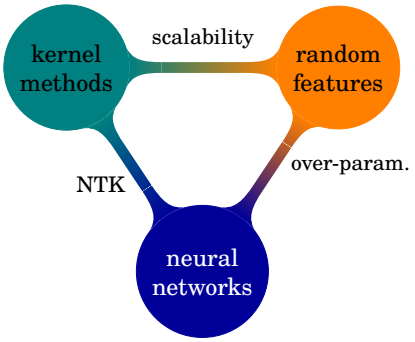

Figure 1: An overview from kernel methods to neural networks.

To be specific, 1) **kernel methods** in hyper-RKHS for kernel learning; 2) kernel approximation for **scalability** via random Fourier features (RFFs); 3) focus on generalization guarantees of **over-parameterized** models beyond classical learning theory, including two-layer NNs in Barron space, deep NNs via neural tangent kernel (NTK) in RKHS, as well as Besov space for Q-function approximation in deep reinforcement learning (RL). Besides, some theoretical-oriented topics with strong application background by student projects I co-supervised, e.g., robustness, neural architecture search can be also studied under our framework 1 in a systematical way.

My current research statement focuses on the following four fundamental questions from theory to application:

- **Q1:** What is generalization guarantees beyond RKHS?
- **Q2:** Why over-parameterized NNs generalize well under SGD training?
- **Q3:** Over-parameterization helps or hurts robustness in NNs?
- **Q4:** How does deep RL work well for function approximation beyond "linear" regime?

Current achievements: My research is able to understand what is the role of over-parameterization from kernel methods to neural networks on robustness, generalization, function approximation, centering around previous three questions. This leads to several scientific contributions at the flagship

journals and conferences in machine learning. Some of these research outputs have already presented on ICASSP 2023 tutorial entitled "*Neural networks: the good, the bad, the ugly*" and CVPR 2023 tutorial entitled "*Deep learning theory for vision researchers*.

RKHS: $f(\boldsymbol{x}) = \langle f, k(\boldsymbol{x}, \cdot)\rangle_{\mathcal{H}}, \forall f \in \mathcal{H}$ hyper-RKHS: $k(\boldsymbol{x}, \boldsymbol{x}') = \langle k, \underline{k}((\boldsymbol{x}, \boldsymbol{x}'), \cdot)\rangle_{\underline{\mathcal{H}}}, k \in \underline{\mathcal{H}}$

- **A1:** Our work [LSH$^+$21] generalizes regularized regression problems with bivariate forms endowed by hyper-RKHS, and provides generalization guarantees of regularized regression algorithms in hyper-RKHS. The convergence rate of the excess risk is proved to be the same with the standard RKHS even though hyper-RKHS is much larger than RKHS, which answered **Q1**. Our analysis is non-trivial due to the *non-trivial independence of pairwise samples*, and thus our proof technique can be also beneficial to pairwise learning for non-iid samples. Besides the theoretical contributions, our framework [LSH$^+$21] is able to provide a promising solution to *How to learn an underlying similarity function from a pre-given data-specific matrix?* which extensively exists in kernel learning, manifold learning, and out-of-sample extension [BPV04].

- **A2:** Regarding generalization of over-parameterized models, our work [LSC22] aims to understand over-parameterized two-layer neural networks trained by stochastic gradient descent (SGD), which coincides with practical neural networks training, and accordingly bridges the theoretical gap of previous work depending on the closed-form solution. *Technically*, our analysis shows how to cope with multiple randomness sources of initialization, label noise, and data sampling (as well as stochastic gradients) with no closed-form solution. *Theoretically*, our results are able to characterize the double descent behavior by the unimodality of variance and monotonic decrease of bias. Our finding shows that the constant step-size SGD setting incurs no loss in convergence rate when compared to the exact minimum-norm interpolator, as a theoretical justification of using SGD in practice, which answered **Q2**.

  In fact, the double descent theory can be extended by our work [LLS21] on high-dimensional kernel regression, of which the curve can be unimodal, monotonically decreasing, and double descent under different regularization schemes.

- **A3:** Based on our generalization results, we are able to address some problems in practice. For example, regarding robustness of neural networks, a large numbers of literature in this community have a contradicting conclusion on the fundamental question: over-parameterization helps or hurts robustness? Our work [ZLCC22] aims to investigate this apparent contradiction in theory, and to close the gap as much as possible.

  We demonstrate the *good* in width, the *bad* depth, the *ugly* in initialization regarding the average robustness of DNNs: in the over-parameterized regime, width helps robustness (*good*); depth (*bad*) helps robustness under LeCun initialization but hurts the robustness in both He-initialization and NTK initialization (*ugly*).

- **A4:** Apart from robustness, we are able to analyse the function approximation in deep RL beyond "linear" regime, e.g., NTK, Eluder dimension. This scheme is powerful in practice, e.g., deep Q-network (DQN) using DNNs for function approximation. To bridge the large theory-practice gap, our work [LVC22] study the value iteration algorithm with deep neural function approximation in general function spaces, equipped with the ϵ-greedy exploration, which broadly captures the key features of DQN. Our analysis framework is based on DNNs (as well as two-layer neural networks) where the target Q function lies in the Besov space [Suz19] or the Barron space [EMW21], respectively, to fully capture the properties of Q-functions in terms of smoothness..

  To answer **Q4** as before mentioned, *technically*, our analysis reformulates the temporal difference error in an $L^2(\mathrm{d}\mu)$-integrable space over a certain averaged measure μ, and then transforms the estimation to a *generalization guarantees* problem under the non-iid setting. *Theoretically*, our results demonstrate that the sublinear regret can be achieved for deep neural function approximation under the ϵ-greedy exploration with reasonably finite width and depth in practice. Besides, the relationship between the problem-dependent smoothness of Q-function and regret bounds is also developed. These results could also motivate practitioners to consider different architectures of implementations of deep RL.

Future directions: In the work mentioned above, I have made contributions to several important directions of over-parameterized models in statistical learning theory and reinforcement learning theory. In fact, in each direction, we are just getting started and there is far more to be done. Below I describe two new directions I am eager to explore. They address topics of wide interest, fit well with my expertise, and inspire multidisciplinary collaboration.

Learning with neural networks beyond RKHS: As mentioned before, RKHS is not a large function space and kernel estimator suffers from CoD. From a functional perspective, one key issue corresponds to what norm can be defined and controlled on the functions defined by neural networks, and what suitable function space is for learning via norm capacity based neural networks. Characterizing the "right" function spaces corresponding to neural networks can provide a way to understand their properties, which is a fundamental and significant problem. This would motivate us for better understanding DNNs in function class, training dynamics, and generalization properties.

Efficient function approximation algorithms in RL: General function approximation in RL, particularly theory for deep RL, continues an interesting and unsolved question. Famous theories of function approximation, such as linear mixture model, bilinear class, Bellman rank, Eluder dimension, and low Bellman Eluder dimension, all lack tangible complexity upper bounds for neural network function approximations. How to design both *statistically* and *computationally* efficient RL algorithms in general function approximation is a longstanding question in this community.

References

[Bac17] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(1):629–681, 2017.

[BPV04] Yoshua Bengio, Jean Francois Paiement, and Pascal Vincent. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In *Advances in Neural Information Processing Systems*, pages 177–184, 2004.

[EMW21] Weinan E, Chao Ma, and Lei Wu. The barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, pages 1–38, 2021.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8571–8580, 2018.

[LLS21] Fanghui Liu, Zhenyu Liao, and Johan A.K. Suykens. Kernel regression in high dimensions: Refined analysis beyond double descent. In *International Conference on Artificial Intelligence and Statistics*, pages 649–657, 2021.

[LSC22] Fanghui Liu, Johan A.K. Suykens, and Volkan Cevher. On the double descent of random features models trained with SGD. In *Advances in Neural Information Processing Systems*, 2022.

[LSH+21] Fanghui Liu, Lei Shi, Xiaolin Huang, Jie Yang, and Johan A.K. Suykens. Generalization properties of hyper-rkhs and its applications. *Journal of Machine Learning Research*, 22(140):1–38, 2021.

[LVC22] Fanghui Liu, Luca Viano, and Volkan Cevher. Understanding deep neural function approximation in reinforcement learning via ϵ-greedy exploration. In *Advances in Neural Information Processing Systems*, 2022.

[Suz19] Taiji Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019.

[ZLCC22] Zhenyu Zhu, Fanghui Liu, Grigorios G Chrysos, and Volkan Cevher. Robustness in deep learning: The good (width), the bad (depth), and the ugly (initialization). In *Advances in Neural Information Processing Systems*, 2022.