To Understand Deep Learning We Need to Understand Kernel Learning

Reproduction of Belkin et al. (2018) on MNIST & Fashion-MNIST

Malo Canu, Yann Chauvard

CS

February 1, 2026

Introduction: The Overfitting Myth

- **The Phenomenon:** Modern DNNs fit training data perfectly (zero loss) yet generalize well.
- **Lack of theory:** Previous generalization bounds depend linearly on $||f^*||_{\mathcal{H}}$ which grows exponentially.
- **The Goal:** Demonstrate that those properties are shared by of kernel machines.
- **Approach:** Test universality across two regimes:
  - **MNIST:** Standard digit recognition benchmark.
  - **Fashion-MNIST:** A "less clean," structurally complex dataset to provide a more rigorous test.

Generalization Study: MNIST vs. Fashion-MNIST

Figure: MNIST Convergence

- **Key Finding:** Both models achieve near-zero training MSE while maintaining high test accuracy (98% for MNIST, 89% for Fashion-MNIST).
- **Dynamics:** Laplacian kernels consistently converge faster than Gaussian kernels due to their "spiky" inductive bias.

Generalization Bound for Gaussian Kernels

- New bound to link exponential increase of the norm of overfitted kernel classifier and data size
- With labeled data $(y_i, x_i)_{i=1,..,n}$ assuming that y is not a deterministic function of x on a non-zero subset, let h a kernel classifier that t-overfit the data, ie achieves zero classification loss and for a fixed portion of training data we have

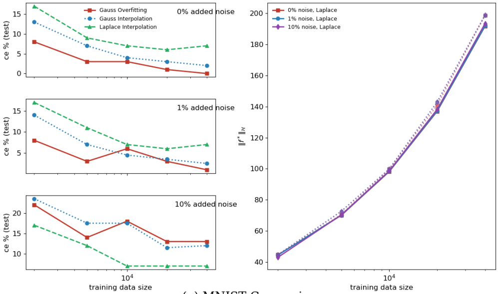
$$\forall i, y_i h(x_i) > t > 0$$

Then, for some constants A, B depending on t and high probability, we have

$$||h|| > Ae^{Bn^{\frac{1}{d}}}$$

with d the dimension of a data point

Robustness to Label Noise



(a) MNIST Comparison

Figure: Noise Robustness RKHS Norm

- **Robustness:** Models manage 10% noise remarkably well, maintaining 92% accuracy.
- **Norm Dynamics:** $\|f^*\|_{\mathcal{H}}$ grows with dataset size, but is largely **insensitive** to noise levels (maybe σ ?).
- **Conclusion:** Kernels decouple global signal from local noise spikes.

Geometry Analysis: The Sinusoid Experiment

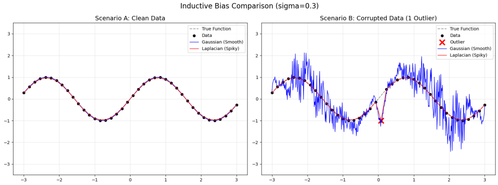


Figure: Fitting a Perturbed Sinusoid

- **Gaussian (Smooth):** Outliers distort the function in a wide neighborhood.
- **Laplacian (Spiky):** Memorizes the outlier with a sharp local peak; preserves global structure.
- **Takeaway:** "Spiky" geometry (like ReLU) is essential for fitting noise without sacrificing generalization.

Conclusion: Inductive Bias is King

1. **Overfitting is a Myth:** Interpolation is a valid regime for high-dimensional data, not a failure.
2. **Minimum Norm:** Success is driven by the algorithm (SGD) selecting the low-complexity solution in the RKHS.
3. **Universality:** Findings hold across simple (MNIST) and complex (Fashion-MNIST) manifolds.
4. **Geometry:** The Laplacian kernel's spikiness mirrors ReLU networks, enabling efficient optimization.