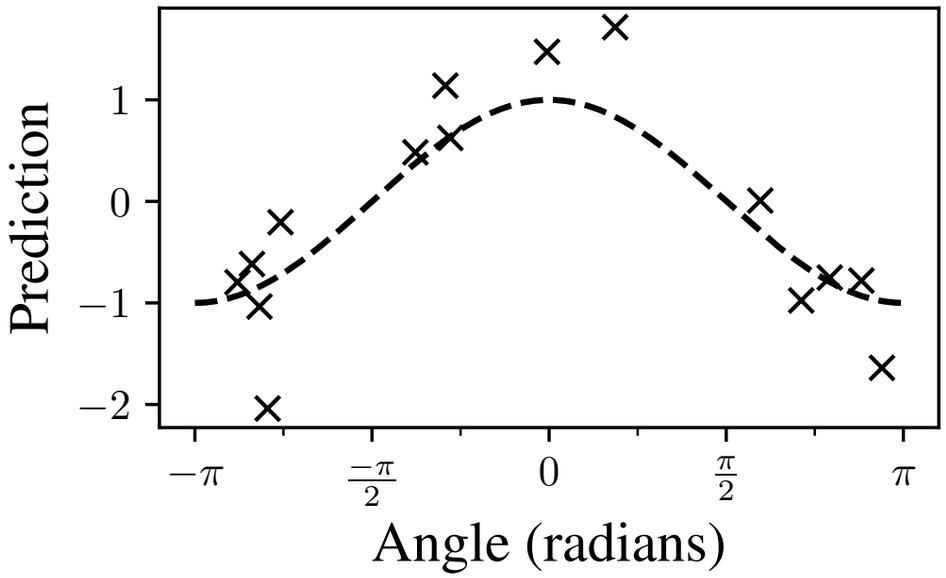
Mind the spikes: Benign overfitting of kernels and neural networks in fixed dimension

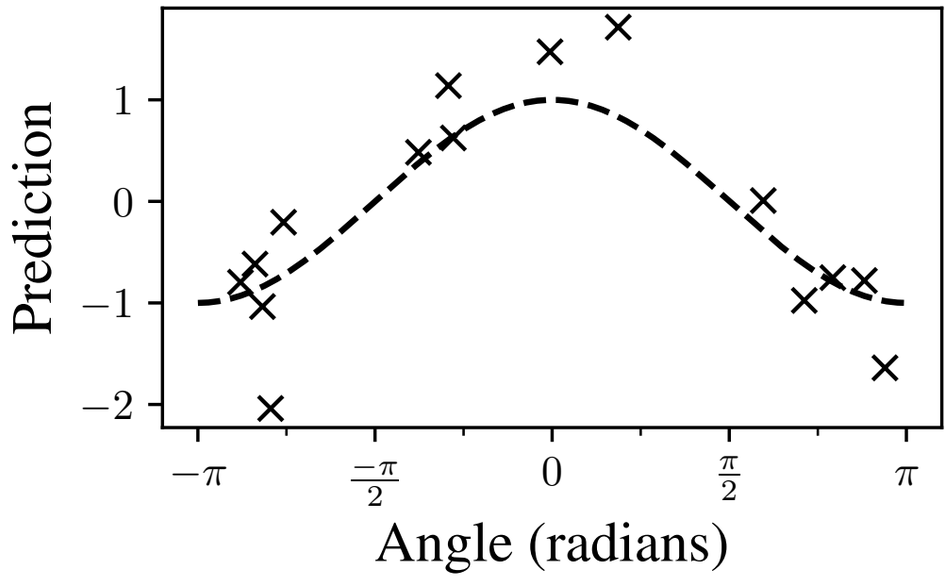
Moritz Haas*, David Holzmüller*, Ulrike von Luxburg, Ingo Steinwart

* denotes equal contribution.

MvL6 10-minute talk

Setting: Fixed continuous distributions

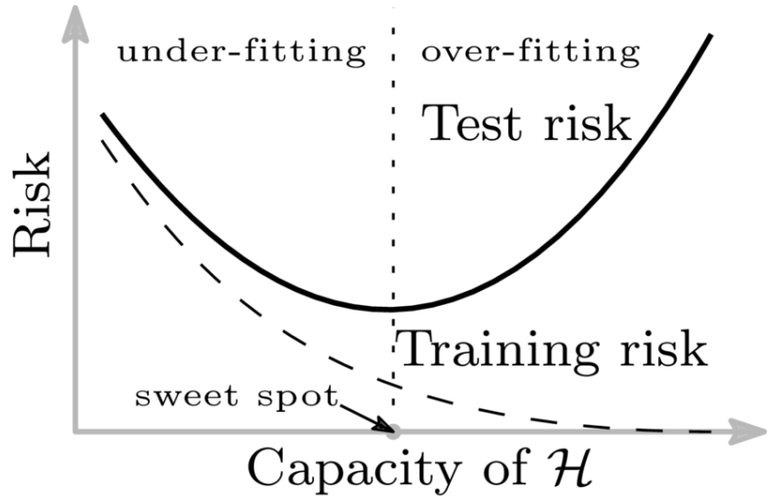


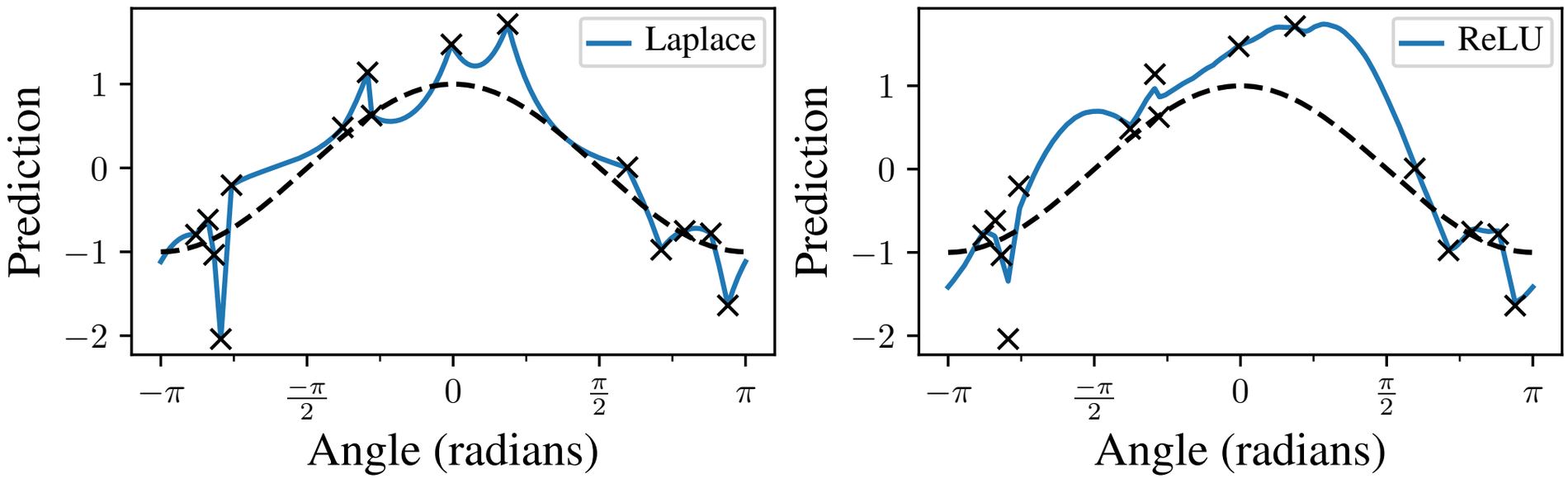


$x_i \overset{iid}{\sim} P_X$ $\quad y_i = f^*(x_i) + \varepsilon_i$ $\quad \varepsilon_i \overset{iid}{\sim} N(0,\sigma),\ \sigma > 0$ $\quad f^* \in C_c^\infty(\Omega)$

$0 < c \leq P^X \leq C < \infty$

Traditional rationale: Do not overfit!





"Min-norm interpolator inconsistent"

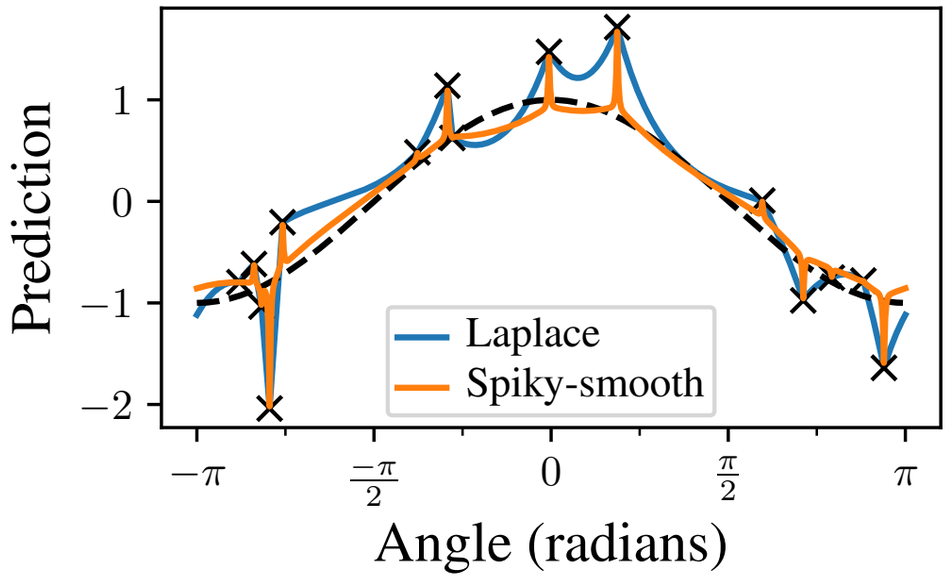
(Rakhlin and Zhai, 2019)

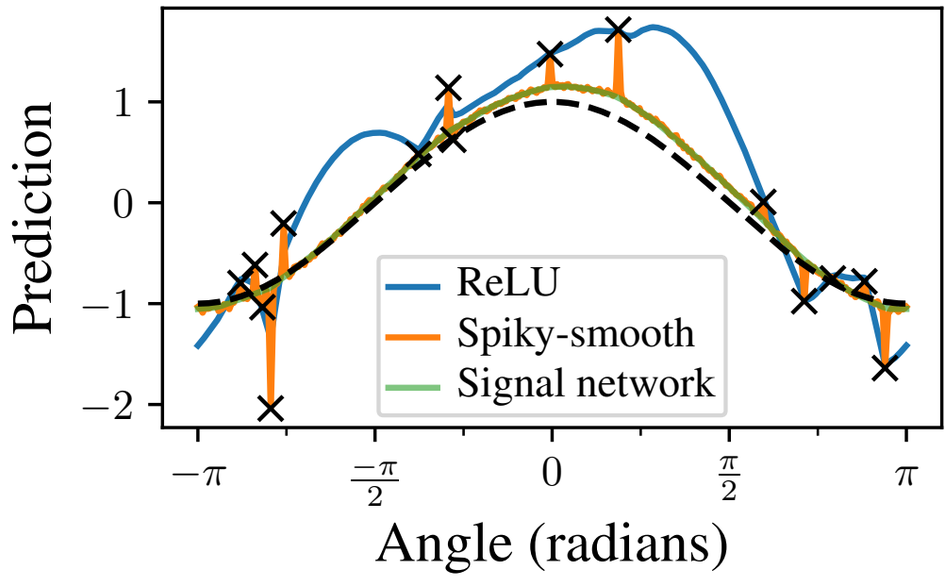
(Buchholz, 2022)

Overfitting with NNGPs/ NTKs also inconsistent?

What about other estimators? Is benign overfitting with kernels/ neural networks in fixed dimension impossible?

Our contributions:





Any overfitting estimator with RKHS-norm comparable to min-norm interpol is inconsistent!

Min-norm interpolator of NNGPs/NTKs is inconsistent!

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

Main Theorem

Let k be a kernel with RKHS equivalent to Sobolev space $H^s(\Omega)$, s>d/2.

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 $C > 0$: $\|\hat{f}_D\|_{H^s} \leq C\, \|\hat{g}_D\|_{H^s}$

Main Theorem

Let k be a kernel with RKHS equivalent to Sobolev space $H^s(\Omega)$, s>d/2.

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.

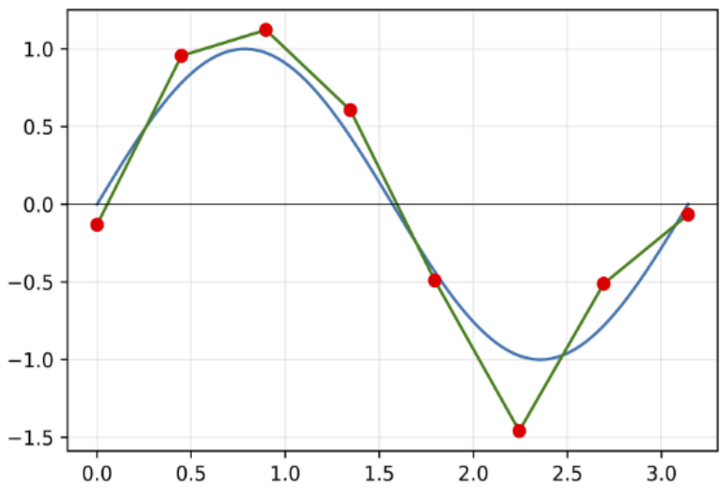
(N) norm-bounded: Exists $C > 0:$ $\|\hat{f}_D\|_{H^s} \leq C\ \|\hat{g}_D\|_{H^s}$

Then, under same distributional assumptions as Buchholz (2022), w.h.p. $\hat{f}_D$ **is inconsistent**, i.e. w.p. 1-O(1/n) over the draw of D,

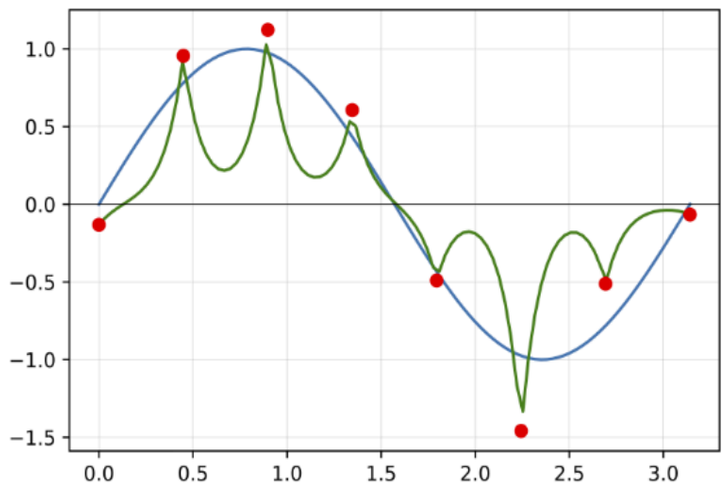
$$\mathbb{E}_{x\sim P^X}\left(\hat{f}_D(x) - f^*(x)\right)^2 \geq c > 0.$$

Proof idea:

Large bandwidths:



Small bandwidths:



Main Theorem

Let k be a kernel with RKHS equivalent to Sobolev space $H^s(\Omega)$, s>d/2.

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 $C > 0:$ $\|\hat{f}_D\|_{H^s} \leq C\ \|\hat{g}_D\|_{H^s}$

Then, under same distributional assumptions as Buchholz (2022), w.h.p. $\hat{f}_D$ **is inconsistent**, i.e. w.p. 1-O(1/n) over the draw of D,

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

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

(N) necessary: next slide

Main Theorem

Let k be a kernel with RKHS equivalent to Sobolev space $H^s(\Omega)$, s>d/2.

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 $C > 0:$ $\|\hat{f}_D\|_{H^s} \leq C\ \|\hat{g}_D\|_{H^s}$

Then, under same distributional assumptions as Buchholz (2022), w.h.p. $\hat{f}_D$ **is inconsistent**, i.e. w.p. 1-O(1/n) over the draw of D,

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

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

(N) necessary: next slide

Other generalizations: $s > d/2$ $Var(y\,|\,x) \geq \sigma^2$ 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.

Achieving benign overfitting with spiky-smooth kernel sequences

Benign overfitting in linear and kernel regression in high dimension is understood:

(Bartlett et al., 2021)

generalises well

min-norm interpol. = Smooth + spiky

interpolates training data and harmless for generalisation

Achieving benign overfitting with spiky-smooth kernel sequences

Benign overfitting in linear and kernel regression in high dimension is understood:

(Bartlett et al., 2021)

generalises well

min-norm interpol. = Smooth + spiky

interpolates training data and harmless for generalisation

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

→ In fixed d, need quickly diverging derivatives to interpolate noise harmlessly.

Achieving benign overfitting with spiky-smooth kernel sequences

Benign overfitting in linear and kernel regression in high dimension is understood:

(Bartlett et al., 2021)

generalises well

min-norm interpol. = Smooth + spiky

interpolates training data and harmless for generalisation

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

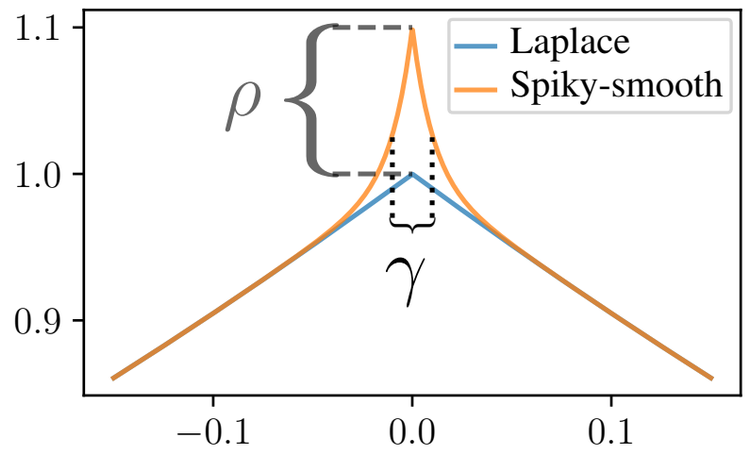
→ In fixed d, need quickly diverging derivatives to interpolate noise harmlessly.

→ **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 min-norm interplant of $k_{\rho,\gamma}$ is consistent.**

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

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

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

$$\kappa(x) = \sum_{i=0}^{\infty} b_i x^i \quad \leftrightarrow$$

$$\omega_{NNGP}(x) = \sum_{i=0}^{\infty} s_i \sqrt{b_i}\, h_i(x),$$

(Simon et al., 2022)

$$\omega_{NTK}(x) = \sum_{i=0}^{\infty} s_i \sqrt{\frac{b_i}{i+1}}\, h_i(x),$$

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 layers

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

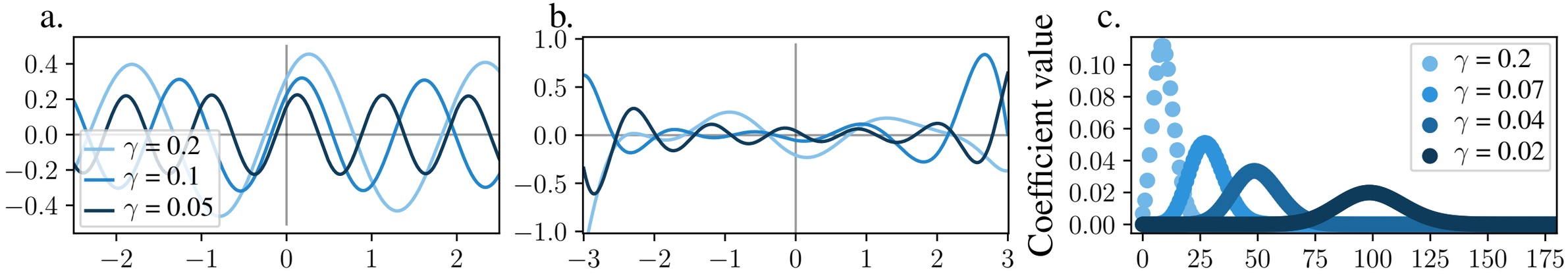
$$\kappa(x) = \sum_{i=0}^{\infty} b_i x^i \quad \leftrightarrow$$

$$\omega_{NNGP}(x) = \sum_{i=0}^{\infty} s_i \sqrt{b_i}\, h_i(x),$$

(Simon et al., 2022)

$$\omega_{NTK}(x) = \sum_{i=0}^{\infty} s_i \sqrt{\frac{b_i}{i+1}}\, h_i(x),$$

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



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

s_i random

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

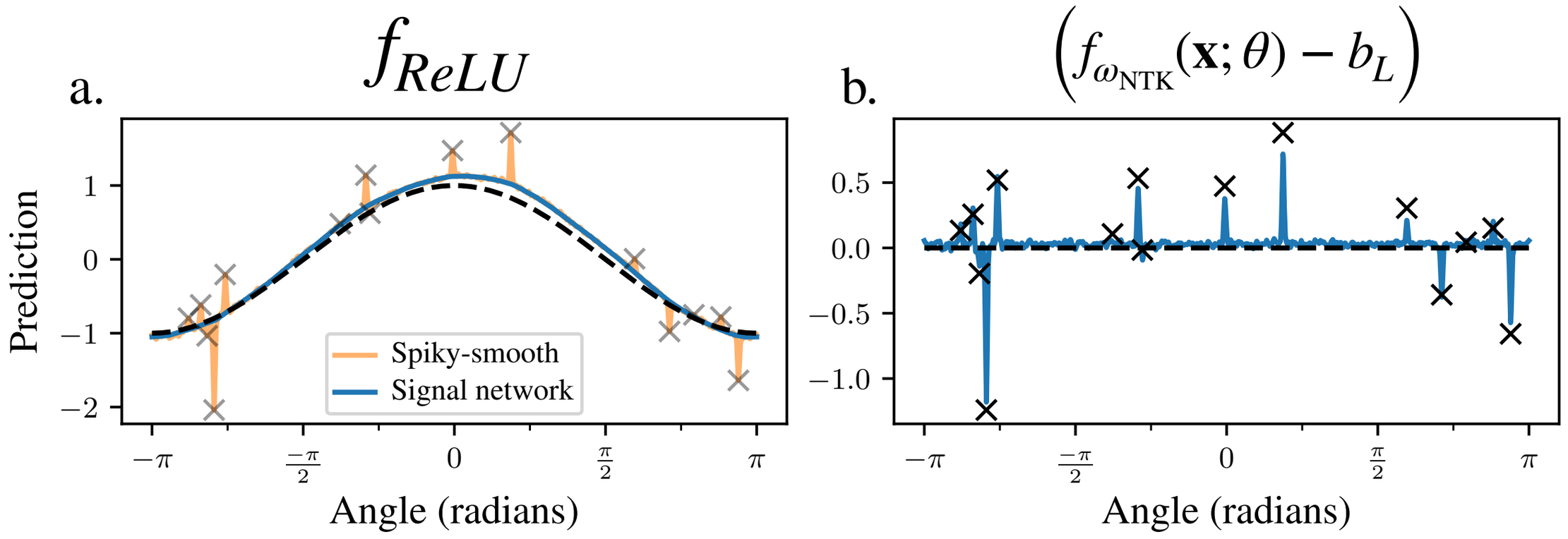
Approximated by

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

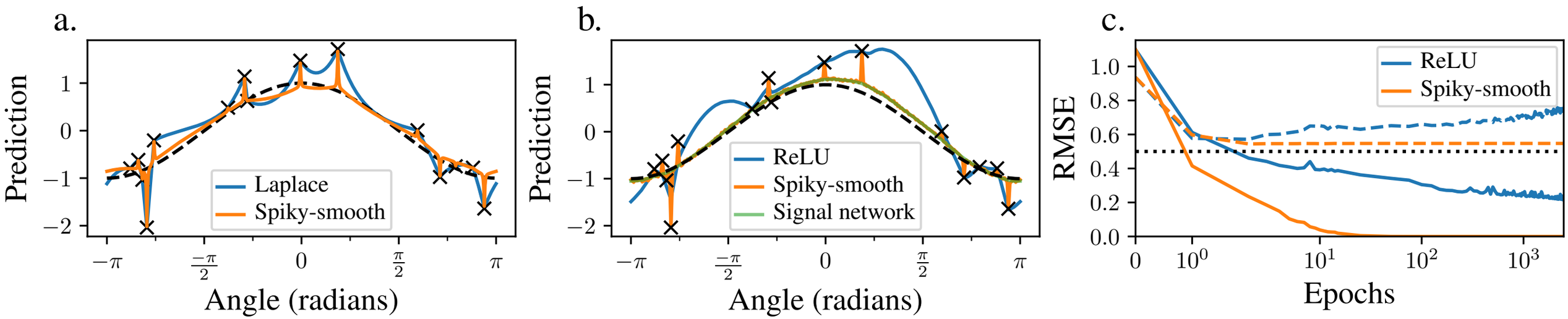
Bonus: Disentangling signal from spike component

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

Activation function *Neural network decomposition*



Conclusions



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

How can we adapt feature learning neural networks to benignly overfit in 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.

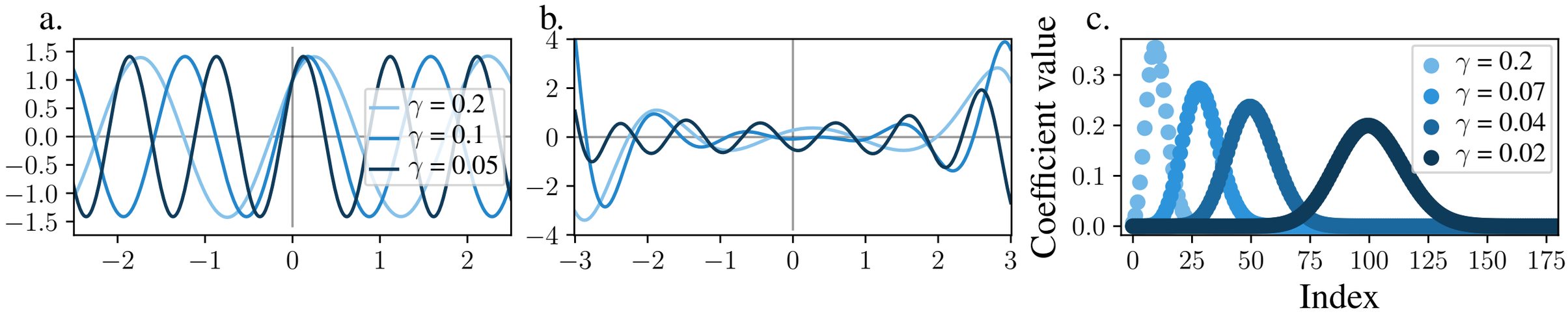
N. Mallinar, J. Simon, A. Abedsoltan, P. Pandit, M. Belkin, P. Nakkiran. **Benign, tempered, or catastrophic: Toward a refined taxonomy of overfitting.** NeurIPS 2022.

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.

Spiky-smooth NNGP Activation Functions

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

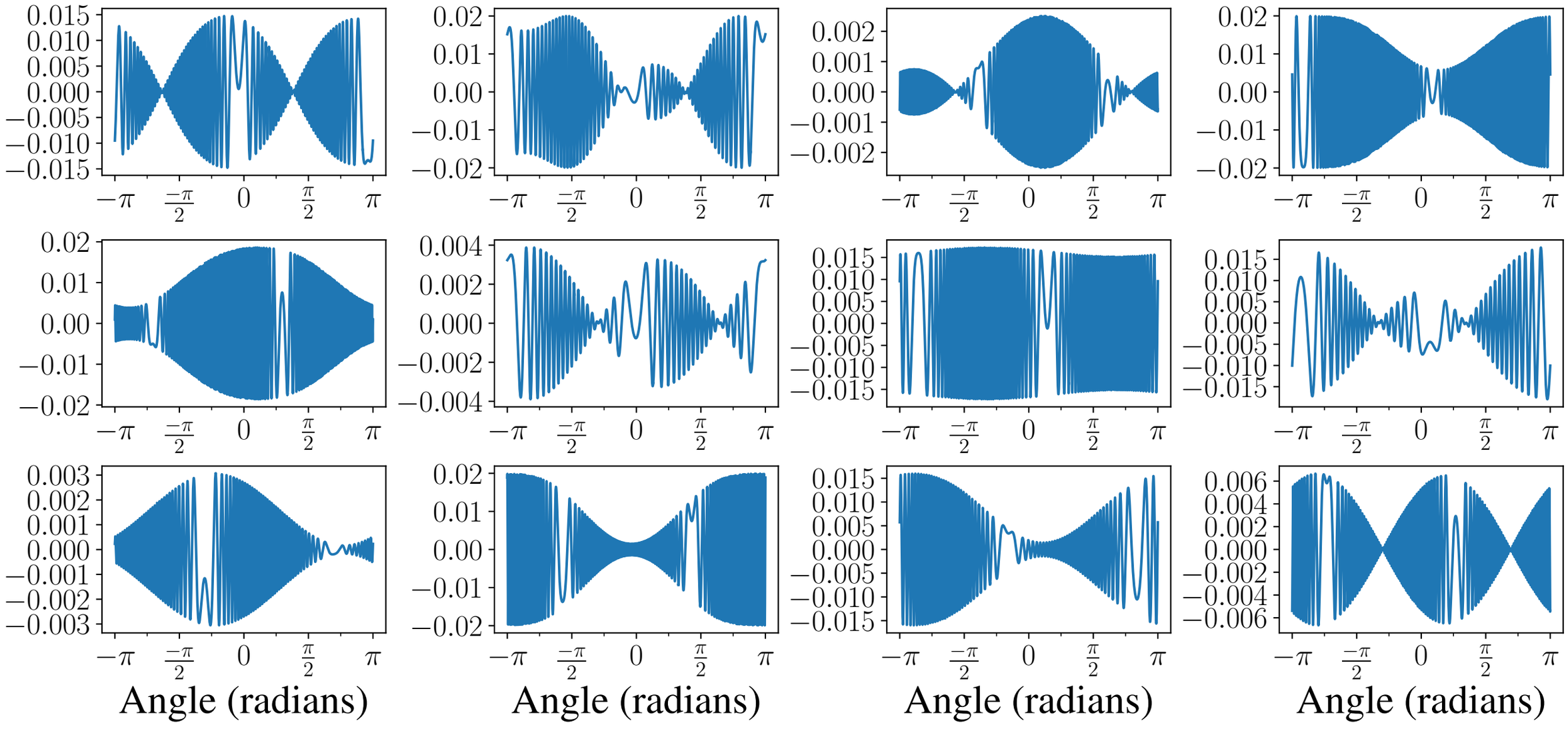


$$\|\omega_{NNGP}^{Gauss}\|_{L_2(N(0,1))} = 1$$

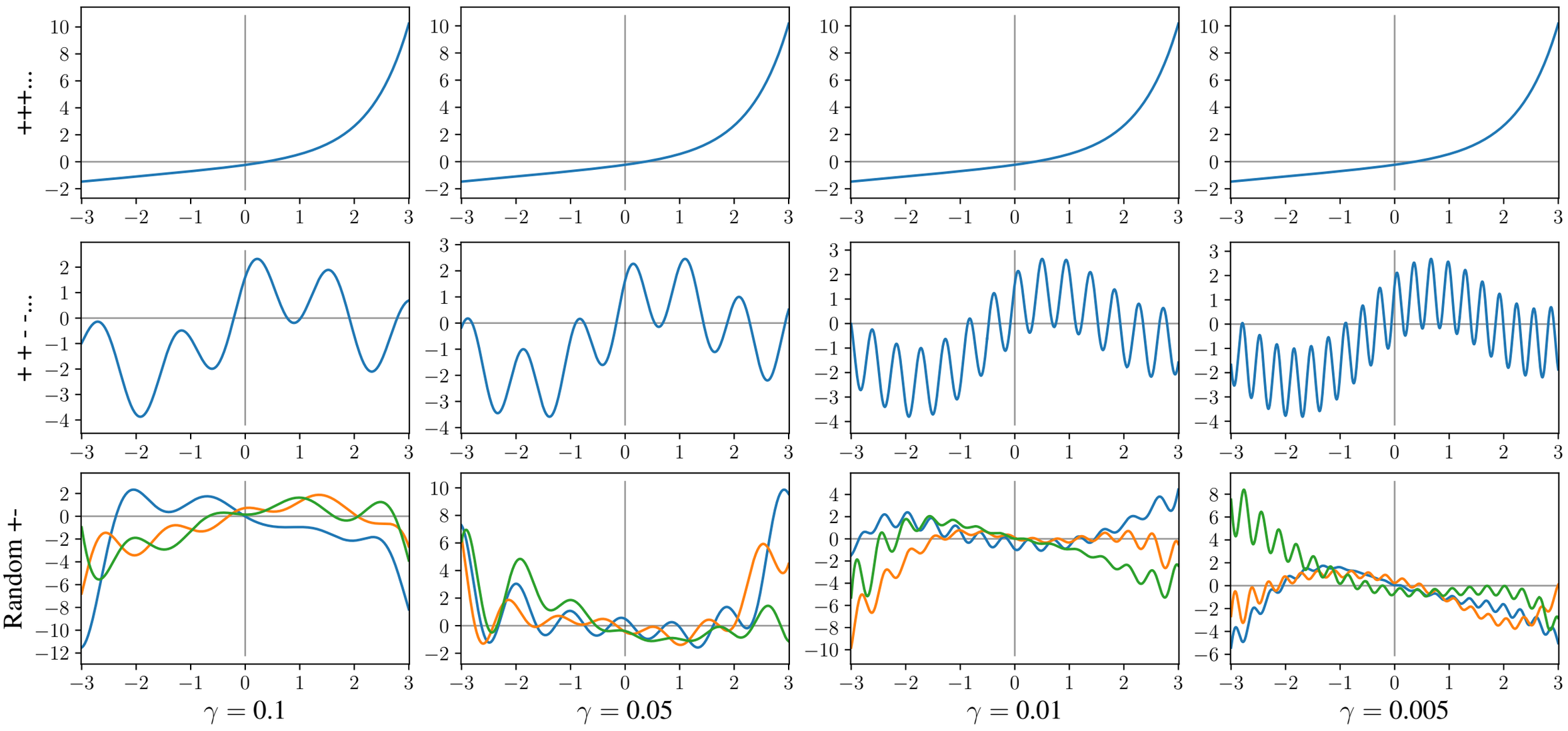
L_2-norm and amplitude invariant to γ.

Constructive and destructive interference?

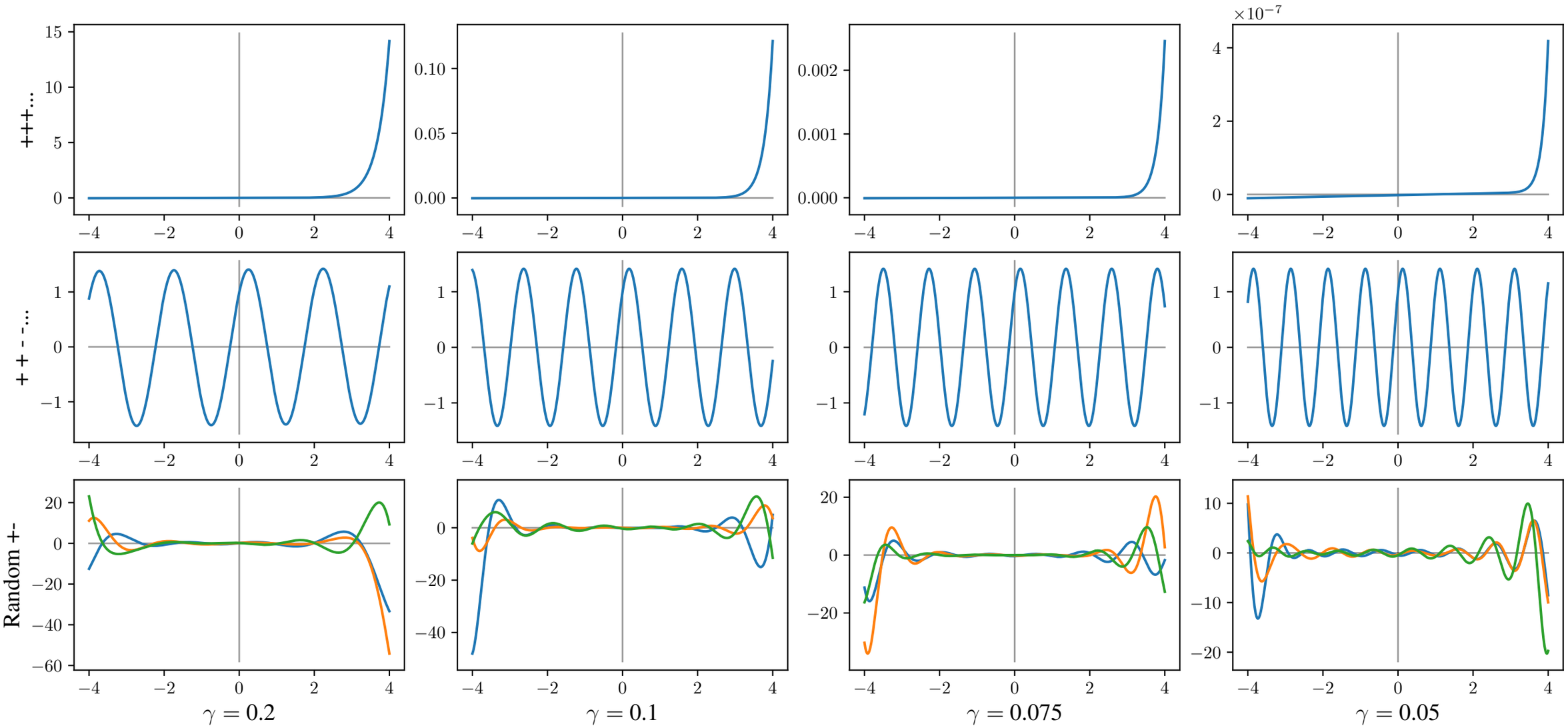
Functions learned by 12 random hidden layer neurons of the spike component network:



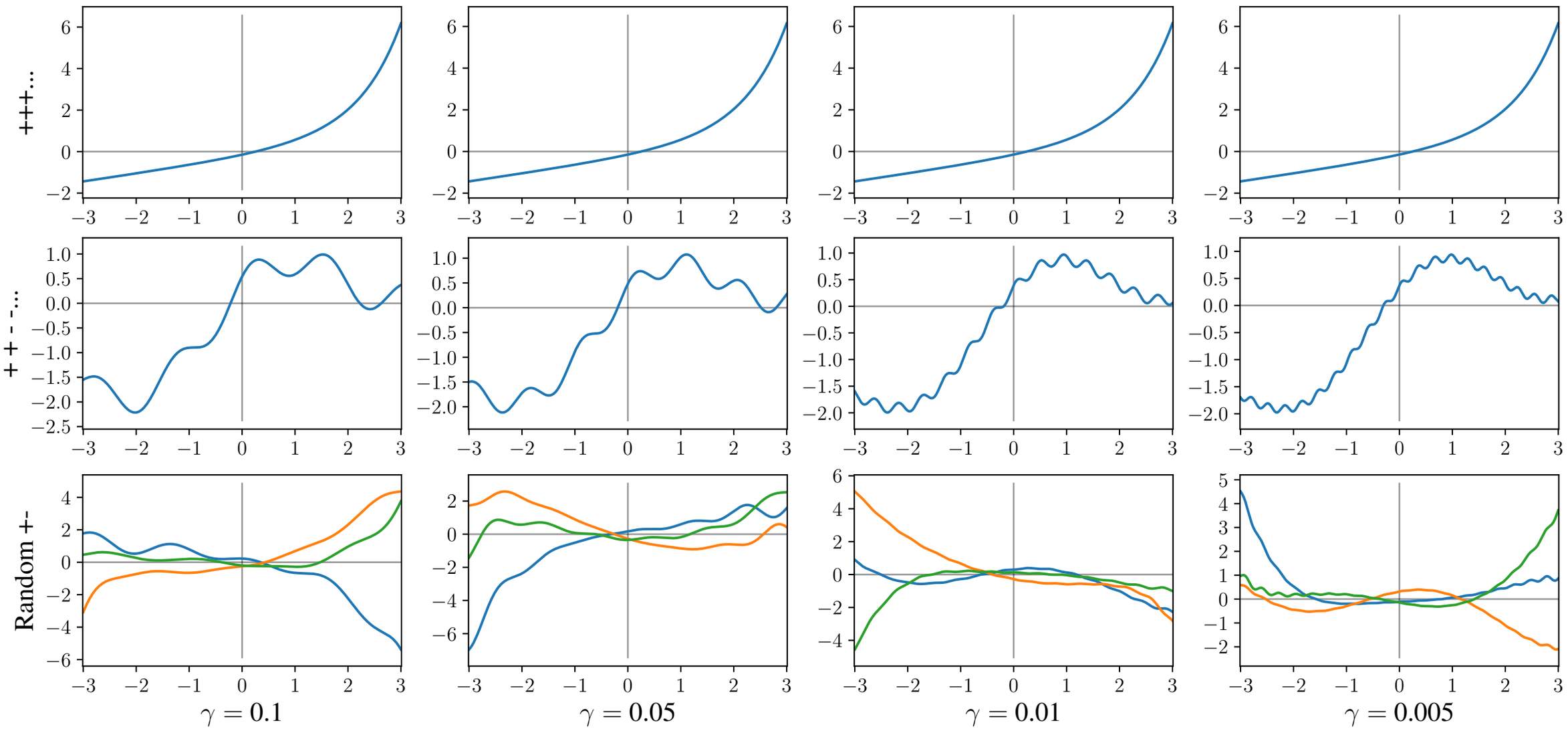
Spiky-smooth NNGP activation functions



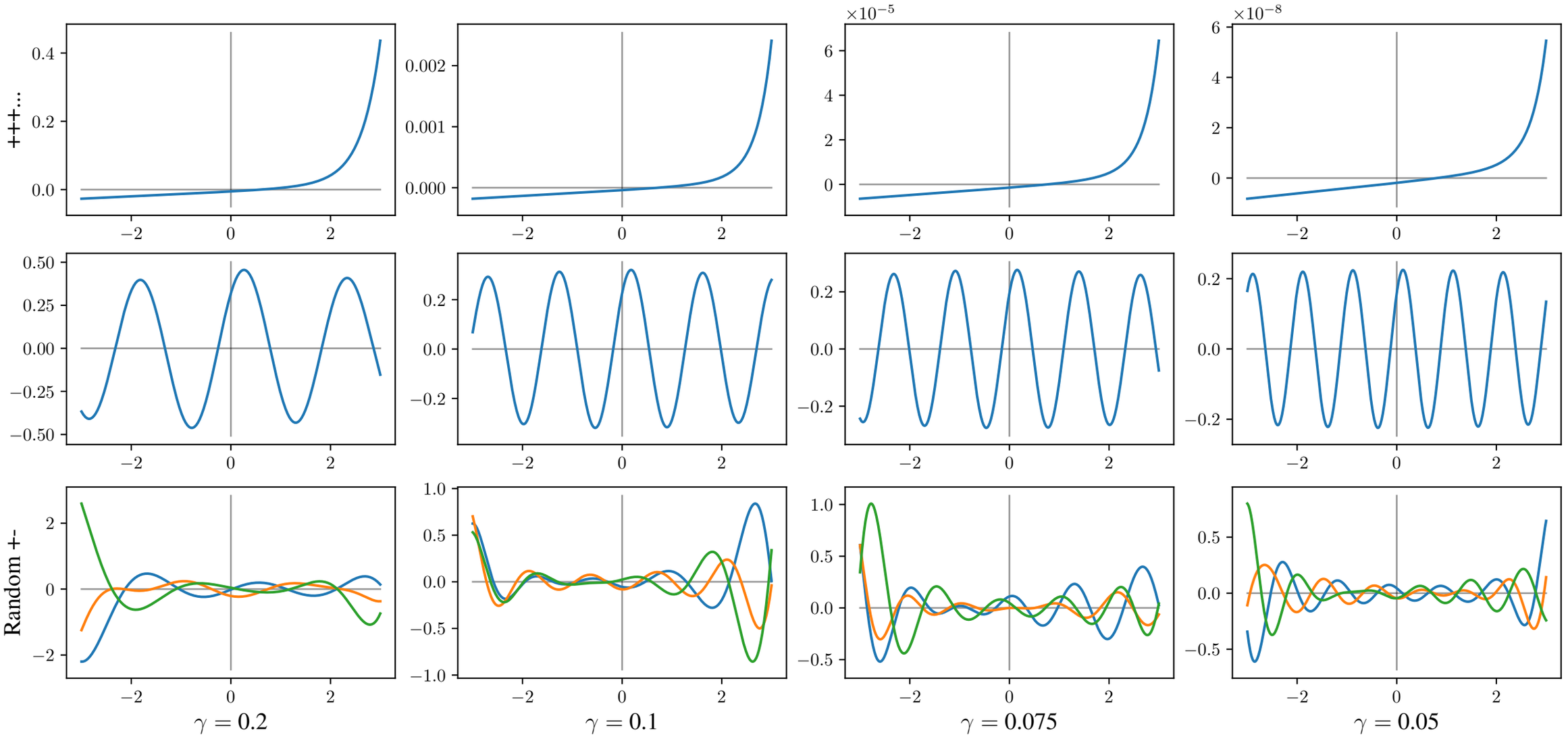
Spiky NNGP activation functions



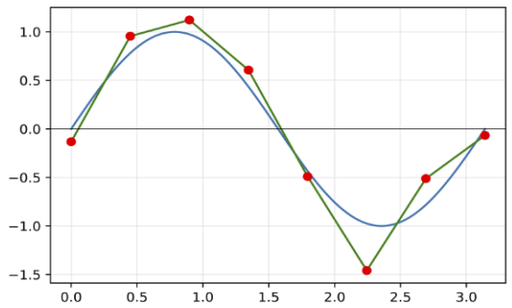
Spiky-smooth NTK activation functions

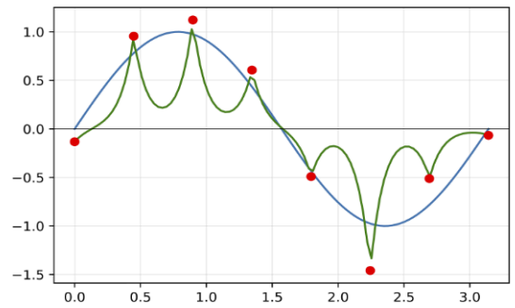


Spiky NTK activation functions



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 $\gamma \in (0,1)$.

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