In machine learning, feature learning is the process of automatically discovering useful representations from data [8]. It reduces traditional reliance on features designed by domain experts and is widely believed to underlie the success of modern models such as neural networks. Yet it remains poorly understood when and why these models produce predictive features and what structure those features capture [10]. My group aims to develop a mathematical foundation to help understand these representations and identify how choices of architecture, optimization, and regularization affect what these models learn.

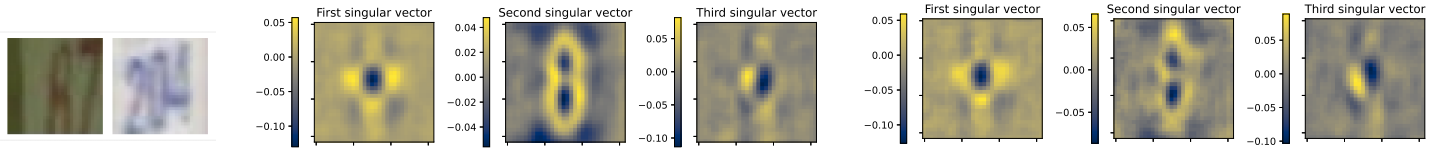
Feature learning has two defining aspects: *composition*, through a predictor $g \circ h$, and *joint optimization*, through the simultaneous learning of the feature map h and predictor g. Our starting point is the *ridge function*—a classical object in approximation theory [9]—which provides a *minimal* model of composition:

$$x \longmapsto f(Ux),$$

where $U \in \mathbf{R}^{d \times d}$ refers to a linear feature map and f is a nonlinear function. We use a reproducing kernel Hilbert space H, e.g., a Sobolev space, as the model class for f. Given $X \in \mathbf{R}^d$ and $Y \in \mathbf{R}$, we introduce the following variational problem [1]:

$$\min_U \min_f \mathbf{E}[(Y - f(UX))^2] + \lambda \|f\|_H^2.$$

Fixing $U = I$ recovers classical kernel ridge regression [6], so the single additional component U *isolates the role of feature learning*. Our central question is *when and why learning a representation reveals predictive structure in the data*. Here, it becomes a precise question about what U recover—analogous to asking what hidden-layer weights learn in a 2-layer neural net.



(a) SVHN samples (b) $f(UX)$ (top 3 singular vectors of U) (c) 2-layer NN (top 3 singular vectors of hidden weights)

Figure 1: On SVHN [7], the $f(UX)$ model learns nontrivial representations, as does a 2-layer neural network [3, Section 10].

Although deceptively simple, the $f(UX)$ problem retains the *nonlinearity and nonconvexity* of its neural network analogue. Surprisingly, it has rich mathematical structure that permits analysis of how features emerge, offering researchers a tractable starting point for identifying feature learning mechanisms and investigating their dependence on architectural choices.

We develop our theory by *varying one ingredient at a time*, addressing four questions about learned representations. First, *what structure in the data does the objective favor?* We characterize population local minima in U, connecting them to variable selection, scale detection, and discrete or clustered data structure [1]. Second, *how does optimization remove noise?* We construct a Riemannian gradient flow with infinitely many monotone quantities that establish monotone suppression of Gaussian noise variables [2]. Third, *how does low-dimensional structure persist with finite samples?* Under the multi-index model, we show that empirical minimizers remain low-rank even without explicit rank or norm penalties on U [3]. Fourth, *which modeling choices affect feature recovery?* For diagonal U, the choice of RKHS H is decisive: local minima U associated with the Laplace RKHS H recover all relevant variables, whereas those associated with the Gaussian RKHS need not [4].

Together, these results establish how the learned representations U depend on the data distribution, architectural choices, and optimization dynamics. By *isolating* architecture, optimization dynamics, finite-sample effects, and regularization *one at a time*, the framework helps explain their individual roles and supplies mathematical benchmarks for investigating analogous phenomena in neural networks, where comparable guarantees remain challenging [2, Postscript].

We envision two complementary directions for this line of research. First, *when can mechanisms identified in kernel models help explain what neural networks learn*? Viewing $f(Ux)$ as a basic abstraction of a neural-net layer—a learned linear transformation U followed by a nonlinear map f—and building on prior work on feature-sparse neural nets [5], my group will investigate conditions under which noise suppression, scale detection, and feature recovery occur in neural-network representations. This will help identify which explanations of feature learning apply across architectures and which depend on specific modeling choices. Second, *how can kernel feature learning be made practical for researchers working with large datasets?* Motivated by the empirical success of feature-learning kernel models on tabular datasets [11, 12], my group will develop scalable methods for optimizing $f(UX)$ using random-feature approximations of f with neural-network optimization techniques. Together, they connect mathematical understanding with practical computation: kernel model $f(UX)$ makes feature learning tractable to analyze, while neural-network optimization methods enable these models practical at scale.

References

Selected work by the group

[1] Yang Li and Feng Ruan, "A Variational Analysis of Kernel Learning with Learnable Linear Transformations," *Foundations of Computational Mathematics*, to appear, 2026.

[2] Yang Li and Feng Ruan, "Gradient flow in the kernel learning problem," *arXiv:2506.08550 [math.OC] (revision at Advances in Mathematics)*, 2025.

[3] Yunlu Chen, Yang Li, Keli Liu, and Feng Ruan, "A Theory of Feature Learning in Kernel Models," *arXiv:2310.11736 [math.ST]*, 2023.

[4] Feng Ruan, Keli Liu, and Michael I. Jordan, "A Compositional Kernel Model for Feature Learning," *Applied and Computational Harmonic Analysis*, to appear, 2026.

[5] Ismael Lemhadri, Feng Ruan, Louis Abraham, and Robert Tibshirani, "LassoNet: A Neural Network with Feature Sparsity," *Journal of Machine Learning Research*, 22(127), 1–29, 2021.

Selected related work

[6] Felipe Cucker and Steve Smale, "On the Mathematical Foundations of Learning," *Bulletin of the American Mathematical Society*, 39(1):1–49, 2002.

[7] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng, "Reading Digits in Natural Images with Unsupervised Feature Learning," *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[8] Yoshua Bengio, Aaron Courville, and Pascal Vincent, "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[9] Allan Pinkus, *Ridge Functions*, Cambridge Tracts in Mathematics, Vol. 205, Cambridge University Press, 2015.

[10] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep Learning," *Nature*, 521:436–444, 2015.

[11] Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin, "Mechanism for Feature Learning in Neural Networks and Backpropagation-Free Machine Learning Models," *Science*, 383(6690), 1461–1467, 2024.

[12] Daniel Beaglehole, David Holzmüller, Adityanarayanan Radhakrishnan, and Mikhail Belkin, "xRFM: Accurate, Scalable, and Interpretable Feature Learning Models for Tabular Data," *International Conference on Learning Representations*, 2026.