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

Moritz Haas*, David Holzmüller*, Ulrike von Luxburg, Ingo Steinwart

* denotes equal contribution.

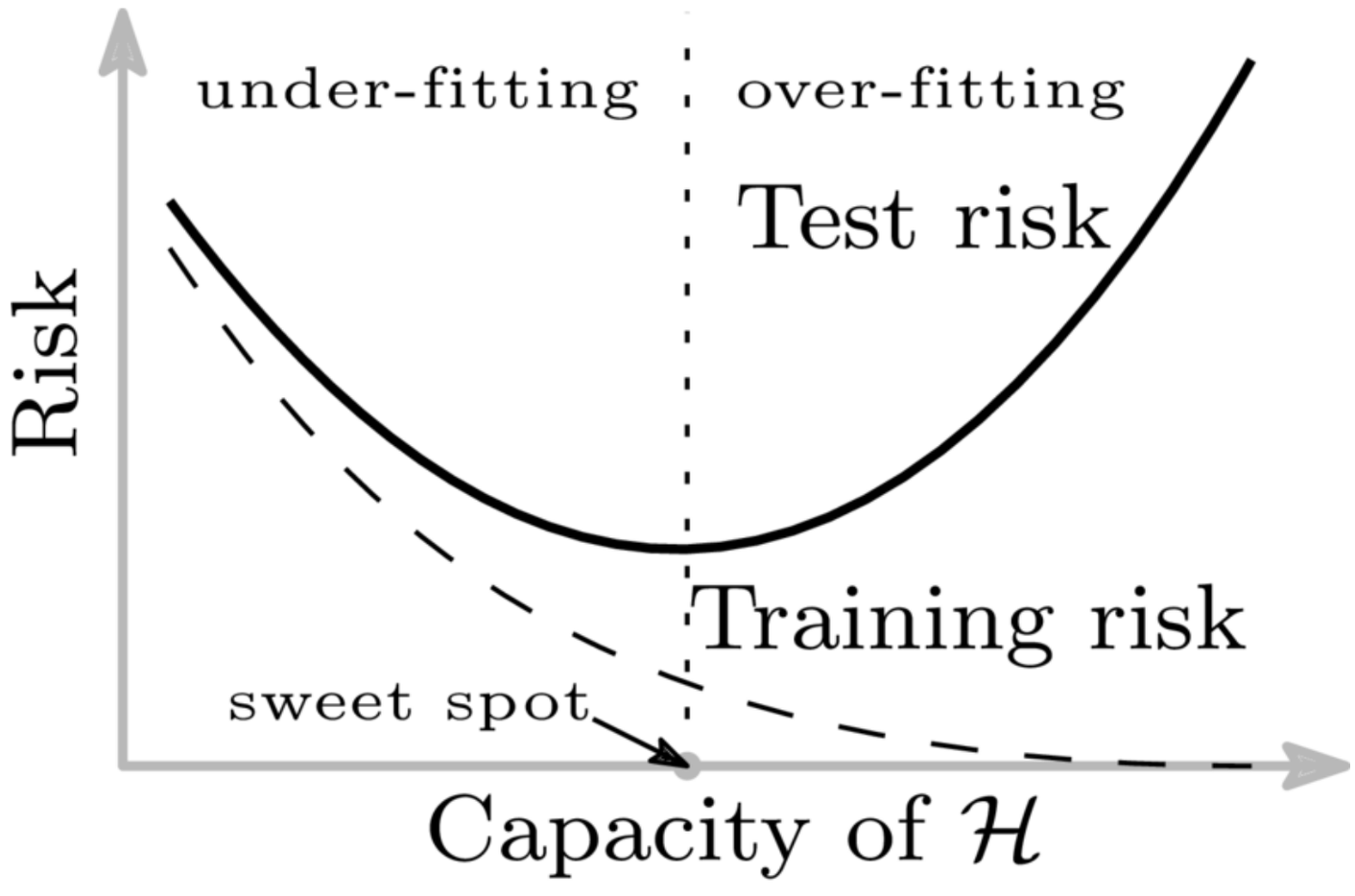
Oberwolfach Workshop

“**Overparametrization**, **Regularization**, Identifiability and Uncertainty **in Machine Learning**”

Benign Overfitting and Double Descent

Hastie, Tibshirani, Friedman. *Elements of Statistical Learning*:

"… interpolating fits... [are] unlikely to predict future data well at all."



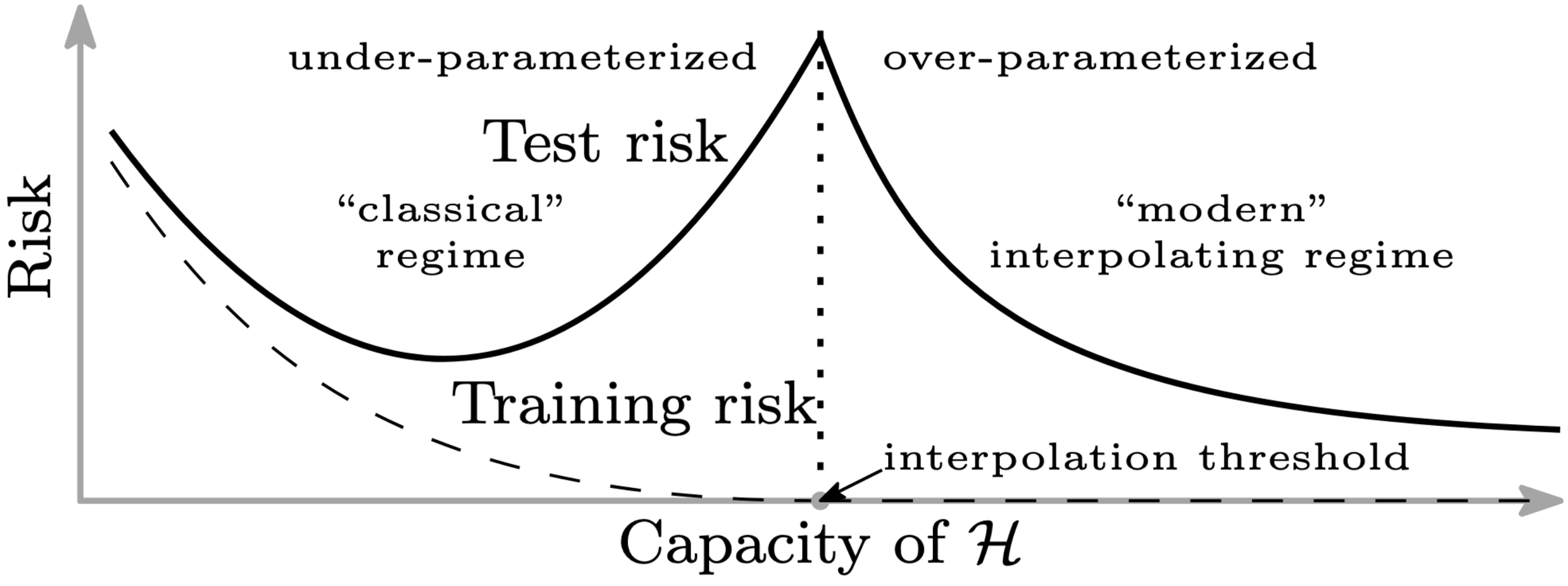
Wyner et al. "Explaining the success of adaboost and random forests as interpolating classifiers." *JMLR 2017*.

Belkin et al. "Reconciling modern machine-learning practice and the classical bias–variance trade-off." *PNAS*, 2019.

Benign Overfitting and Double Descent

Hastie, Tibshirani, Friedman. *Elements of Statistical Learning*:

"… interpolating fits... [are] unlikely to predict future data well at all."



Wyner et al. "Explaining the success of adaboost and random forests as interpolating classifiers." *JMLR 2017*.

Belkin et al. "Reconciling modern machine-learning practice and the classical bias–variance trade-off." *PNAS*, 2019.

Historic Note: Double Descent already in 1989

6. Overfitting *vs.* α.

A very strange observation is that, in the case of PIS, the generalization rate is not a monotonic increasing function of α (see fig. 3). This can be interpreted by the fact that the relative number of small eigenvalues responsible for the explosion of terms in (5) first

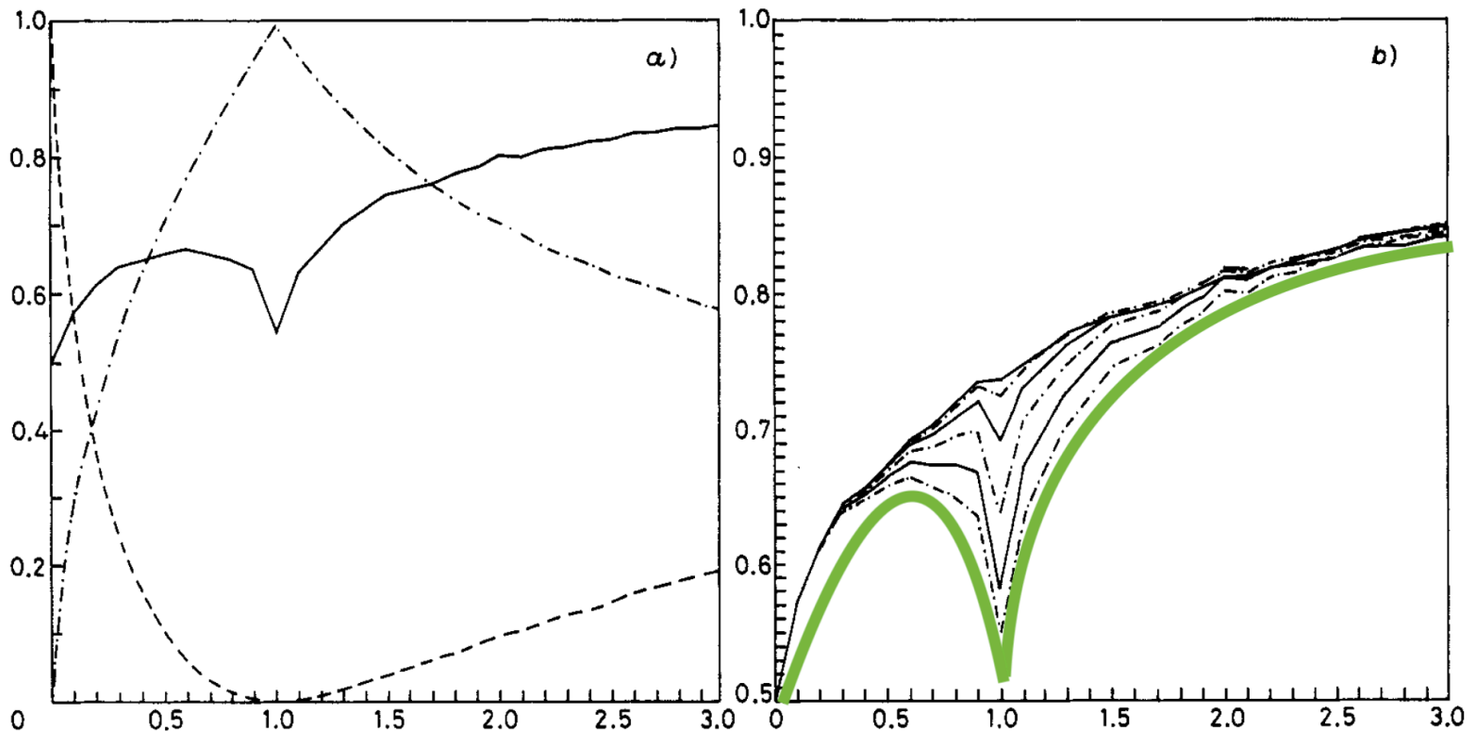


Fig. 3. – *a*) Generalization rate for $k = 1$ (continuous curve), minimum value λ_{min} (dashed curve) and standard deviation (dot-dashed curve) of the normalized eigenvalues of H, *vs.* α. The classification to learn is the MDP ($N = 101$, 46 draws). *b*) Generalization rate for several values of k (0.5, 0.6, 0.7, 0.8, 0.9, 1), *vs.* α, for the MDP ($N = 101$, 46 draws). The curves with deepening valleys correspond to growing values of k.

Vallet, F., J-G. Cailton, and Ph Refregier. "Linear and nonlinear extension of the pseudo-inverse solution for learning boolean functions." ***Europhysics Letters*** **9.4 (1989): 315.**

Setting: Kernel Regression

Given n labeled data points $D := \{(x_i, y_i)\}_{i=1\ldots,n} \subset \mathbb{R}^d \times \mathbb{R}$**.**

Kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ **induces RKHS** $\mathscr{H}$**.**

Setting: Kernel Regression

Given n labeled data points $D := \{(x_i, y_i)\}_{i=1\ldots,n} \subset \mathbb{R}^d \times \mathbb{R}$**.**

Kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ **induces RKHS** $\mathscr{H}$**.**

(Kernel) "ridgeless" regression ($\lambda \to 0$**):**

$$\min_{f \in \mathscr{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathscr{H}}^2$$

is solved by minimum-norm interpolant (MNI)

$$\hat{f}_{\textbf{MNI}} = \arg\min_{f \in \mathscr{H}} \|f\|_{\mathscr{H}} \quad \textbf{s.t.} \quad y_i = f(x_i) \quad \forall i \in [n].$$

Setting: Kernel Regression

Given n labeled data points $D := \{(x_i, y_i)\}_{i=1\ldots,n} \subset \mathbb{R}^d \times \mathbb{R}$**.**

Kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ **induces RKHS** $\mathscr{H}$**.**

(Kernel) "ridgeless" regression ($\lambda \to 0$**):**

$$\min_{f \in \mathscr{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathscr{H}}^2$$

is solved by minimum-norm interpolant (MNI)

$$\hat{f}_{\textbf{MNI}} = \arg\min_{f \in \mathscr{H}} \|f\|_{\mathscr{H}} \quad \textbf{s.t.} \quad y_i = f(x_i) \quad \forall i \in [n].$$

If $\mathbf{K} := \left(k(x_i, x_j)\right)_{i,j=1\ldots,n}$ **invertible, then** $\hat{f}_{\textbf{MNI}}(x) = k(x, \mathbf{X})\, \mathbf{K}^{-1} y\,.$

Previous Work on Benign Overfitting with Kernels

With $n \to \infty$, can MNI generalize near optimally?

	Data Dimension	Distributional assumptions	Generalization
(Liang and Rakhlin, 2018) (Liang et al., 2020) (Ghorbani et al., 2021) (Mei and Montanari, 2022)	$d \to \infty$	**Favorable**	**MNI can generalize near-optimally**
(Rakhlin and Zhai, 2019) (Buchholz, 2022)	d **fixed**	**Weak**	**MNI is inconsistent**

$$x_i \overset{iid}{\sim} P_X \qquad 0 < c \leq P_X \leq C < \infty$$

$$y_i = f^*(x_i) + \varepsilon_i,$$

$$f^* \in C_c^\infty(\Omega) \qquad \varepsilon_i \overset{iid}{\sim} N(0,\sigma),\ \sigma > 0$$

Previous Work on Benign Overfitting with Kernels

With $n \to \infty$, can MNI generalize near optimally?

Overfitting more harmful in fixed dimension d?

Is there any hope?

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

d fixed **Weak** **MNI is inconsistent**

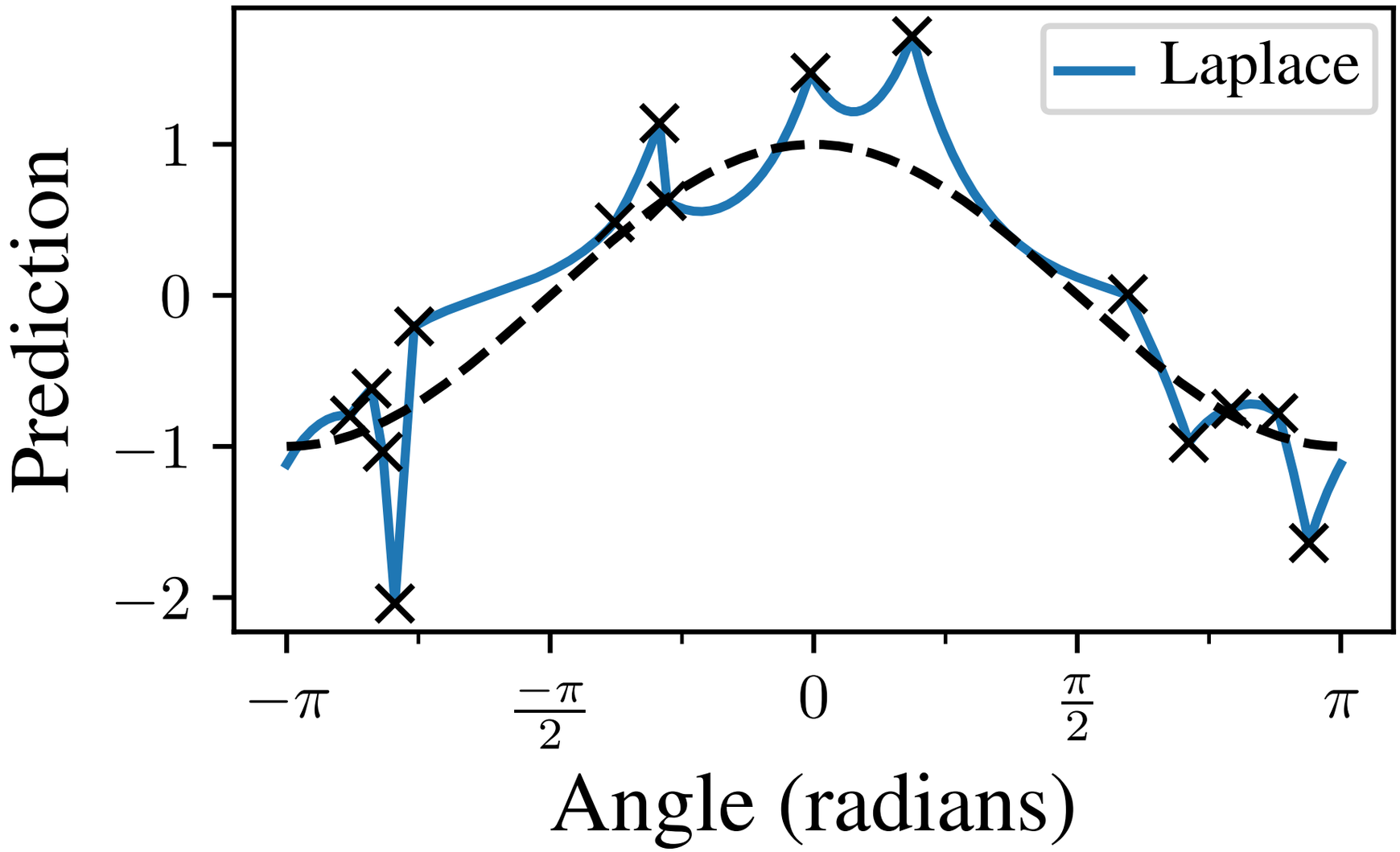
$$x_i \overset{iid}{\sim} P_X \qquad 0 < c \leq P_X \leq C < \infty$$

$$y_i = f^*(x_i) + \varepsilon_i,$$

$$f^* \in C_c^\infty(\Omega) \qquad \varepsilon_i \overset{iid}{\sim} N(0,\sigma),\ \sigma > 0$$

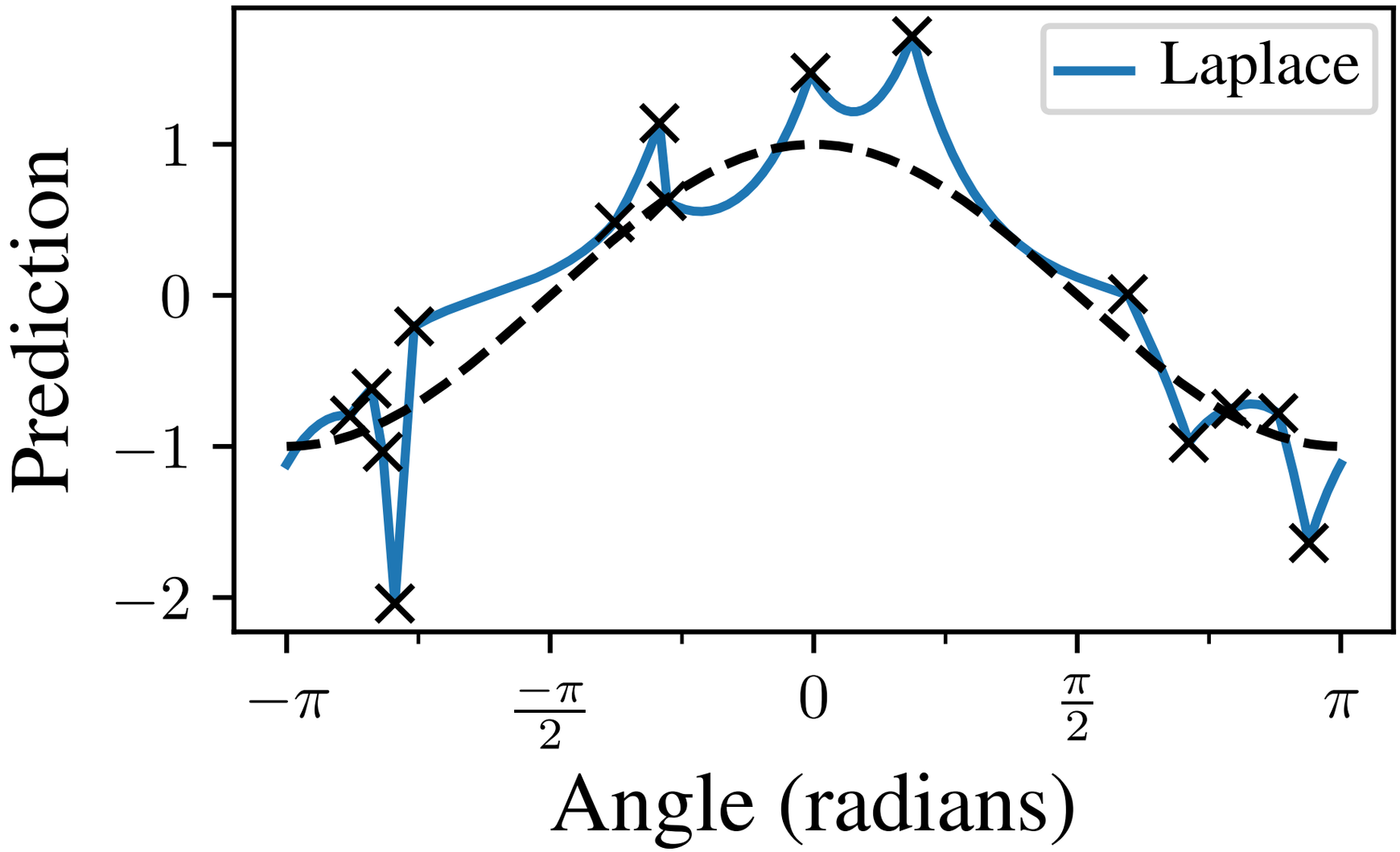
“Min-norm interpolant inconsistent”

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



“Min-norm interpolant inconsistent”

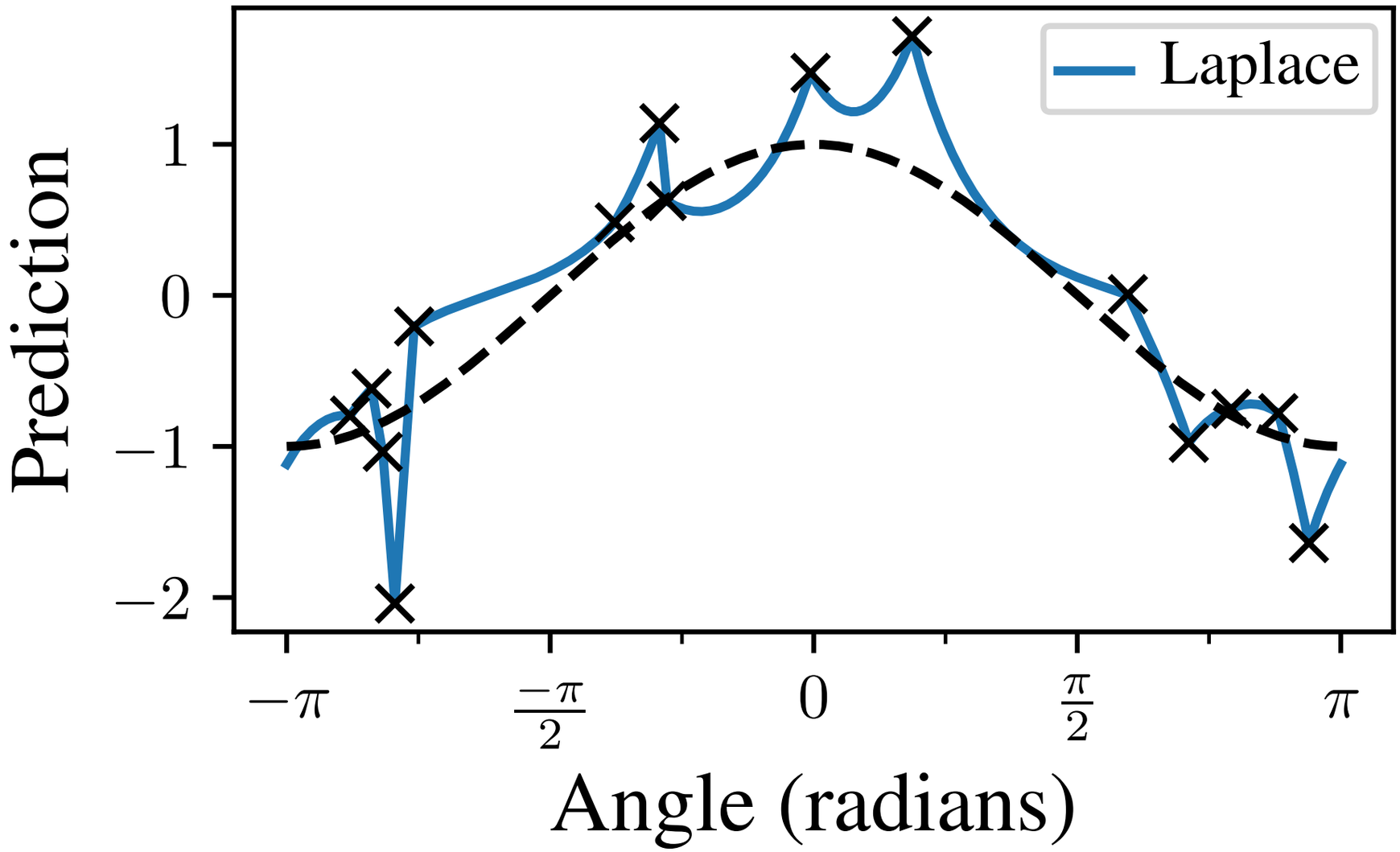
(Rakhlin and Zhai, 2019)
(Buchholz, 2022)



What about other estimators?

“Min-norm interpolant inconsistent”

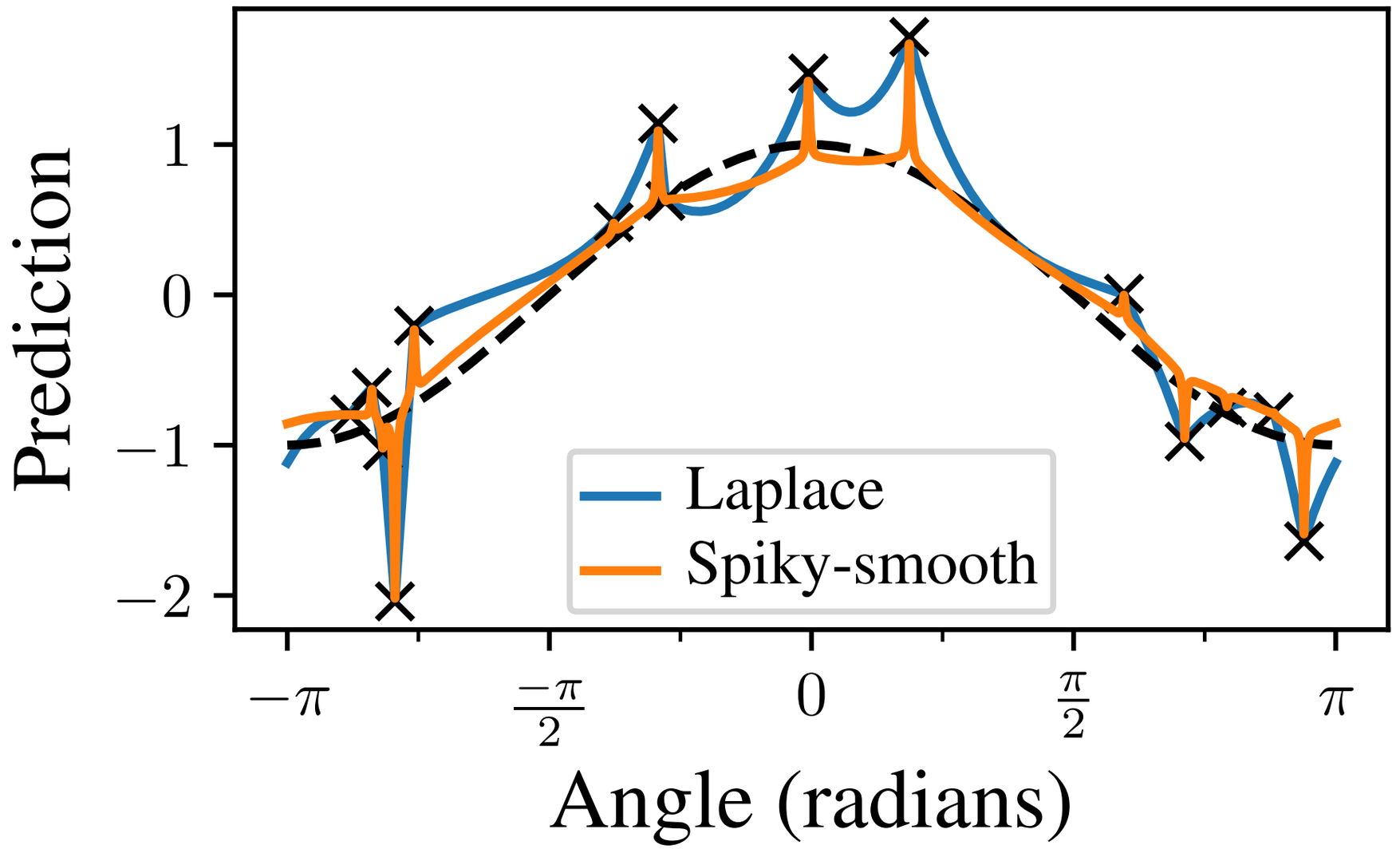
(Rakhlin and Zhai, 2019)
(Buchholz, 2022)



What about other estimators?

Is benign overfitting with kernels/neural networks in fixed dimension impossible?

Our contributions:



Overfitting with common estimators is inconsistent!

With the right ‘spiky-smooth’ kernels/activation functions we achieve rate-optimal generalization while interpolating the training set!

Main Inconsistency Theorem

Theorem (informal):

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

(O) Overfitting: Exists $c_{fit} \in (0,1]$: $\quad$ Trainerror$(\hat{f}) \leq (1 - c_{fit})\,\sigma^2 \qquad \forall D.$

(N) norm-bounded: Exists $C > 0$: $\quad \|\hat{f}\| \leq C\,\|\hat{f}_{\mathbf{MNI}}\|.$

Main Inconsistency Theorem

Theorem (informal):

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

(O) Overfitting: Exists $c_{fit} \in (0,1]$: $\quad$ Trainerror$(\hat{f}) \leq (1 - c_{fit})\,\sigma^2 \qquad \forall D.$

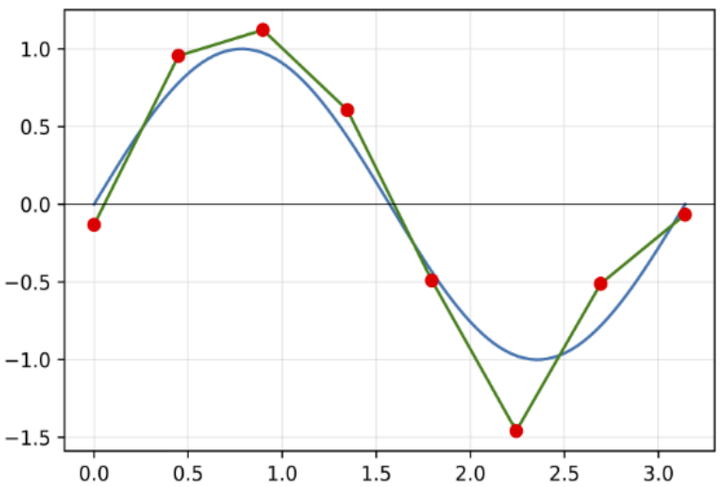
(N) norm-bounded: Exists $C > 0$: $\quad \|\hat{f}\| \leq C\, \|\hat{f}_{\mathbf{MNI}}\|.$

Then w.h.p. $\hat{f}$ is inconsistent, i.e.

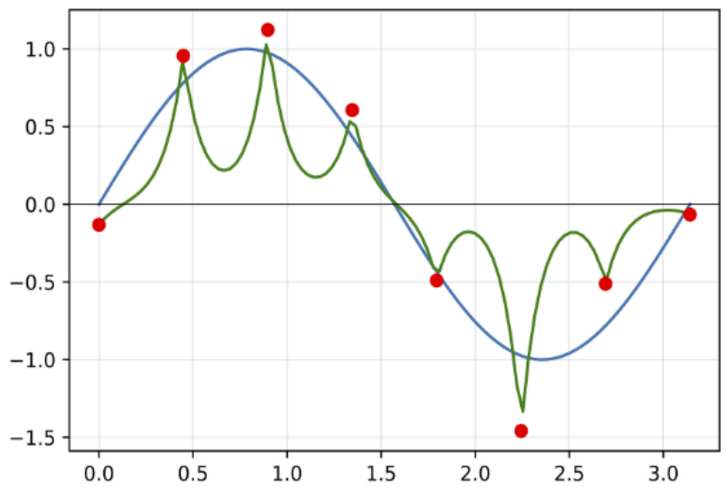
$$\mathbb{E}\left(\hat{f} - f^*\right)^2 \geq c > 0.$$

Proof idea:

Large bandwidths:



Small bandwidths:



Main Inconsistency Theorem

Theorem (informal):

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

(O) Overfitting: Exists $c_{fit} \in (0,1]$: $\quad$ Trainerror$(\hat{f}) \leq (1 - c_{fit})\,\sigma^2 \qquad \forall D.$

(N) norm-bounded: Exists $C > 0$: $\quad \|\hat{f}\| \leq C\, \|\hat{f}_{\mathbf{MNI}}\|.$

Then w.h.p. $\hat{f}$ is inconsistent, i.e.

$$\mathbb{E}\left(\hat{f} - f^*\right)^2 \geq c > 0.$$

(O) necessary: Optimally regularized ridge regression consistent with minimax optimal rates

(N) necessary: next slide

Main Inconsistency Theorem

Theorem (informal):

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

(O) Overfitting: Exists $c_{fit} \in (0,1]$: Trainerror$(\hat{f}) \leq (1 - c_{fit})\,\sigma^2 \qquad \forall D.$

(N) norm-bounded: Exists $C > 0$: $\|\hat{f}\| \leq C\,\|\hat{f}_{\mathbf{MNI}}\|.$

Then w.h.p. $\hat{f}$ is inconsistent, i.e.

$$\mathbb{E}\left(\hat{f} - f^*\right)^2 \geq c > 0.$$

(O) necessary: Optimally regularized ridge regression consistent with minimax optimal rates

(N) necessary: next slide

Other generalizations: more kernels | $Var(y\,|\,x) \geq \sigma^2$ for all x | $supp(\Omega) \subseteq \mathbb{S}^d$

ReLU NTK RKHS equivalent to $H^{\frac{d+1}{2}}(\mathbb{S}^d)$.

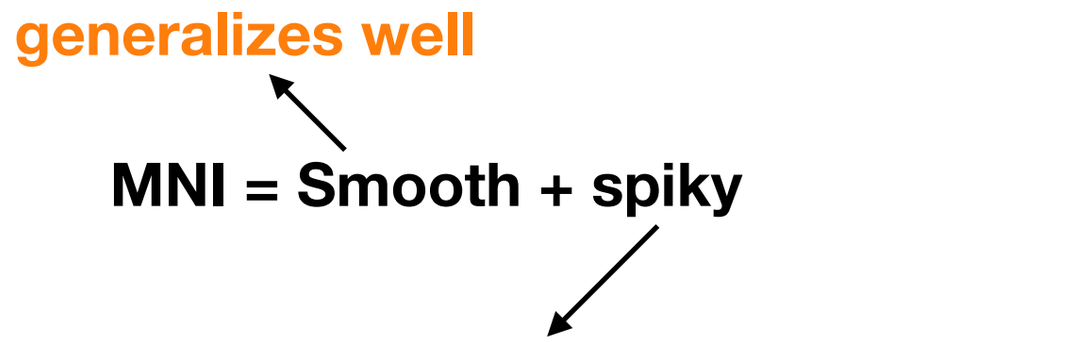
(Chen and Xu, 2021)
(Bietti and Bach, 2021)

Corollary: Under assumptions as above,
overfitting with (deep) ReLU NTKs/NNGPs is inconsistent.

Spiky-smooth kernel sequences

Linear regression in high dimension:

(Bartlett et al., 2021)



Spiky-smooth kernel sequences

Linear regression in high dimension:

(Bartlett et al., 2021)

generalizes well

MNI = Smooth + spiky

interpolates training data and harmless for generalization

Typical distance between training points: $n^{-1/d}$.

Spiky-smooth kernel sequences

Linear regression in high dimension:

(Bartlett et al., 2021)

generalizes well

MNI = Smooth + spiky

interpolates training data and harmless for generalization

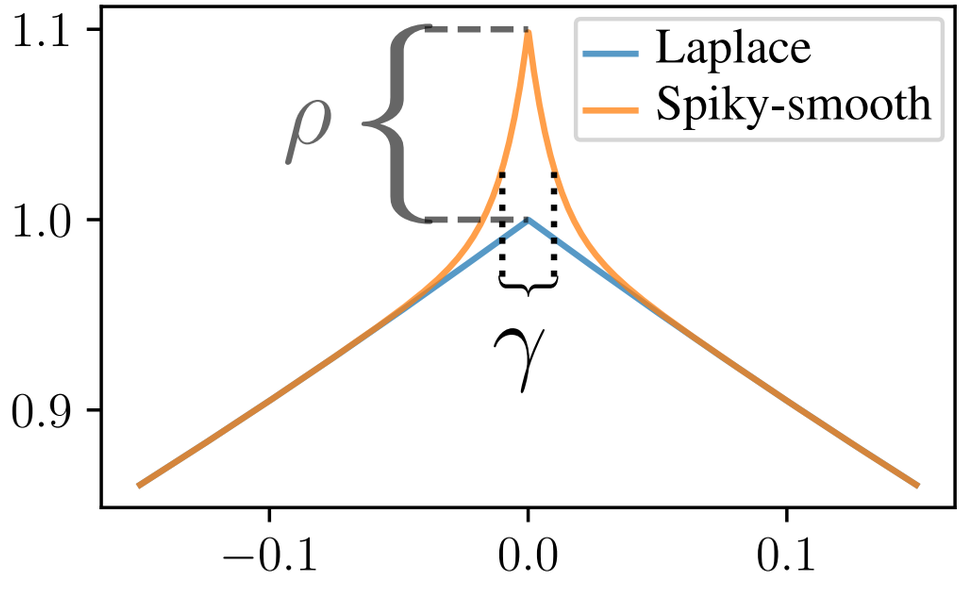
Typical distance between training points: $n^{-1/d}$.

Spiky-smooth kernel sequence:

"quasi-regularization"

$$k_{\rho,\gamma} = \tilde{k} + \rho \cdot k_{\gamma}$$

"spike bandwidth"



Spiky-smooth kernel sequences

Linear regression in high dimension:

(Bartlett et al., 2021)

generalizes well

MNI = Smooth + spiky

interpolates training data and harmless for generalization

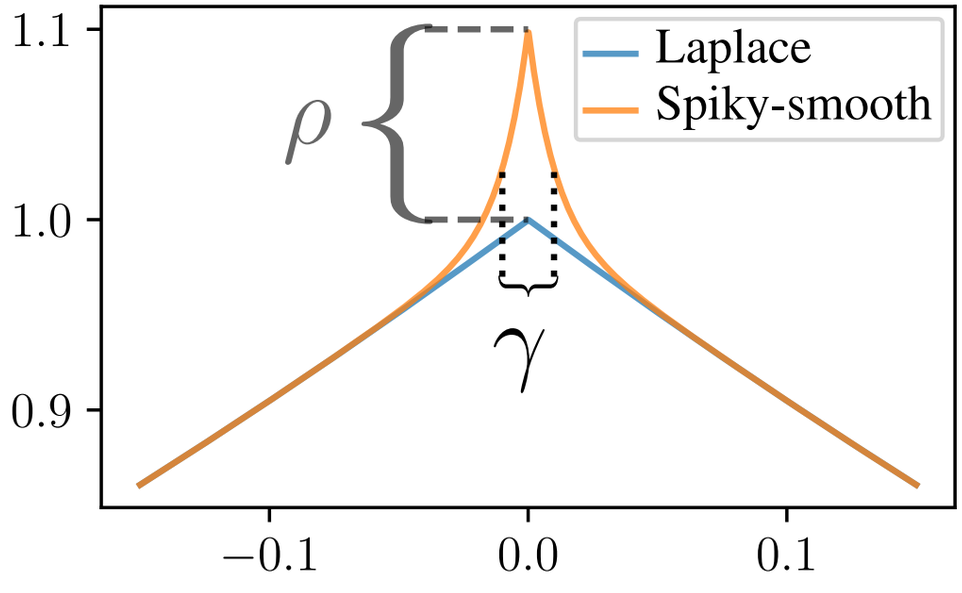
Typical distance between training points: $n^{-1/d}$.

Spiky-smooth kernel sequence:

"quasi-regularization"

$$k_{\rho,\gamma} = \tilde{k} + \rho \cdot k_\gamma$$

"spike bandwidth"



Theorem: If P_X atom-free, $\tilde{k}$ universal, $\rho_n \to 0$ and $n\rho_n \to \infty$, k_γ Laplace kernel with $\gamma_n \leq n^{-\frac{2+\alpha}{d}}((9/4 + \alpha/2)\ln n)^{-1}$, **then the MNI of k_{ρ_n,γ_n} is consistent.**

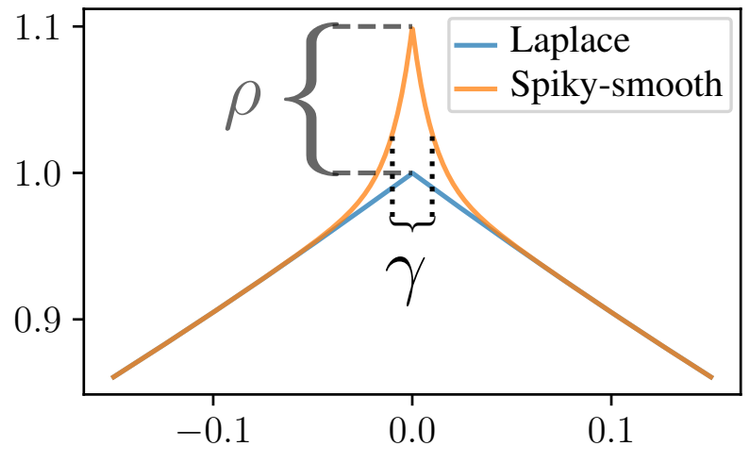
Spiky-smooth kernel sequences

Spiky-smooth kernel sequence:

$$k_{\rho,\gamma} = \tilde{k} + \rho \cdot k_\gamma$$

"quasi-regularization" — ρ

"spike bandwidth" — γ



Theorem: If P_X atom-free, $\tilde{k}$ universal, $\rho_n \to 0$ and $n\rho_n \to \infty$, k_γ Laplace kernel with $\gamma_n \leq n^{-\frac{2+\alpha}{d}}((9/4 + \alpha/2)\ln n)^{-1}$, **then the MNI of k_{ρ_n,γ_n} is consistent.**

Proof idea:

$$\hat{f}_0^k(x) = (\tilde{k} + k_\gamma)(x, \mathbf{X})(\tilde{\mathbf{K}} + \rho_n \mathbf{K}_\gamma)^{-1} y$$

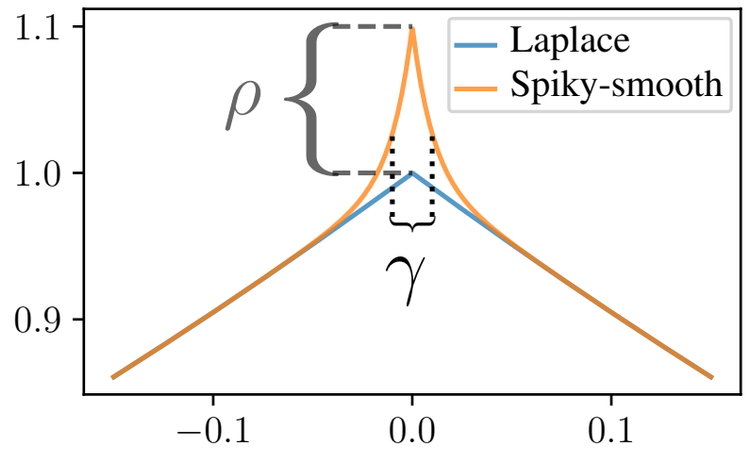
Spiky-smooth kernel sequences

Spiky-smooth kernel sequence:

$$k_{\rho,\gamma} = \tilde{k} + \rho \cdot k_\gamma$$

"quasi-regularization" — ρ

"spike bandwidth" — γ



Theorem: If P_X atom-free, $\tilde{k}$ universal, $\rho_n \to 0$ and $n\rho_n \to \infty$, k_γ Laplace kernel with $\gamma_n \leq n^{-\frac{2+\alpha}{d}}((9/4 + \alpha/2)\ln n)^{-1}$, **then the MNI of k_{ρ_n,γ_n} is consistent.**

Proof idea:

$$\hat{f}_0^k(x) = (\tilde{k} + k_\gamma)(x, \mathbf{X})(\tilde{\mathbf{K}} + \rho_n \mathbf{K}_\gamma)^{-1} y$$

$$= \tilde{k}(x, \mathbf{X})(\tilde{\mathbf{K}} + \rho_n \underbrace{\mathbf{K}_\gamma}_{\to \mathbf{I}})^{-1} y + \underbrace{\rho_n k_\gamma(x, \mathbf{X})}_{\to 0}(\tilde{\mathbf{K}} + \rho_n \mathbf{K}_\gamma)^{-1} y$$

Let $x, x_1, \ldots, x_n \overset{iid}{\sim} P_X$, then w.h.p.:

$$\mathbf{K}_\gamma := k_\gamma(\mathbf{X}, \mathbf{X}) \approx \mathbf{I}$$

$$k_\gamma(x, x_1) \approx 0$$

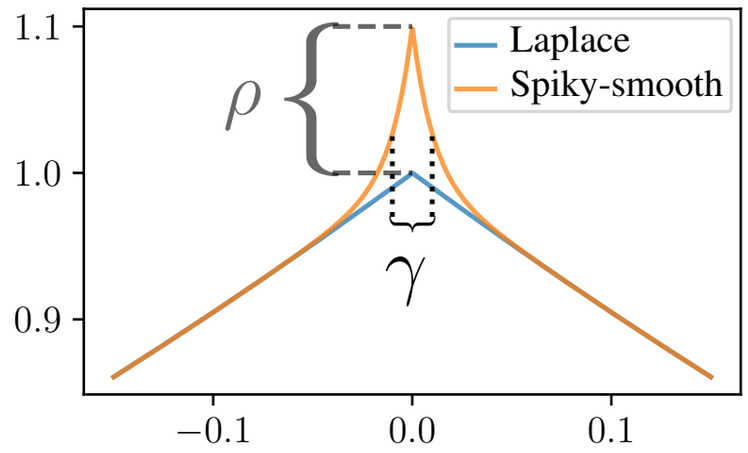
Spiky-smooth kernel sequences

Spiky-smooth kernel sequence:

$$k_{\rho,\gamma} = \tilde{k} + \rho \cdot k_\gamma$$

"quasi-regularization" — ρ

"spike bandwidth" — γ



Theorem: If P_X atom-free, $\tilde{k}$ universal, $\rho_n \to 0$ and $n\rho_n \to \infty$, k_γ Laplace kernel with $\gamma_n \leq n^{-\frac{2+\alpha}{d}}((9/4 + \alpha/2)\ln n)^{-1}$, **then the MNI of k_{ρ_n,γ_n} is consistent.**

Proof idea:

$$\hat{f}_0^k(x) = (\tilde{k} + k_\gamma)(x, \mathbf{X})(\tilde{\mathbf{K}} + \rho_n \mathbf{K}_\gamma)^{-1} y$$

$$= \tilde{k}(x, \mathbf{X})(\tilde{\mathbf{K}} + \rho_n \underbrace{\mathbf{K}_\gamma}_{\to \mathbf{I}})^{-1} y + \underbrace{\rho_n k_\gamma(x, \mathbf{X})}_{\to 0}(\tilde{\mathbf{K}} + \rho_n \mathbf{K}_\gamma)^{-1} y$$

$$= \hat{f}_{\rho_n}^{\tilde{k}}(x)$$

Let $x, x_1, \ldots, x_n \overset{iid}{\sim} P_X$, then w.h.p.:

$$\mathbf{K}_\gamma := k_\gamma(\mathbf{X}, \mathbf{X}) \approx \mathbf{I}$$

$$k_\gamma(x, x_1) \approx 0$$

Mimick kernel ridge regression while interpolating the training set!

Translation to neural networks: Add tiny fluctuations to the activation function

In kernel regime, deep equals shallow → focus on 2 layer NNs (Bietti and Bach, 2021)

Rotation-invariant kernels on $\mathbb{S}^d$ ↔ Activation function of NN in NTK regime

$$\kappa(x) = \sum_{i=0}^{\infty} b_i x^i \quad \leftrightarrow \quad \omega_{NTK}(x) = \sum_{i=0}^{\infty} s_i \sqrt{\frac{b_i}{i+1}}\, h_i(x), \quad \text{(Simon et al., 2022)}$$

where $s_i \in \{-1, +1\}$ and h_i Hermite polynomials.

Translation to neural networks: Add tiny fluctuations to the activation function

In kernel regime, deep equals shallow → focus on 2 layer NNs (Bietti and Bach, 2021)

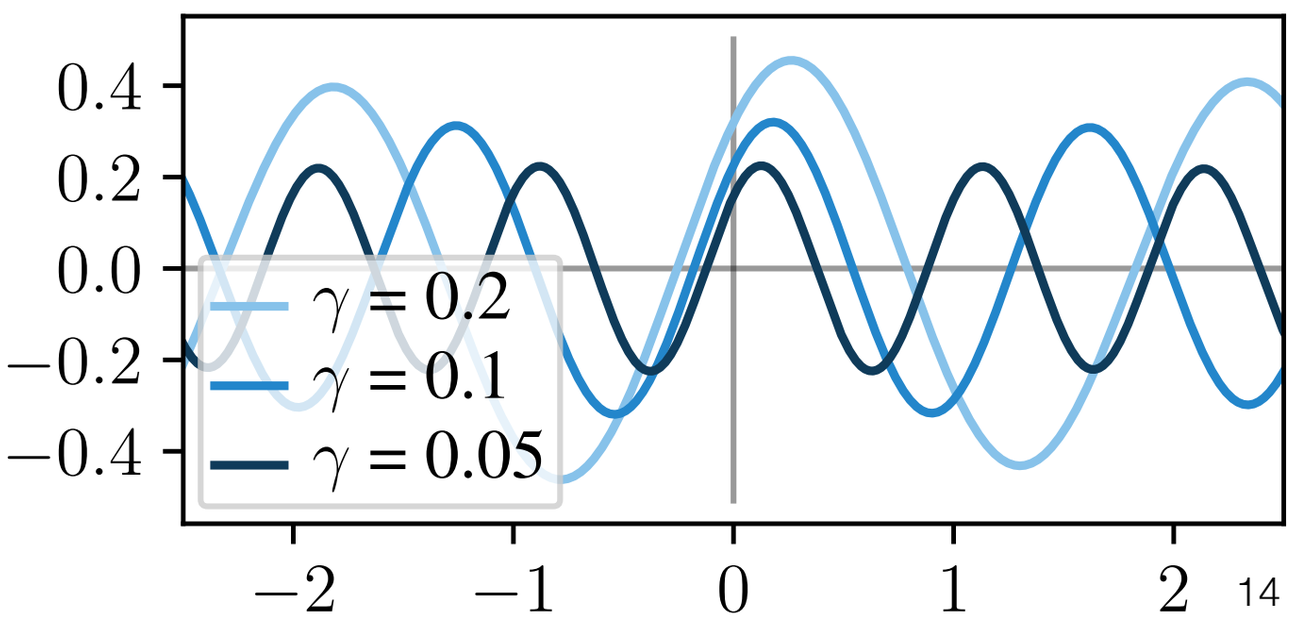
Rotation-invariant kernels on $\mathbb{S}^d$ ↔ **Activation function of NN in NTK regime**

$$\kappa(x) = \sum_{i=0}^{\infty} b_i x^i \quad \leftrightarrow \quad \omega_{NTK}(x) = \sum_{i=0}^{\infty} s_i \sqrt{\frac{b_i}{i+1}}\, h_i(x), \quad \text{(Simon et al., 2022)}$$

where $s_i \in \{-1, +1\}$ and h_i Hermite polynomials.

Add Gaussian kernel as spike k_γ ↔ **Add ω_{NTK}^{Gauss} to activation function**

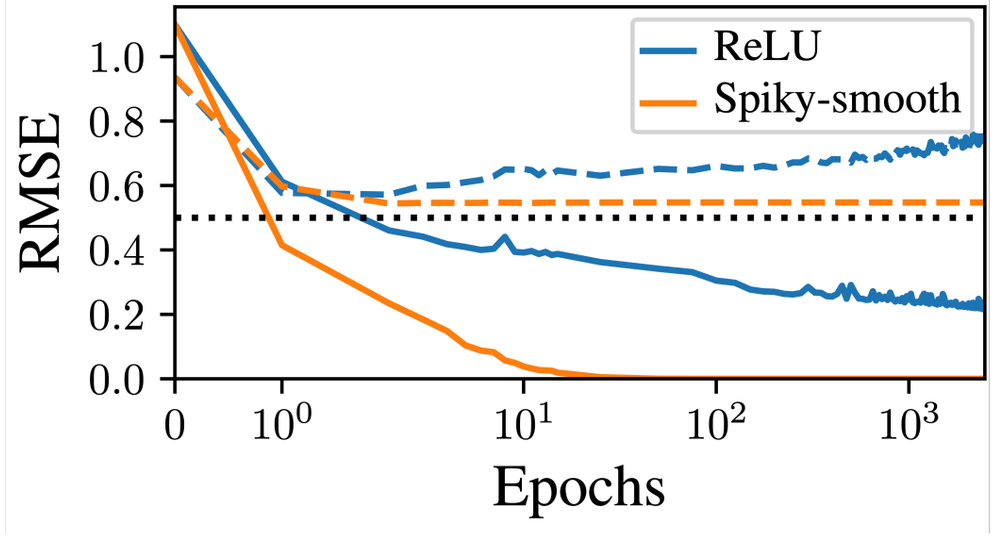
$$\omega_{\mathrm{NTK}}^{Gauss}(x;\gamma) := \sqrt{\gamma} \cdot \sin\left(\sqrt{2/\gamma} \cdot x + \pi/4\right)$$

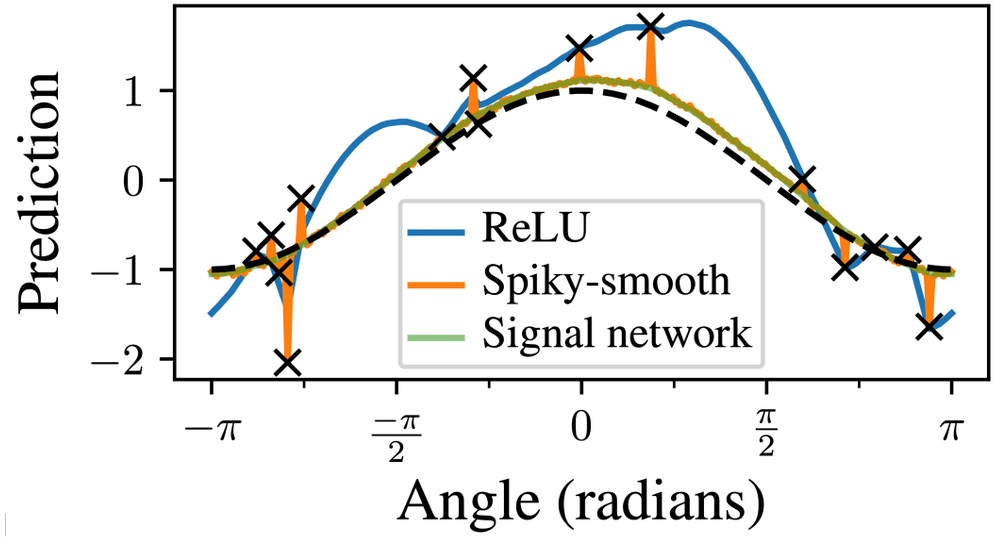


Training Finite Neural Networks

Train 2-layer network with $\left\{\right.$ **ReLU activation**

$$\sigma_{spsm}(x) = ReLU(x) + \rho\,\omega_{\mathrm{NTK}}(x)$$





Needs early stopping

Just train to 0 training error

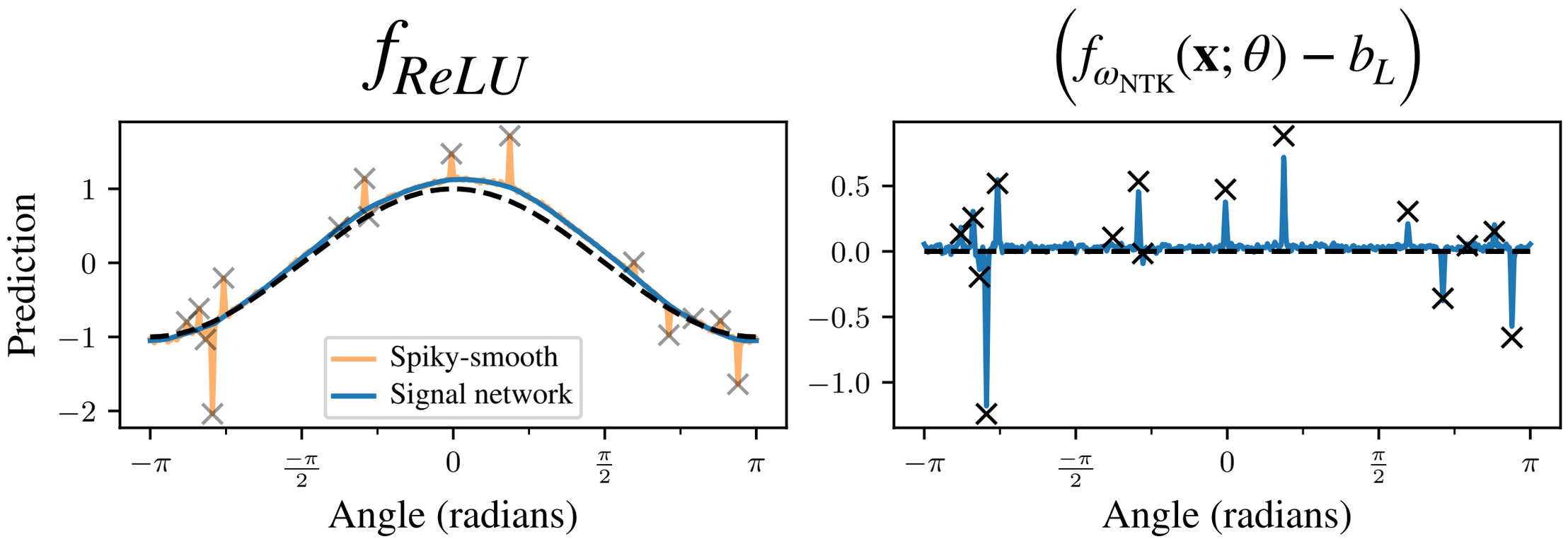
Interpolates too smoothly

Interpolates by forming spikes not harming generalization

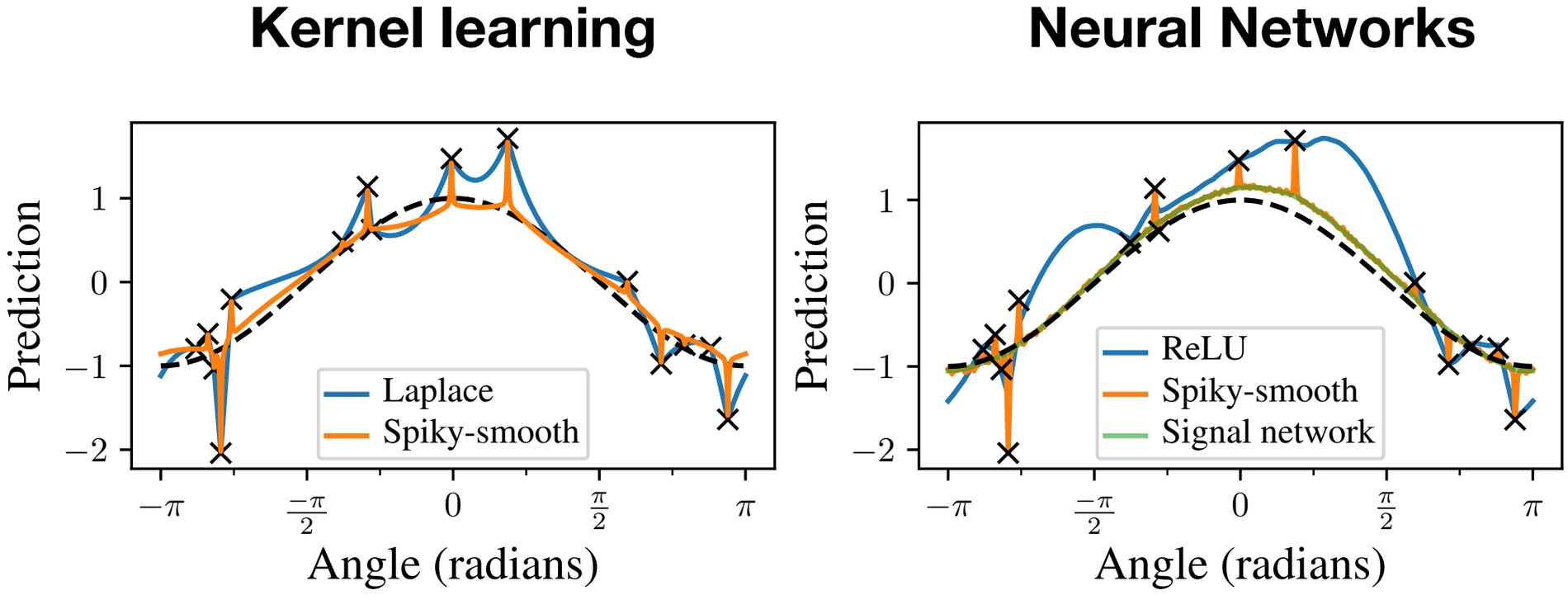
Bonus: Disentangling signal from spike component

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

Activation function *Neural network decomposition*

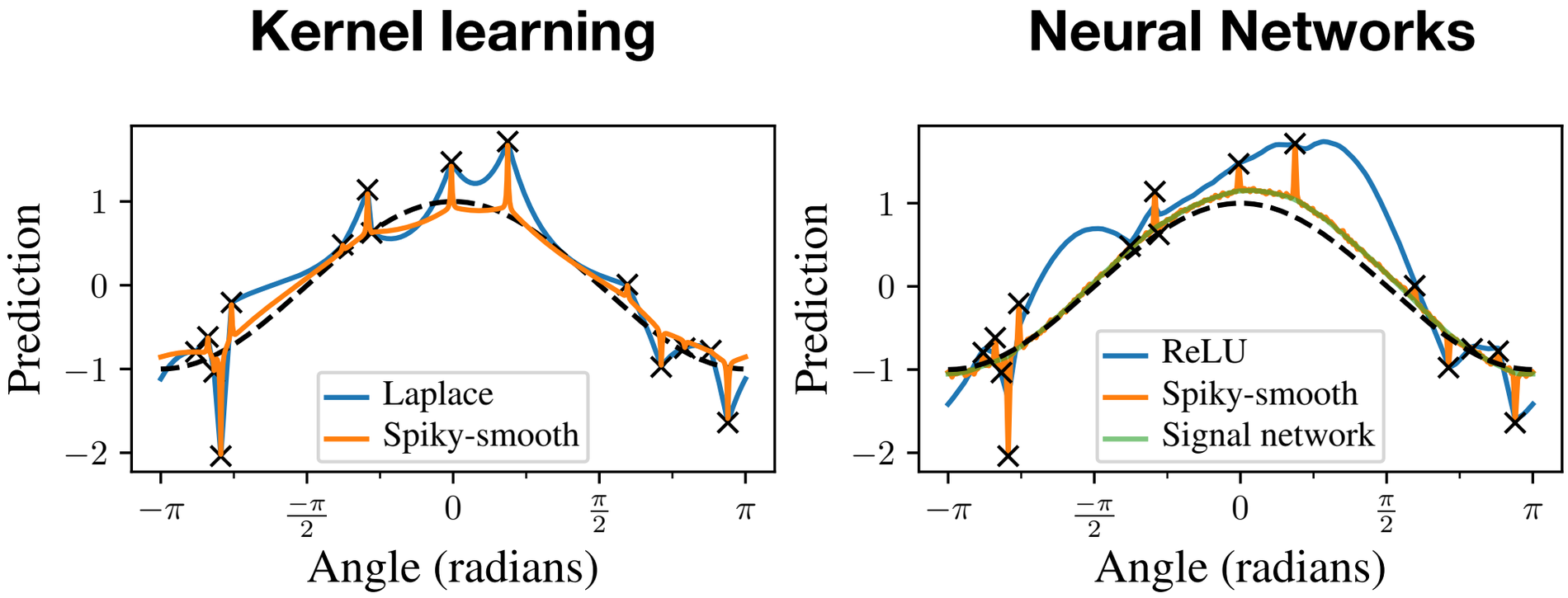


Conclusions



- Common estimators will overfit harmfully in fixed dimension,
- But spiky-smooth estimators and activation functions can generalize as well as optimal regularization

Conclusions

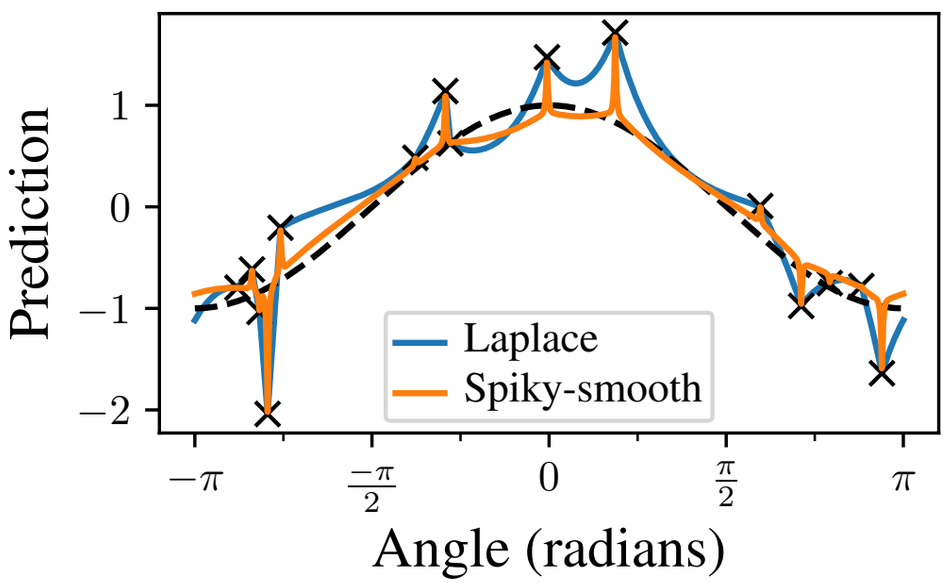


- Common estimators will overfit harmfully in fixed dimension,
- But spiky-smooth estimators and activation functions can generalize as well as optimal regularization

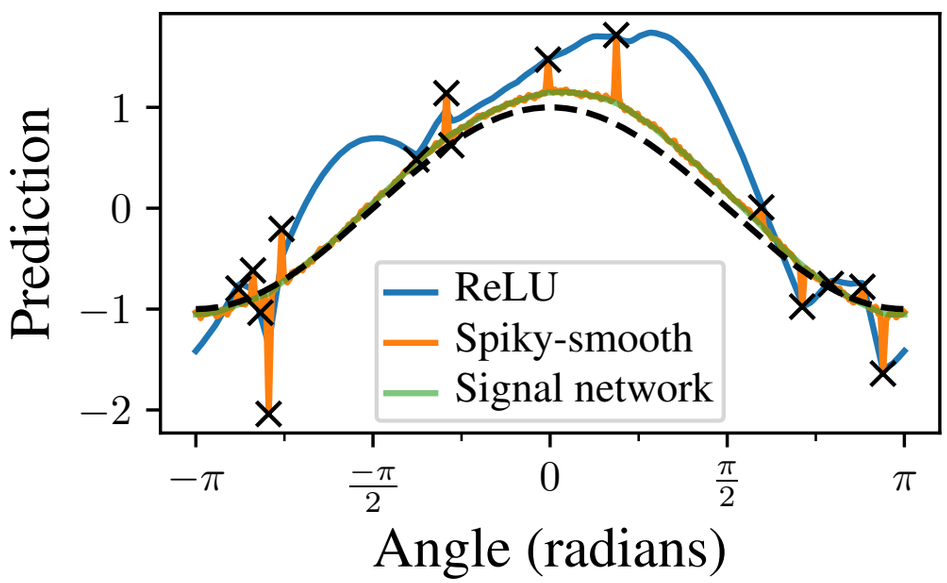
→ **Here, overfitting neither intrinsically helpful nor harmful with the right choice of estimator/kernel/activation function**

Conclusions

Kernel learning



Neural Networks



- Common estimators will overfit harmfully in fixed dimension,
- But spiky-smooth estimators and activation functions can generalize as well as optimal regularization

→ **Here, overfitting neither intrinsically helpful nor harmful with the right choice of estimator/kernel/activation function**

How can we adapt feature learning neural networks to overfit benignly in realistic settings arbitrary dimensions?

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. Daniely, R. Frostig, Y. Singer. **Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity.** NeurIPS 2016.

B. Ghorbani et al. **Linearized two-layers neural networks in high dimension.** The Annals of Statistics 49 (2), 1029-1054, 2021.

T. Liang, A. Rakhlin. **Just interpolate: Kernel" ridgeless" regression can generalize.** arXiv:1808.00387, 2018.

T. Liang, A. Rakhlin, X. Zhai. **On the Multiple Descent of Minimum-Norm Interpolants and Restricted Lower Isometry of Kernels**. COLT 2020.

M. Loog et al. **A brief prehistory of double descent.** PNAS 2020.

N. Mallinar et al. **Benign, tempered, or catastrophic: Toward a refined taxonomy of overfitting.** NeurIPS 2022.

S. Mei, A. Montanari. **The generalization error of random features regression: Precise asymptotics and the double descent curve**. Communications on Pure and Applied Mathematics 75 (4), 667-766 (2022).

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.

Further Related Work

First observations of double descent that I know of:

Vallet, F., J-G. Cailton, and Ph Refregier. "Linear and nonlinear extension of the pseudo-inverse solution for learning boolean functions." *Europhysics Letters* 9.4 (1989): 315.

Opper, Manfred, et al. "On the ability of the optimal perceptron to generalise." *Journal of Physics A: Mathematical and General* 23.11 (1990).

Benign overfitting in linear regression requires many similarly unimportant directions, in particular $d \gg n$. In infinite dimension, needs spectral decay $\lambda_k \propto k^{-1}\log^{\alpha}(k),\ \alpha > 1$.

Bartlett et al. “Benign overfitting in linear regression.” *PNAS* (2020).

Benign overfitting in kernel regression requires kernel spectral decay $\lambda_k \propto k^{-1}\log^{\alpha}(k),\ \alpha > 1$:

Mallinar, Neil, et al. "Benign, tempered, or catastrophic: Toward a refined taxonomy of overfitting." *NeurIPS* (2022). (*semi-rigorous result)

Asymptotic analysis for random feature ridge regression in proportional limit:

Mei, Song, and Andrea Montanari. "The generalization error of random features regression: Precise asymptotics and the double descent curve." *Communications on Pure and Applied Mathematics* 75.4 (2022): 667-766.

Nadaraya-Watson estim. achieves optimal non-param. rates with singular kernels:

Belkin, Mikhail, Alexander Rakhlin, and Alexandre B. Tsybakov. "Does data interpolation contradict statistical optimality?" *AISTATS* 2019.

Benign overfitting for classification is much more generic than for regression:

Shamir, Ohad. "The implicit bias of benign overfitting." *COLT*, 2022.

Muthukumar, Vidya, et al. "Classification vs regression in overparameterized regimes: Does the loss function matter?" *JMLR* 22.222 (2021): 1-69.

Multiple descent in high-dim kernel regression:

Liang et al. “On the Multiple Descent of Minimum-Norm Interpolants and Restricted Lower Isometry of Kernels.” *COLT* (2020).

RF and NT models learn degree-l polynomial parts of target fct. When $d^l \ll n \ll d^{l+1}$:

Ghorbani, Behrooz, et al. "Linearized two-layers neural networks in high dimension.” *The Annals of Statistics* 49.2 (2021): 1029-1054.

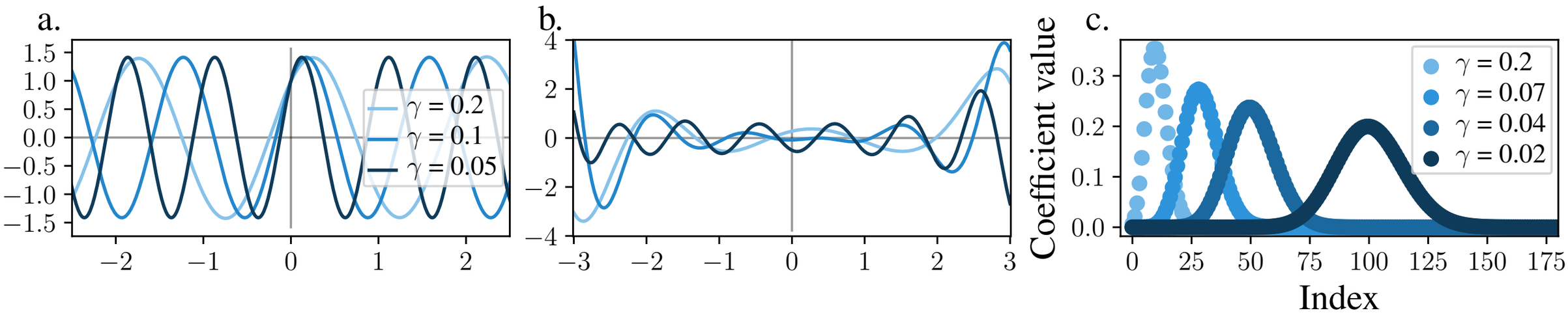
Spiky-smooth NNGP Activation Functions

Rotation-invariant kernels on $\mathbb{S}^d$ ⟷ **Activation function of NN in NTK regime**

$$\kappa(x) = \sum_{i=0}^{\infty} b_i x^i \quad \longleftrightarrow \quad \omega_{NNGP}(x) = \sum_{i=0}^{\infty} s_i \sqrt{b_i}\, h_i(x), \quad \text{(Daniely et al., 2016)}$$

where $s_i \in \{-1, +1\}$ and h_i Hermite polynomials.

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



L_2-norm and amplitude invariant to γ: $\|\omega_{NNGP}^{Gauss}\|_{L_2(N(0,1))} = 1$

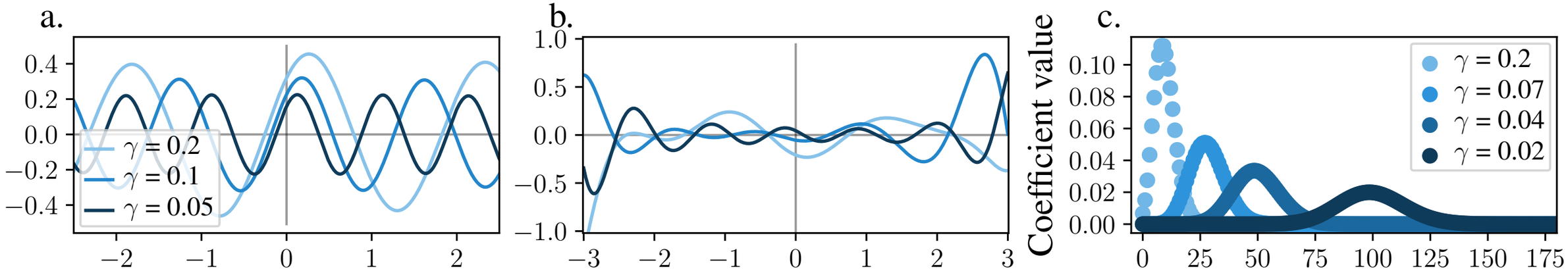
Spiky-Smooth NTK Activation Functions

In kernel regime, deep equals shallow → focus on 2 layers

Rotation-invariant kernels on $\mathbb{S}^d$ ↔ **Activation function of NN in NTK regime**

$$\kappa(x) = \sum_{i=0}^{\infty} b_i x^i \quad \leftrightarrow \quad \omega_{NTK}(x) = \sum_{i=0}^{\infty} s_i \sqrt{\frac{b_i}{i+1}}\, h_i(x), \quad \text{(Simon et al., 2022)}$$

where $s_i \in \{-1, +1\}$ and h_i Hermite polynomials.



$s_i = +1$ iff $\lfloor i/2 \rfloor$ even

s_i random

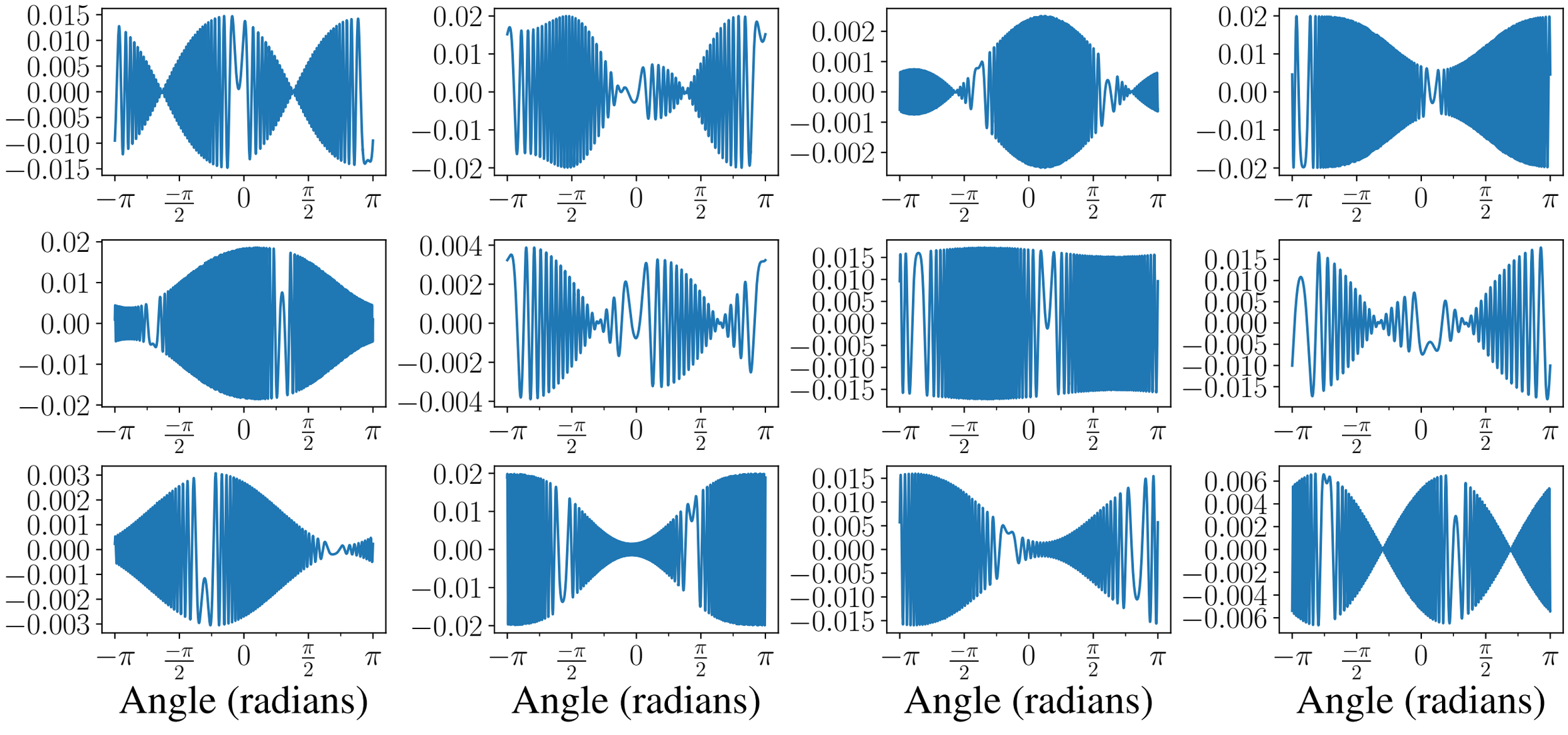
$$\|\omega_{NTK}^{Gauss}\|_{L_2(N(0,1))} \approx \frac{\gamma}{2}$$

Approximated by

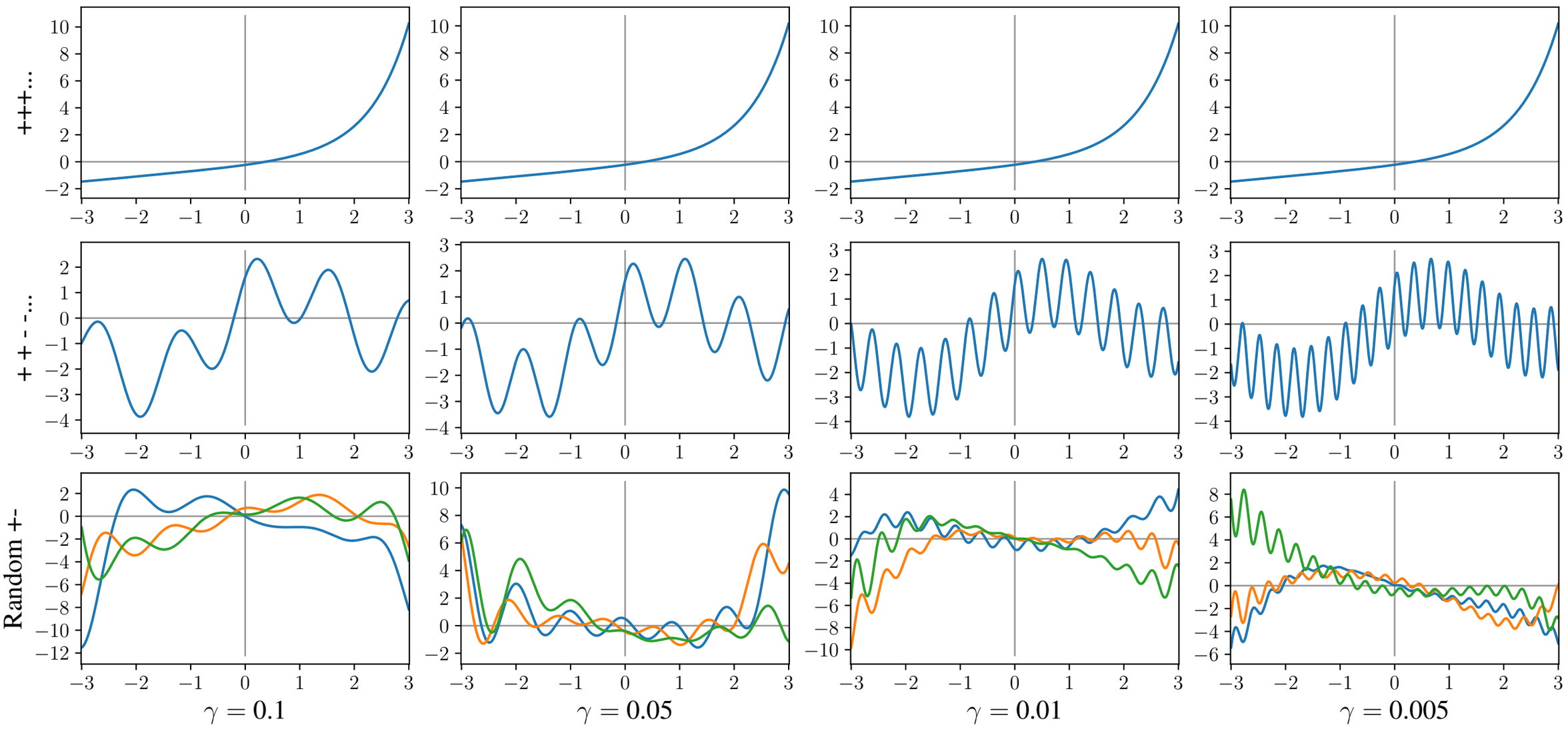
$$\omega_{\mathrm{NTK}}^{Gauss}(x;\gamma) := \sqrt{\gamma} \cdot \sin\left(\sqrt{2/\gamma} \cdot x + \pi/4\right) = \sqrt{\gamma/2}\left(\sin\left(\sqrt{2/\gamma} \cdot x\right) + \cos\left(\sqrt{2/\gamma} \cdot x\right)\right).$$

Constructive and destructive interference?

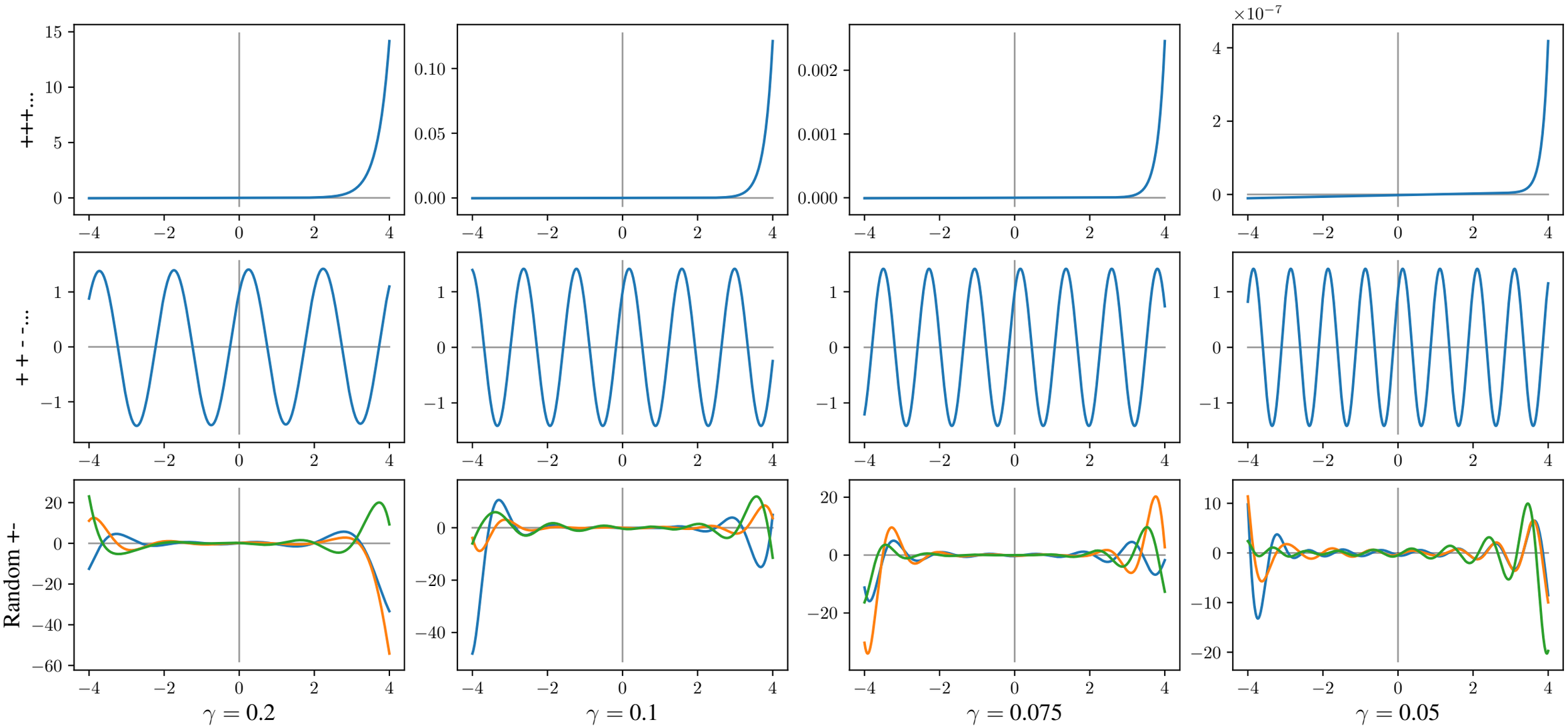
Functions learned by 12 random hidden layer neurons of the spike component network:



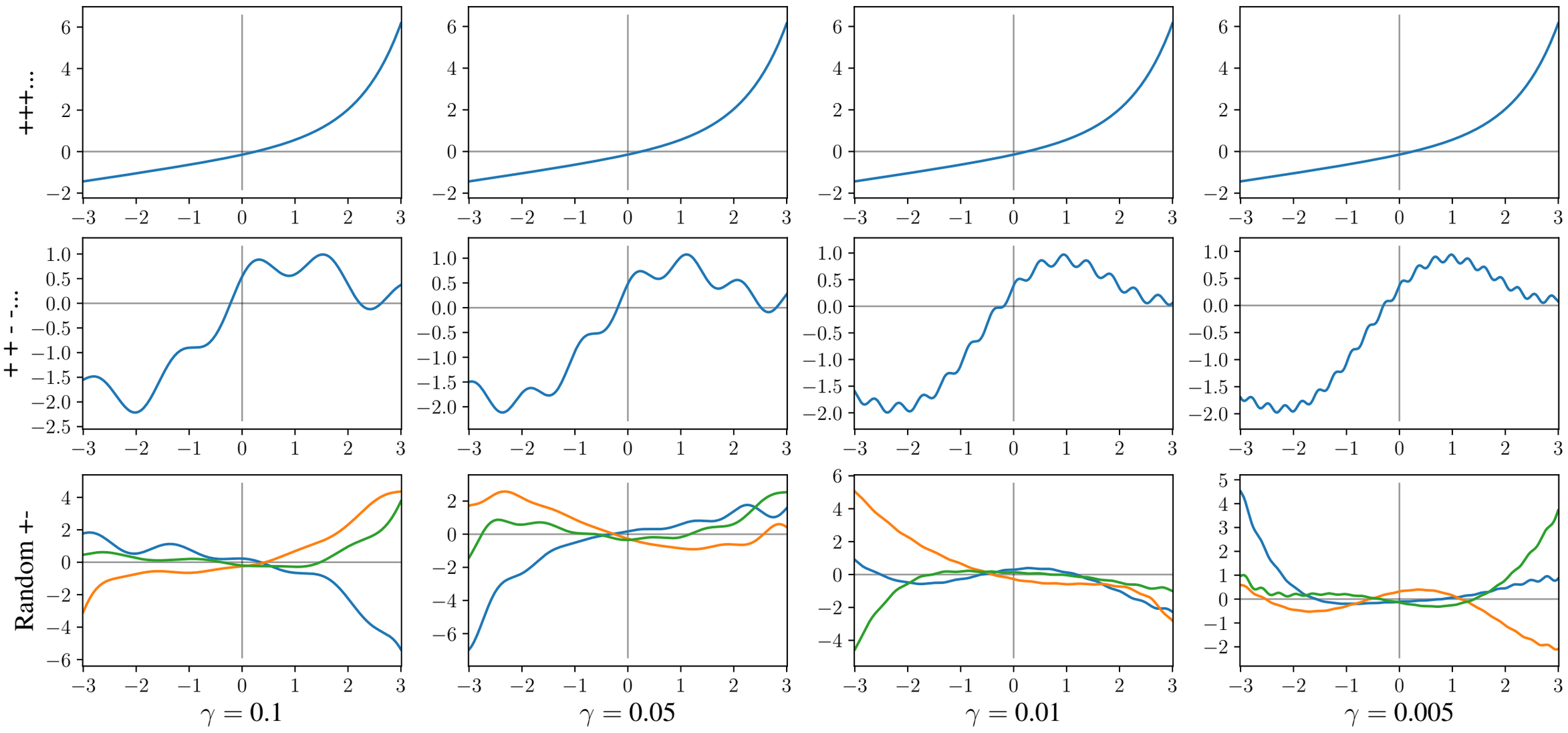
Spiky-smooth NNGP activation functions



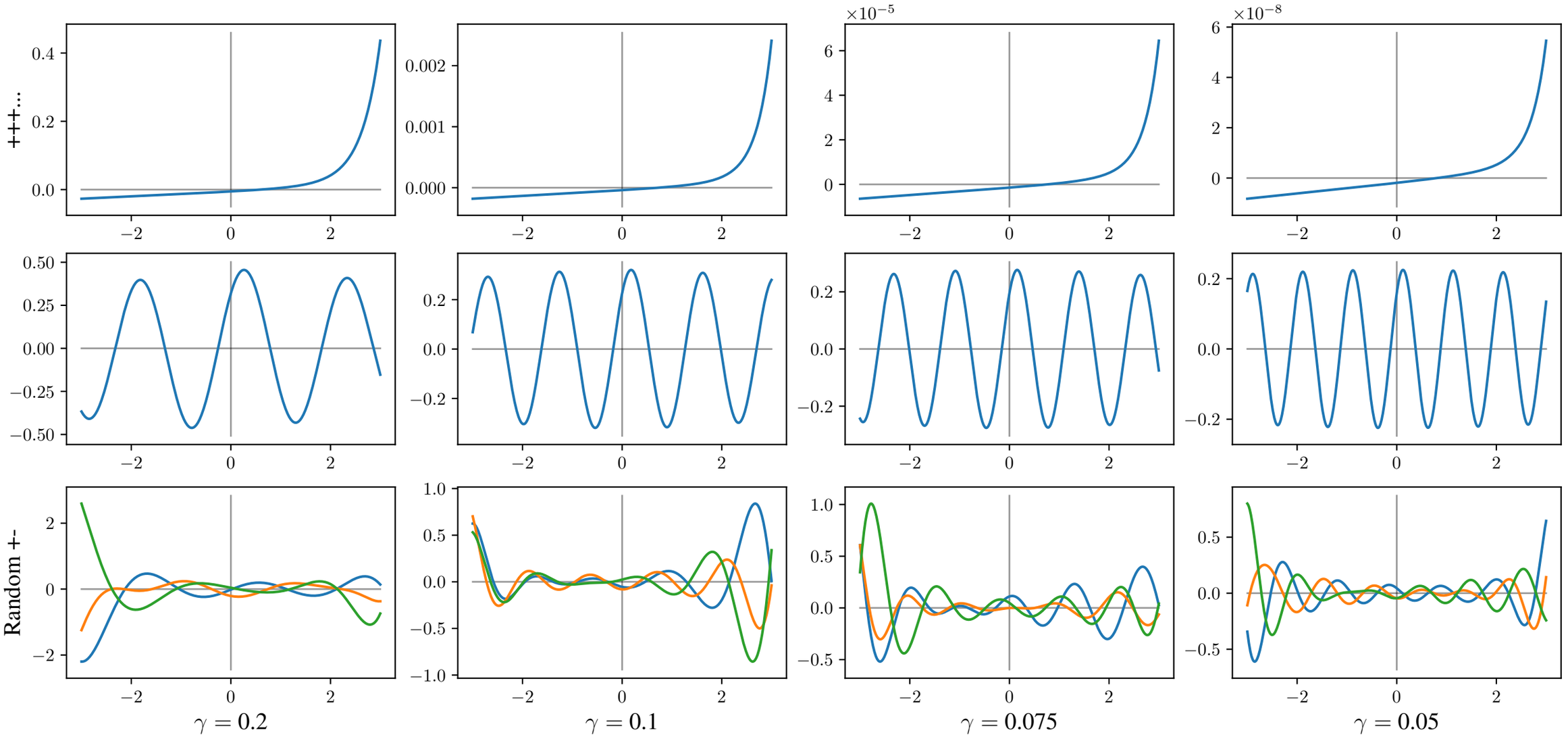
Spiky NNGP activation functions



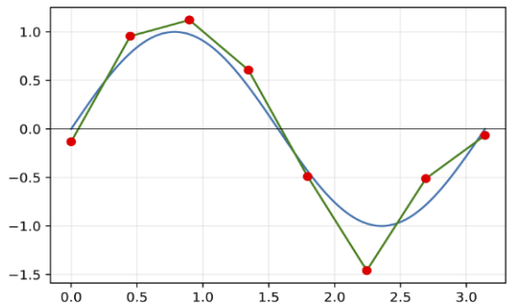
Spiky-smooth NTK activation functions

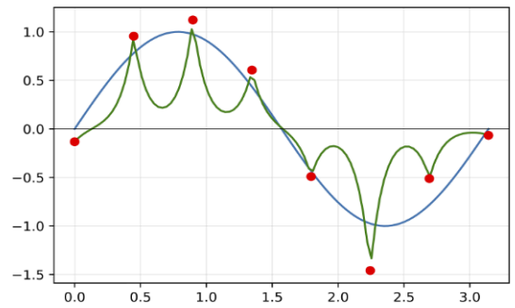


Spiky NTK activation functions



Earlier Work





- Benign overfitting mostly studied in high-dimensional limits, but what about fixed dimension d?
- (Rakhlin and Zhai, 2019) and (Buchholz, 2022):

Theorem (Buchholz, 2022):

Let k be a translation-inv. kernel $k_\gamma(x,y) = \gamma^{-d} k\left(\frac{x-y}{\gamma}\right)$ whose Fourier transform is given by

$\hat{k}_1(\xi) = (1+|\xi|^2)^{-s}$ for some $s \in \left(\frac{d}{2}, \frac{3d}{4}\right]$. (then $\|f\|_{\mathscr{H}}^2 = \int \frac{|\hat{f}(\xi)|^2}{\hat{k}(\xi)} d\xi = \|f\|_{H^s}^2$)

Let $\Omega \subseteq \mathbb{R}^d$ be a bounded open Lipschitz domain, $supp(P^X) = \bar{\Omega}$, $0 < c \leq P^X \leq C < \infty$.

The training data $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ consists of iid points $x_i \sim P^X$,

$$y_i = f^*(x_i) + \varepsilon_i, \qquad \text{where } \varepsilon_i \sim N(0,\sigma^2) \text{ iid and } f^* \in C_c^\infty(\Omega)\backslash\{0\}.$$

Then, with prob. $1 - O(1/n)$, the **min-norm. interpolant $\hat{g}_{D,\gamma}$ is inconsistent,**

$$\mathbb{E}_{x\sim P^X}\left(\hat{g}_{D,\gamma}(x) - f^*(x)\right)^2 \geq c > 0,$$

where c is independent of n and $\gamma \in (0,1)$.

Main Inconsistency Theorem

Let k be a kernel with RKHS equivalent to Sobolev space $H^s(\Omega)$, $s \in \left(\frac{d}{2}, \frac{3d}{4}\right)$.
Assume $\hat{f}_D$ estimator 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 $C > 0$: $\|\hat{f}_D\|_{H^s} \leq C\ \|\hat{f}_0\|_{H^s}$

Then, under our distrib. assump., w.h.p. 1-O(1/n) over draw of D, $\hat{f}_D$ **is inconsistent,**

$$\mathbb{E}_{x \sim P^X}\left(\hat{f}_D(x) - f^*(x)\right)^2 \geq c > 0.$$

(O) necessary: Optimally regularised ridge regression consistent with minimax optimal rates

(N) necessary: next slide

Other generalizations: $s > d/2$ $Var(y \mid x) \geq \sigma^2$ for all x, $supp(\Omega) \subseteq \mathbb{S}^d$

ReLU NTK RKHS equivalent to $H^{\frac{d+1}{2}}(\mathbb{S}^d)$.

(Chen and Xu, 2021)
(Bietti and Bach, 2021)

Corollary: Under assumptions as above,
overfitting with (deep) ReLU NTKs/NNGPs is inconsistent.

Neural Tangent Kernels

- First-order Taylor expansion of NNs:
  $f_\theta(x) = f_{\theta_t}(x) + \langle \nabla_\theta f_{\theta_t}(x), \theta - \theta_t \rangle + O(\|\theta - \theta_t\|^2)$

- Can show that gradient flow $\dot{\theta}_t = -\nabla L(\theta_t)$ is equivalent to gradient flow with empirical/finite-width NTK
  $k_t(x, x') = \langle \nabla_\theta f_{\theta_t}(x), \nabla_\theta f_{\theta_t}(x') \rangle$

- k_t is random (due to random initialization) and time-dependent

- In the infinite-width limit: $k_t = k_0$ deterministic

Explanation: Laziness at Infinite Width

Arora et al. (2019):

Neural network with activation function ϕ at layer h:

$$\text{Preactivations } f^{(h)}(x) = W^{(h)} g^{(h-1)}(x) \in \mathbb{R}^{d_h}, \quad g^{(h)}(x) = \sqrt{\frac{c_\phi}{d_h}} \phi\left(f^{(h)}(x)\right) \in \mathbb{R}^{d_h}.$$

Initialise $W^{(h)} \overset{iid}{\sim} N(0,1) \longrightarrow$ Conditioned on $f^{(h-1)}$, $f^{(h)}$ is a centred Gaussian process.

Explanation: Laziness at Infinite Width

Arora et al. (2019):

Neural network with activation function ϕ at layer h:

$$\text{Preactivations } f^{(h)}(x) = W^{(h)} g^{(h-1)}(x) \in \mathbb{R}^{d_h}, \quad g^{(h)}(x) = \sqrt{\frac{c_\phi}{d_h}} \phi\left(f^{(h)}(x)\right) \in \mathbb{R}^{d_h}.$$

Initialise $W^{(h)} \overset{iid}{\sim} N(0,1)$ ⟶ Conditioned on $f^{(h-1)}$, $f^{(h)}$ is a centred Gaussian process.

⟶ **At infinite width becomes deterministic limit with recursive definition:**

$$\Sigma^{(0)}(x, x') = x^T x', \qquad \Sigma^{(h)}(x, x') = c_\phi \mathop{\mathbb{E}}_{(u,v)\sim N\left(0, \mathscr{P}^{(h-1)}(x, x')\right)} \phi(u)\phi(v),$$

Explanation: Laziness at Infinite Width

Arora et al. (2019):

Neural network with activation function ϕ at layer h:

$$\text{Preactivations } f^{(h)}(x) = W^{(h)} g^{(h-1)}(x) \in \mathbb{R}^{d_h}, \quad g^{(h)}(x) = \sqrt{\frac{c_\phi}{d_h}} \phi\left(f^{(h)}(x)\right) \in \mathbb{R}^{d_h}.$$

Initialise $W^{(h)} \overset{iid}{\sim} N(0,1)$ ⟶ Conditioned on $f^{(h-1)}$, $f^{(h)}$ is a centred Gaussian process.

⟶ **At infinite width becomes deterministic limit with recursive definition:**

$$\Sigma^{(0)}(x, x') = x^T x', \qquad \Sigma^{(h)}(x, x') = c_\phi \mathop{\mathbb{E}}_{(u,v)\sim N\left(0, \mathscr{P}^{(h-1)}(x,x')\right)} \phi(u)\phi(v),$$

Where marginal distribution of (x,x') at layer h: $\mathscr{P}^{(h-1)}(x, x') := \begin{pmatrix} \Sigma^{(h-1)}(x,x) & \Sigma^{(h-1)}(x,x') \\ \Sigma^{(h-1)}(x',x) & \Sigma^{(h-1)}(x',x') \end{pmatrix}$

Explanation: Laziness at Infinite Width

Arora et al. (2019):

Neural network with activation function ϕ at layer h:

$$\text{Preactivations } f^{(h)}(x) = W^{(h)} g^{(h-1)}(x) \in \mathbb{R}^{d_h}, \quad g^{(h)}(x) = \sqrt{\frac{c_\phi}{d_h}} \phi\left(f^{(h)}(x)\right) \in \mathbb{R}^{d_h}.$$

Initialise $W^{(h)} \overset{iid}{\sim} N(0,1)$ ⟶ Conditioned on $f^{(h-1)}$, $f^{(h)}$ is a centred Gaussian process.

⟶ **At infinite width becomes deterministic limit with recursive definition:**

$$\Sigma^{(0)}(x,x') = x^T x', \qquad \Sigma^{(h)}(x,x') = c_\phi \mathop{\mathbb{E}}_{(u,v)\sim N\left(0,\mathscr{P}^{(h-1)}(x,x')\right)} \phi(u)\phi(v),$$

Where marginal distribution of (x,x') at layer h: $\mathscr{P}^{(h-1)}(x,x') := \begin{pmatrix} \Sigma^{(h-1)}(x,x) & \Sigma^{(h-1)}(x,x') \\ \Sigma^{(h-1)}(x',x) & \Sigma^{(h-1)}(x',x') \end{pmatrix}$

Lee et al. (2019): $k_{NTK}^{(h)}(x,x') = \Sigma^{(h)}(x,x') + k_{NTK}^{(h-1)}(x,x') \cdot \mathop{\mathbb{E}}_{(u,v)\sim N\left(0,\mathscr{P}^{(h-1)}(x,x')\right)} \phi'(u)\phi'(v)\,.$