Mind the spikes: Benign overfitting of kernels and neural networks in fixed dimension

Moritz Haas*[1], David Holzmüller*[2], Ulrike von Luxburg[1], Ingo Steinwart[2]

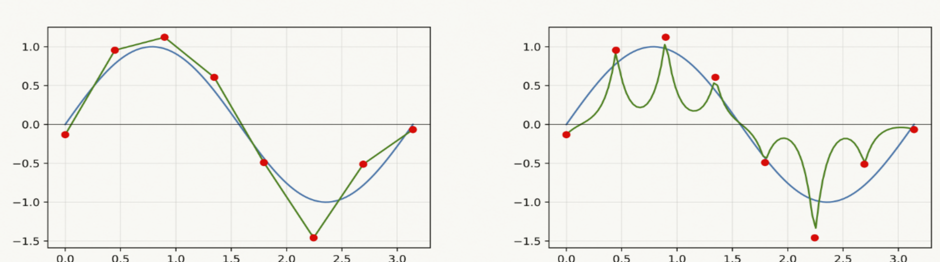
[1] University of Tübingen and Tübingen AI Center, [2] Faculty of Mathematics and Physics, Institute for Stochastics and Applications, University of Stuttgart.

* denotes equal contribution.

Can interpolating models generalize well?

- Traditional view: For good generalization, do not overfit.
- Some neural networks trained to interpolation still generalize well. Why?
- In high-dimensional limits, *benign overfitting* understood for linear and kernel models
- In fixed/low dimension, narrative so far: Overfitting kernels inconsistent

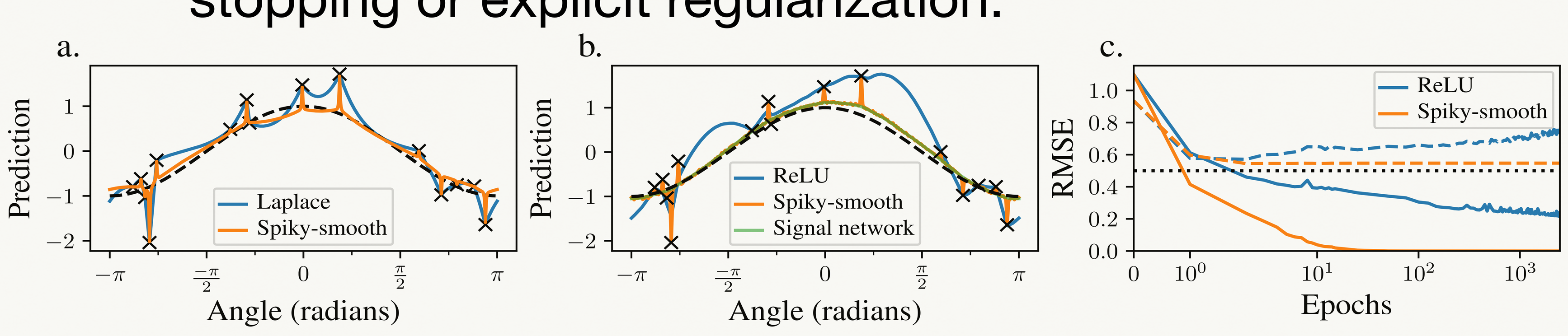
(Rakhlin and Zhai, 2019) (Buchholz, 2022)



Large bandwidth: interpolating noise accumulates errors — *Small bandwidth: Too small between training points*

We show: **Benign overfitting is possible with kernels and neural networks in fixed dimension!**

➡ Simply train to overfit, no need for early stopping or explicit regularization:



Kernel regression — *Neural network regression* — *Training process of neural networks*

Generalized inconsistency results

Not only over bounded, open subsets of $\mathbb{R}^d$ but also over $\mathbb{S}^d$

ReLU NTK RKHS equivalent to $H^{\frac{d+1}{2}}(\mathbb{S}^d)$. (Chen and Xu, 2021) (Bietti and Bach, 2021) →

Corollary: Inconsistency of overfitting (deep) ReLU NNGPs and NTKs

$Var(y|x) \geq \sigma^2$ for all x

Theorem (Buchholz): *Let k be kernel with RKHS equivalent to Sobolev space H^s, $s \in$ ~~[d/2,3d/4]~~. Then under label noise and mild distrib. assump., w.h.p. the* ***min-norm interpol.*** *$\hat{g}_D$ in the RKHS is* ***inconsistent****.*

$s > \frac{d}{2}$

Assume $\hat{f}_D$ in the RKHS fulfills:

(O) Overfitting: Exists $c_{fit} \in (0,1]$: Trainerror$(\hat{f}_D) \leq (1-c_{fit})\,\sigma^2$ for all training sets D.

(N) norm-bounded: Exists $C > 0$: $\|\hat{f}_D\|_{H^s} \leq C \|\hat{g}_D\|_{H^s}$

How to achieve benign overfitting in arbitrary dimension

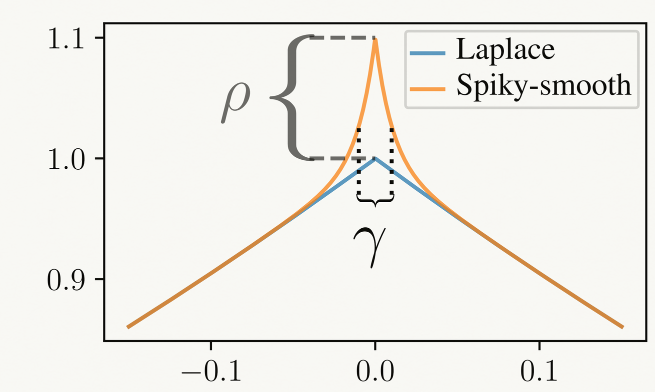
In high dimension (Bartlett et al. 2021): **"min-norm interpol. = smooth + spiky"**

"generalizes well" (smooth) — "interpolates noise in training data with low volume spikes" (spiky)

➡ Break assumption (N) by introducing sharper spikes.

Spiky-smooth kernel: $k_{\rho,\gamma} = \tilde{k} + \rho \cdot k_\gamma$

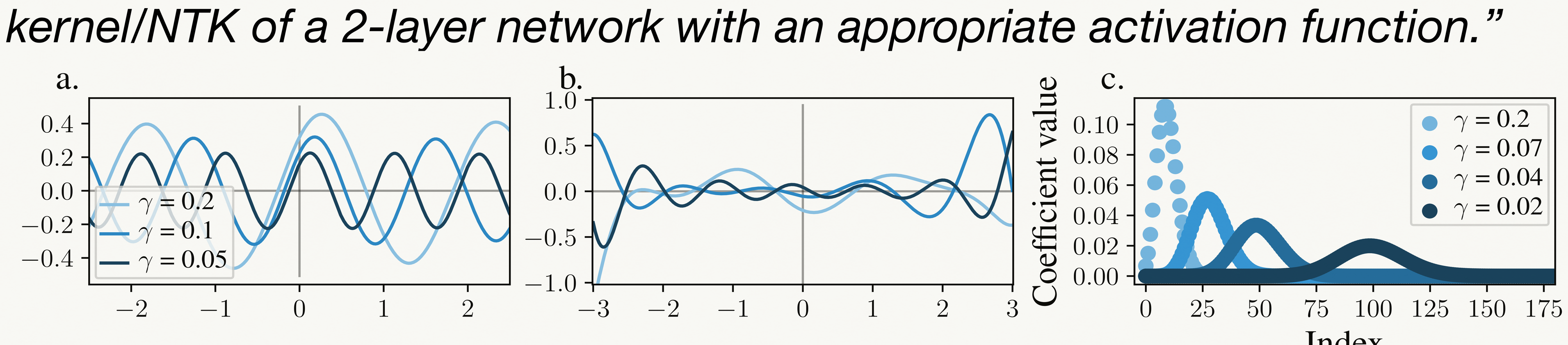
"quasi-regularization" (ρ) — "spike bandwidth" (γ)



Theorem: *Given atom-free distribution and Sobolev target function, choose $\gamma \to 0$ fast enough, $\rho \to 0$ as for kernel ridge regression, then* ***min-norm interpol. of $k_{\rho,\gamma}$ achieves optimal convergence rate.***

Neural networks: Add tiny fluctuations to activation function!

Simon et al (2022): "*Every dot-product kernel on $\mathbb{S}^d$ $\forall d \in \mathbb{N}$ is the NNGP kernel/NTK of a 2-layer network with an appropriate activation function.***"**

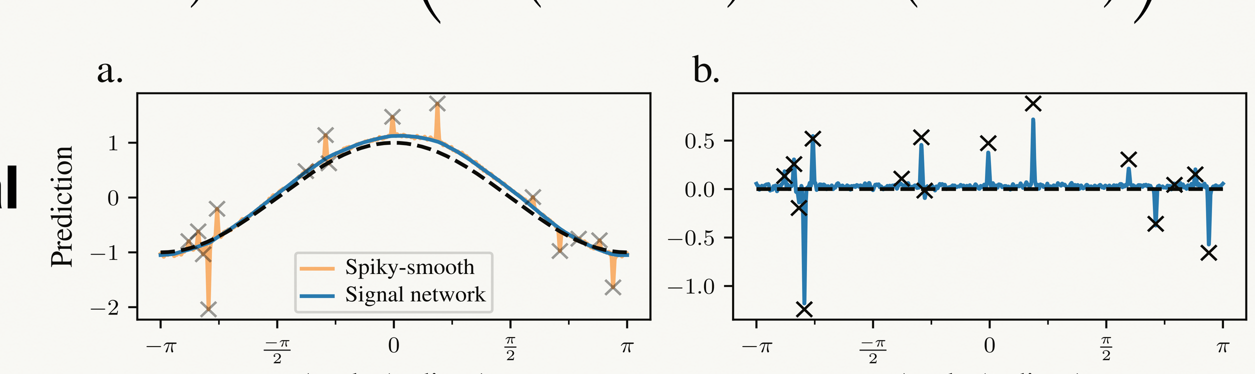


Additive component is approx. small, shifted, high-freq. sin★ — *Or may explode for $|x| \to \infty$* — *Add high-freq. components in Hermite series*

$$\omega_{\mathrm{NNGP}}(x;\gamma) := \sqrt{2}\cdot\sin\left(\sqrt{2/\gamma}\cdot x + \pi/4\right) = \sin\left(\sqrt{2/\gamma}\cdot x\right) + \cos\left(\sqrt{2/\gamma}\cdot x\right),$$

$$★\ \omega_{\mathrm{NTK}}(x;\gamma) := \sqrt{\gamma}\cdot\sin\left(\sqrt{2/\gamma}\cdot x + \pi/4\right) = \sqrt{\gamma/2}\left(\sin\left(\sqrt{2/\gamma}\cdot x\right) + \cos\left(\sqrt{2/\gamma}\cdot x\right)\right).$$

Bonus: **Disentangle signal from spike component**



$$\sigma_{spsm}(x) = ReLU(x) + \omega_{\mathrm{NTK}}(x) \;\rightarrow\; f_{spsm}(\mathbf{x};\theta) = f_{ReLU}(\mathbf{x};\theta) + \left(f_{\omega_{\mathrm{NTK}}}(\mathbf{x};\theta) - b_L\right)$$

Activation function — *Neural network decomposition*

Conclusion

- Harmful overfitting is a generic phenomenon in fixed dimension,
- But can be fixed with spiky-smooth estimators and activation functions

Future work: How can we design activation functions for complex architectures and datasets?

References.
P. Bartlett, A. Montanari, A. Rakhlin. **Deep learning: a statistical viewpoint.** Acta Numerica 2021.
A. Bietti, F. Bach. **Deep Equals Shallow for ReLU Networks in Kernel Regimes.** ICLR 2021.
S. Buchholz. **Kernel Interpolation in Sobolev Spaces is Not Consistent in Low Dimensions.** COLT 2022.
L. Chen, S. Xu. **Deep Neural Tangent Kernel and Laplace Kernel Have the Same RKHS.** ICLR 2021.
A. Rakhlin, X. Zhai. **Consistency of Interpolation with Laplace Kernels is a High-dimensional Phenomenon.** COLT 2019.
J. Simon, S. Anand, M. DeWeese. **Reverse Engineering the Neural Tangent Kernel.** ICML 2022.

Acknowledgements. Funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 - Project 390740016 and EXC 2064/1 - Project 390727645.

imprs-is — EBERHARD KARLS UNIVERSITÄT TÜBINGEN — Universität Stuttgart — machine learning new perspectives for science