Efficient Anti-Symmetrization of a Neural Network Layer by Taming the Sign Problem

Nilin Abrahamsen * [1,2] and Lin Lin † [1,3]

[1]Department of Mathematics, University of California, Berkeley, USA
[2]The Simons Institute for the Theory of Computing, University of California, Berkeley, USA
[3]Lawrence Berkeley National Laboratory, University of California, Berkeley, USA

Abstract. Explicit antisymmetrization of a neural network is a potential candidate for a universal function approximator for generic antisymmetric functions, which are ubiquitous in quantum physics. However, this procedure is a priori factorially costly to implement, making it impractical for large numbers of particles. The strategy also suffers from a sign problem. Namely, due to near-exact cancellation of positive and negative contributions, the magnitude of the antisymmetrized function may be significantly smaller than before anti-symmetrization. We show that the anti-symmetric projection of a two-layer neural network can be evaluated efficiently, opening the door to using a generic antisymmetric layer as a building block in anti-symmetric neural network Ansatzes. This approximation is effective when the sign problem is controlled, and we show that this property depends crucially the choice of activation function under standard Xavier/He initialization methods. As a consequence, using a smooth activation function requires re-scaling of the neural network weights compared to standard initializations.

1 Introduction

Simulation of quantum chemistry from first principles depends on the accurate modeling of fermionic system comprised of the electrons. The Pauli exclusion principle dictates that fermionic wavefunctions must be antisymmetric with respect to particle exchange. This antisymmetry poses challenges; for instance, as the number of fermions increases, the effective parameterization of such wavefunctions becomes exceedingly complex for many systems. The antisymmetry condition also results in near-exact cancellation between positive and negative contributions when computing observables. This leads to the so-called fermionic sign problem (FSP), which was originally discovered in quantum Monte Carlo (QMC) simulations [2, 6, 12].

Over the last decade, the scientific community has witnessed a surge in the development of methods employing neural networks (NNs) as universal function approximators. This surge is due to advancements in software tools, hardware capabilities, and algorith-

---

*Corresponding author. `nilin@berkeley.edu`
†`linlin@math.berkeley.edu`

mic improvements. These developments have had a significant impact on the modeling of fermionic systems [3–5, 8, 9, 13].

However, constructing a universal NN representation for antisymmetric functions that does not suffer from the curse of dimensionality is still an open question. In the absence of symmetry constraints, even a simple structure such as a two-layer NN can act as a universal function approximator. In theory, one could explicitly antisymmetrize such a two-layer NN to parameterize universal antisymmetric functions. Such an explicitly antisymmetrized NN structure has been recently studied in QMC calculations, which can yield effectively the exact ground state energy for small atoms and molecules [5]. However, the computational cost of this antisymmetrization procedure appears a priori to grow factorially with the system size.

In this paper, we give a procedure to efficiently evaluate the explicit anti-symmetrization of a two-layer neural network using a quadrature procedure. This is surprising due to the factorially many terms in the definition of the anti-symmetrization. For this statement to be meaningful we require that the sign problem is controlled, meaning that the anti-symmetrization does not make the original function vanish due to cancellations. We demonstrate that with the standard Xavier/He initialization, the sign problem is controlled when the activation function in the neural network is rough. Examples of a rough (respectively smooth) activation function in the ReLU (respectively sigmoid). Alternatively, this statement implies that to avoid the sign problem with the sigmoid activation, the weights in the first layer need to be asymptotically larger than the standard Xavier/He initializations.

Among all activation functions, the exponential activation function (real or complex) plays a special role in our analysis. This is because antisymmetrizing a two-layer NN with an exponential activation function gives rise to a determinant (called a Slater determinant), which can be evaluated in polynomial time. By exploring the Fourier representation of a (rough) activation function, we can approximately express the explicitly antisymmetrized two-layer NN as a linear combination of polynomially (with respect to the system size and inverse precision) many Slater determinants. This overcomes the factorial scaling barrier, and gives rise to a polynomial-time algorithm for approximate evaluation of antisymmetrized two-layer neural networks (Theorem 3.3).

1.1 Related work

The representation of anti-symmetric functions is extensively studied in physics, where a widely used class of Ansatzes for anti-symmetric functions takes the form of a sum of Slater determinants. Slater determinants can span a dense subset of the anti-symmetric space but the representation is very inefficient. Indeed, even in the case of a finite single-particle state space $|\Omega| = \mathcal{O}(n)$ we would require $\binom{|\Omega|}{n}$ Slater determinants to span the anti-symmetric space. [17] finds certain anti-symmetric functions that cannot be efficiently approximated using a simple sum of Slater determinants, but can be effectively expressed using a more complex Ansatz called the Slater-Jastrow form.

In the machine learning literature there is a rich body of works related to permutation-invariant data, i.e. when the input data is a set [11, 14–16]. But the literature on anti-

symmetrized neural networks is sparse. [1] gave approximation bounds for the class of anti-symmetric functions in the Barron space, which can be viewed as the set of functions that can be expressed as infinite two-layer neural networks.

2 Problem setting

The wave function $\psi(x_1, \ldots, x_n)$ of a system of n indistinguishable particles in a d-dimensional space, $d = 1, 2, 3$, satisfies permutation symmetry of $|\psi|$ under interchange of the n inputs $x_i \in \mathbb{R}^d$. Fermions are indistinguishable particles which satisfy the Pauli exclusion principle and correspond to an antisymmetric wave function ψ. For a permutation $\pi \in S_n$ with sign $(-1)^\pi$, we have $\pi(\psi) = (-1)^\pi \psi$, where we have defined $\pi(\psi) : \mathbb{R}^{nd} \mapsto \mathbb{C}$ by

$$\pi(\psi)(x) := \psi(x_{\pi(1)}, \ldots, x_{\pi(n)}) \quad \text{for} \quad x \in \mathbb{R}^{nd}.$$

For any $f : \mathbb{R}^{nd} \to \mathbb{C}$ we can define its explicit antisymmetrization

$$\mathcal{A}f = \frac{1}{\sqrt{n!}} \sum_{\pi \in S_n} (-1)^\pi \pi(f). \tag{2.1}$$

As will be shown later, the prefactor $1/\sqrt{n!}$ is the natural scaling in the antisymmetrization process.

A function defined on $\mathbb{R}^d$ is called a single-particle function. Let ρ be a fast-decaying probability density on $\mathbb{R}^d$ and let $\rho_n = \rho^{\otimes n}$ be a product of single-particle densities. For simplicity of analytic computation, we may take ρ to be the density of the standard Gaussian $\mathcal{N}(0, I_d)$ (called a Gaussian envelope). We represent an n-particle fermionic wave function as

$$\psi = \sqrt{\rho_n} \odot \mathcal{A}f = \mathcal{A}(\sqrt{\rho_n} \odot f), \tag{2.2}$$

where $\odot$ denotes multiplication of function values. The wave function should be normalized as $\|\psi\|^2 = 1$, where $\|\cdot\|$ is the L^2-norm on $\mathbb{R}^{nd}$. Eq. (2.2) implies $\|\psi\| = \|\mathcal{A}f\|_\rho$ where $\|\cdot\|_\rho$ is the norm induced by the inner product

$$\langle f \,|\, g \rangle_\rho = \int \bar{f}(x) g(x) d\rho_n(x).$$

If $f = f_1 \otimes \cdots \otimes f_n$ is a product of single-particle functions, then so is $\phi = \phi_1 \otimes \cdots \otimes \phi_n$, where $\phi_i = \sqrt{\rho} \odot f_i$. In this case $\psi(x) = (\mathcal{A}\phi)(x)$ is a determinant (called the Slater determinant) denoted by $\phi_1 \wedge \cdots \wedge \phi_n$ and defined by

$$(\phi_1 \wedge \cdots \wedge \phi_n)(x_1, \ldots, x_n) = \frac{1}{\sqrt{n!}} \det\left[(\phi_i(x_j))_{ij}\right].$$

The normalization in Eq. (2.1) is such that if ϕ_i are orthonormal functions on $L^2(\mathbb{R}^d)$ then ψ is normalized by Pythagoras' theorem, $\|\psi\| = \|\mathcal{A}f\|_\rho = 1$.

By letting f range over a universal class of functions on $\mathbb{R}^{nd}$ we obtain a universal class of antisymmetric functions $\psi = \sqrt{\rho_n} \odot \mathcal{A}f$ which are not in general normalized. However, this approach has two important drawbacks a priori:

1. The procedure of normalizing ψ becomes numerically unstable if cancellations in (2.1) cause the magnitudes of $\mathcal{A}f$ to be too small compared to f. This can be viewed as a manifestation of the fermionic sign problem in this setting.

2. The sum (2.1) has $n!$ terms, making it in general intractable to evaluate the sum for all but small values of n.

We consider the case when f is given by a two-layer NN

$$f_{W,a,b}(x) = \sum_{k=1}^{m} a_k \tau\big(w^{(k)} \cdot x + b_k\big), \tag{2.3}$$

where $\tau : \mathbb{R} \to \mathbb{R}$ is some activation function, $w^{(k)} \in \mathbb{R}^{nd}$, and $b_k, a_k \in \mathbb{R}$ for each $k = 1, \ldots, m$.

To illustrate the sign problem, Fig. 2.1 shows the magnitude $\|\mathcal{A}f\|_\rho^2$ for four activation functions exp, tanh, ReLU ($\tau_{\text{ReLU}}(y) = \max\{0, y\}$), and the Heaviside step function ($\tau_{\text{HS}}(y) = 1_{y>0}$). As the system size n increases, the norm decreases with respect to n for all activation functions. However, the decay rate depends on the smoothness of the activation function. The deterioration of the sign problem is much more severe for smooth activation functions (exp, tanh) than for rough activation functions (ReLU, Heaviside). We aim to quantify this effect and investigate more precisely how the magnitude of $\mathcal{A}\tau_w$ depends on τ.
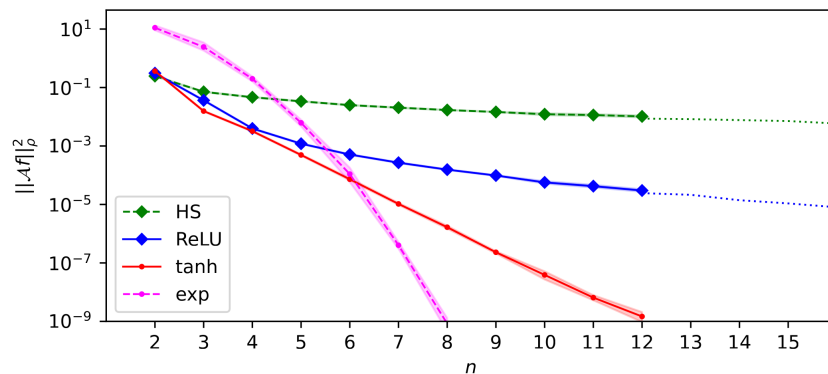

Figure 2.1: Log-plot of $\mathbb{E}[\|\mathcal{A}f\|_\rho^2]$ as a function of n for different activation functions τ: exp (magenta), tanh (red), ReLU (blue), and Heaviside step function (green). The weights are sampled from the Xavier initialization (Definition 3.4) with $d = 3$, $m = nd$. Shaded areas represent 90% confidence regions. Values for $n \leq 12$ are computed by direct antisymmetrization, and dotted lines on the right are computed from (5.6).

3 Main results

We state our results in terms of the Fourier transform $\hat{\tau}$ of the activation function. A typical activation function does not have finite integral over $\mathbb{R}$, so its Fourier transform is not defined as a convergent integral but rather in the sense of tempered distributions [10].

We will not need the precise definition of $\widehat{\tau}$ but only that it satisfies the Fourier inversion formula in the sense that for $0 < \epsilon < 1$,

$$\tau(y) = \frac{1}{\sqrt{2\pi}} \int_{|\theta|>\epsilon} \widehat{\tau}(\theta)e^{i\theta y}d\theta + p(y) + C_\epsilon + \mathcal{O}_{\epsilon\to 0}(\epsilon g(y)), \tag{3.1}$$

where p is a polynomial of bounded degree and g is bounded by a polynomial. In particular we will be able to take $p, C_\epsilon \equiv 0, g(y) = |y|$ for $\tau = \tanh$ and $p(y) = y/2$, $C_\epsilon = 1/(\pi\epsilon), g(y) = y^2$ for $\tau = \text{ReLU}$. The integral in (3.1) converges for these activation functions since $\int_{|\theta|>\epsilon} |\widehat{\tau}(\theta)|d\theta < \infty$:

Table 3.1: Fourier transforms of different activation functions.

$\tau(y)$	$\widehat{\tau}(\theta)$	Fourier tail decay K
$\text{ReLU}(y)$	$\frac{-1}{\sqrt{2\pi}\cdot\theta^2} + \sqrt{\pi/2}i\delta'(\theta)$	3 (rough)
$\tanh(y)$	$\frac{-i\sqrt{\pi/2}}{\sinh(\pi\theta/2)}$	∞ (smooth)

Consider the decomposition the an activation function into low- and high-frequency parts as follows.

Definition 3.1. *For $\tau : \mathbb{R} \to \mathbb{C}$ define its high-pass $\tau^{\text{HP}(t)}$ at threshold $t > 0$ by*

$$\tau^{\text{HP}(t)}(y) = \frac{1}{\sqrt{2\pi}} \int_{|\theta|>t} \widehat{\tau}(\theta)e^{i\theta y}d\theta. \tag{3.2}$$

Define its low-pass as the remainder $\tau^{\text{LP}(t)} = \tau - \tau^{\text{HP}(t)}$.

Eq. (3.1) says that $\tau^{\text{LP}(\epsilon)} = p + C_\epsilon + \mathcal{O}(\epsilon g)$ as $\epsilon \to 0$.

Definition 3.2. *For an activation function τ define its frequency tail $\mathfrak{T}_{\widehat{\tau}} : (0,\infty) \to [0,\infty)$ by*

$$\mathfrak{T}_{\widehat{\tau}}(t) = \int_{|\theta|\geq t} |\widehat{\tau}(\theta)|^2 d\theta. \tag{3.3}$$

We define the tail decay $K \geq 0$ of $\widehat{\tau}$ as the largest K such that $\mathfrak{T}_{\widehat{\tau}}(t) = \mathcal{O}(t^{-K})$ as $t \to \infty$. More precisely,

$$K = \limsup_{t\to\infty} \frac{-\log \mathfrak{T}_{\widehat{\tau}}(t)}{\log t}.$$

Definition 3.3 (Smooth and Rough Activation Functions)**.**

1. *A function $\tau : \mathbb{R} \to \mathbb{C}$ is smooth if its Fourier transform $\widehat{\tau}$ has tail decay ∞, i.e. if $\mathfrak{T}_{\widehat{\tau}}(t) = t^{-\omega(1)}$ decays faster than polynomially as $t \to \infty$. In particular, any activation function with $\widehat{\tau}(\theta) = \theta^{-\omega(1)}$ is smooth.*

2. *τ is rough if there exists $k > 1$ and non-zero constants $z_+, z_- \in \mathbb{C}$ such that $\widehat{\tau}(\theta) = z_+\Theta(|\theta|^{-k})$[1] as $\theta \to \infty$ and $\widehat{\tau}(\theta) = z_-\Theta(|\theta|^{-k})$ as $\theta \to -\infty$. In this case, τ has Fourier tail decay $K = 2k - 1$.*

---

[1]Here $\Theta(b_t)$ is to be understood as non-negative by definition. More generally we use the standard $\mathcal{O}(\cdot)$-notation: Write $a_t = \mathcal{O}(b_t)$ and $b_t = \Omega(|a_t|)$ if $|a_t| \leq Cb_t$ for all t and some constant $C > 0$. Write $\tilde{b}_t = \Theta(b_t)$ if $\tilde{b}_t = \mathcal{O}(b_t)$ and $\tilde{b}_t = \Omega(b_t)$. Write $a_t = \tilde{\mathcal{O}}(b_t)$ if $a_t = \mathcal{O}(b_t|\log b_t|^{\mathcal{O}(1)})$. Write $a_t = o(b_t)$ and $b_t = \omega(|a_t|)$ if $|a_t| = \epsilon_t b_t$ for some ϵ_t that converges to 0.

Our results below for rough activation functions hold for a more general definition of roughness which allows the Fourier transform to have varying phase. We give this definition of generalized rough activation functions in Appendix C.

We consider the two-layer network (2.3) with randomly initialized weights using two standard initialization strategies. It is typical to initialize the biases to zero.

Definition 3.4. *We say that*

$$f_{W,a}(x) = \sum_{k=1}^{m} a_k \tau\big(w^{(k)} \cdot x\big)$$

is chosen with the Xavier initialization or He initialization if the $mnd + m$ weights $W = (w_{ij}^{(k)})$, $a = (a_i)$ are chosen independently from $a_k \sim \mathcal{N}(0, \tilde{c}/m)$ for each $k = 1, \dots, m$, and $w_{ij}^{(k)} \sim \mathcal{N}(0, \tilde{c}/(nd))$ for $k = 1, \dots, m, i = 1, \dots, n$, and $j = 1, \dots, d$. Here $\tilde{c} = 1$ corresponds to the Xavier initialization and $\tilde{c} = 2$ to the He initialization.

Theorem 3.1 (Upper bound, sign problem deteriorates super-polynomially for smooth activation functions). *Let $f_{W,a} : \mathbb{R}^{nd} \to \mathbb{C}$ be given by a two-layer neural network (2.3) with activation function τ and let $\rho = \mathcal{N}(0, I_{nd})$ be the Gaussian envelope function. If τ is smooth and $f_{W,a}$ is sampled from the Xavier or He initializations, then with probability $1 - o(1)$ over W,*

$$\mathbb{E}_{a|W}\big[\|\mathcal{A}f_{W,a}\|_{\mathcal{N}}^2\big] = n^{-\omega(1)}.$$

Given an activation function $\tau : \mathbb{R} \to \mathbb{C}$ and weight vector $w \in \mathbb{R}^{nd}$ define $\tau_w : \mathbb{R}^{nd} \to \mathbb{C}$ by $\tau_w(x) = \tau(w \cdot x)$. Let $\ell_w = \tau_w^{\mathrm{LP}(t)}$ be the low-passed part of the activation function. The upper bound of $\|\mathcal{A}f\|_{\mathcal{N}}^2$ relies on the fact that the norm of the antisymmetrized function $\|\mathcal{A}\ell_w\|_{\mathcal{N}}^2$ decreases exponentially with respect to n when $t = \mathcal{O}(\sqrt{n/\log n})$ (Lemma 6.4). The high-passed part can be directly controlled by the tail decay of the activation function.

Theorem 3.2 (Lower bound, sign problem deteriorates at most polynomially for rough activation functions). *Let $f_{W,a} : \mathbb{R}^{nd} \to \mathbb{C}$ be given by a two-layer neural network (2.3) with activation function τ and weights sampled from the Xavier or He initialization, and let $\rho = \mathcal{N}(0, I_{nd})$ be the Gaussian envelope. If τ is rough or generalized rough with tail decay K, then with probability $1 - o(1)$ over W,*

$$\mathbb{E}_{a|W}\big[\|\mathcal{A}f_{W,a}\|_{\mathcal{N}}^2\big] = \tilde{\Omega}\big(n^{-(1+\frac{2}{d})K}\big),$$

where $\tilde{\Omega}$ denotes a lower bound up to $\log$-factors. In particular the magnitude of the antisymmetrized NN decays no faster than polynomially in n.

To prove Theorem 3.2 we show that for the high-passed part $h = \tau^{\mathrm{HP}(T)}$ at sufficiently large threshold, the norm of the antisymmetized function $\|\mathcal{A}h_w\|_{\mathcal{N}}^2$ can be approximated by the norm $\|h_w\|_{\mathcal{N}}^2$ before antisymmetrization (Lemma 6.1). The lower bound with polynomial scaling in n can also be viewed as evidence that the prefactor $1/\sqrt{n!}$ in Eq. (2.1) is the appropriate scaling.

We then show that when f is given by a two-layer NN with a rough activation function, $\mathcal{A}f$ can be computed efficiently to any inverse-polynomial precision relative to $\sqrt{\mathbb{E}[\|\mathcal{A}f\|_{\mathcal{N}}^2]}$:

Theorem 3.3 (Deterministic polynomial time algorithm for approximate evaluation of explicitly antisymmetrized two-layer neural network)**.** *Let $f_{W,a} : \mathbb{R}^{nd} \to \mathbb{C}$ be given by a two-layer neural network as in Theorem 3.2 with a rough or generalized rough activation function, and let $\epsilon = n^{-O(1)}$. There exists a deterministic polynomial-time algorithm $(W, a, b, x) \mapsto S_{W,a,b}(x)$ whose output is exactly antisymmetric in x and such that with probability $1 - o(1)$ over W,*

$$\mathbb{E}_{a|W}\big[\|S_{W,a,0} - \mathcal{A}f_{W,a}\|_{\mathcal{N}}^2\big] \le \epsilon \, \mathbb{E}_{a|W}\big[\|\mathcal{A}f_{W,a}\|_{\mathcal{N}}^2\big].$$

We consider the examples of activation functions in Table 3.1. tanh is smooth, so $\|\mathcal{A}f\|_{\mathcal{N}}^2$ decays super-polynomially with n with the tanh or sigmoid activation function. The ReLU activation is rough with tail decay 3. By Theorem 3.2, $\|\mathcal{A}f\|_{\mathcal{N}}^2$ is of order $\tilde{\Omega}(n^{-(3+6/d)})$ with the ReLU activation function. By Theorem 3.3, there exists an efficient algorithm to compute the output of f with inverse polynomial relative error when the activation function is chosen to be ReLU.

The approximation algorithm in Theorem 3.3 involves the approximate evaluation of an integral over frequencies θ.

Remark 3.1. In the setting of the standard Xavier/He initializations, Theorems 3.1-3.3 show that a rough activation function is required to avoid the sign problem. On the other hand it is still desirable to use smooth activation functions to obtain a smooth wavefunction. In the context of a smooth activation function, our results show that an initialization should be used in which the weights in the first layer are larger than those on the typical Xavier/He initializations. This re-scaling of the first layer need only be by an algebraic[2] factor $r = n^{O(1)}$. In this setting the approximation algorithm of Theorem 3.3 is unchanged except that the infra-red truncation t of the integral is replaced by t/r.

Using the smooth activation function tanh as an example, the modification in Remark 3.1 is equivalent to replacing tanh with $y \mapsto \tanh(ry)$ where r grows with n.

4 Reduction and generic weights

We now present an outline of the proofs of the main theorems. Additional details follow in Section 9.1 and Appendix A. We first reduce estimates of the magnitude of $\mathcal{A}f$ to estimates on the magnitude of $\mathcal{A}\tau_w$.

Lemma 4.1. *For $f_{W,a}(x)$ given by the network* (2.3) *with the He initialization (Definition 3.4),*

$$\mathbb{E}\big[\|\mathcal{A}f_{W,a}\|_\rho^2 \mid W\big] = \frac{\tilde{c}}{m} \sum_{k=1}^{m} \|\mathcal{A}\tau_{w^{(k)}}\|_\rho^2.$$

In particular,

$$\mathbb{E}\big[\|\mathcal{A}f_{W,a}\|_\rho^2\big] = \tilde{c}\,\mathbb{E}\big[\|\mathcal{A}\tau_w\|_\rho^2\big], \quad w \sim \mathcal{N}\left(0, \frac{\tilde{c}}{nd}I_{nd}\right) \tag{4.1}$$

does not depend on m.

---

[2]This is because the re-scaling factor can be chosen as the frequency threshold used to prove Theorem 3.2.

Proof. Expand

$$\|\mathcal{A}f_{W,a}\|_\rho^2 = \sum_{k,l=1}^m a_k a_l \langle \mathcal{A}\tau_{w^{(k)}} \mid \mathcal{A}\tau_{w^{(l)}} \rangle_\rho.$$

The $a_k \sim \mathcal{N}(0, \tilde{c}/m)$'s are independent so $\mathbb{E}[a_k a_l] = \delta_{kl}\tilde{c}/m$, and

$$\mathbb{E}\big[\|\mathcal{A}f_{W,a}\|_\rho^2 \mid W\big] = \frac{\tilde{c}}{m} \sum_{k,l=1}^m \delta_{kl}\mathbb{E}\big[\langle \mathcal{A}\tau_{w^{(k)}} \mid \mathcal{A}\tau_{w^{(l)}} \rangle_\rho\big] = \frac{\tilde{c}}{m} \sum_{k=1}^m \|\mathcal{A}\tau_{w^{(k)}}\|_\rho^2. \qquad (4.2)$$

Taking the expectation over W yields

$$\mathbb{E}_{W,a}\big[\|\mathcal{A}f_{W,a}\|_\rho^2\big] = \tilde{c}\,\mathbb{E}\big[\|\mathcal{A}\tau_v\|_\rho^2\big].$$

The proof is complete. $\square$

The next definition characterizes the weights of the first layer with high probability under the Xavier/He initializations.

Definition 4.1. *Fix a constant $C > 1$. We say that $w \in \mathbb{R}^{nd}$ is typical if $\tilde{c}/2 \le \|w\|^2 \le 2\tilde{c}$ and*

$$\|w\|_\infty := \max_{ij} |w_{ij}| \le C\sqrt{\frac{\log(nd)}{nd}}.$$

In particular a typical w has $\|w\| = \Theta(1)$. We view d as a constant, so

$$\|w\|_\infty = \mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right)$$

for typical w. For lower bounds we need an additional property of $w = w^{(k)}$ sampled from the Xavier/He initializations, namely that the $w_i \in \mathbb{R}^d$ are sufficiently separated. To see why this is needed, take the example where $w_i = w_j$ for some $i \ne j$ which would imply $\mathcal{A}\tau_w \equiv 0$. We formalize the separation property using the following quantity:

Definition 4.2. *For $w \in \mathbb{R}^{nd}$ write*

$$\delta_w = \frac{1}{2} \min_{1 \le i < j \le n} \|w_i \pm w_j\|,$$

where the minimum is over both choices of sign $\pm$.

We then define weights with typical separation:

Definition 4.3. *Fix a function $\delta(n) = o(n^{-(1/2+2/d)})$ as $n \to \infty$ (for concreteness we let $\delta(n) = n^{-(1/2+2/d)}/\sqrt{\log n}$). We say that v has typical separation if $\delta_v \ge \delta(n)$. We say that $W \in \mathbb{R}^{m \times nd}$ is typical if each $w^{(k)}$ is typical and at least half the $w^{(k)}$ have typical separation.*

In Appendix A we show that He/Xavier initialized weights generically have typical separation.

Lemma 4.2. *Let d be constant, let $m = \mathcal{O}(n^{C'})$ for some constant C', and let f be sampled as in Definition 3.4. Then W is typical with probability $1 - o(1)$ for some constant C in Definition 4.1 depending on C'. For such W,*

$$\frac{\tilde{c}}{2} \inf_{w \in S'} \|\mathcal{A}\tau_w\|_\rho^2 \leq \mathbb{E}_{a|W}\left[\|\mathcal{A}f\|_\rho^2\right] \leq \tilde{c} \sup_{w \in S} \|\mathcal{A}\tau_w\|_{\rho'}^2 \tag{4.3}$$

where S is the set of typical $w \in \mathbb{R}^{nd}$ and $S' \subset S$ is the set of typical w which have typical separation.

5 Overlap kernel induced by Fourier decomposition

Consider the case when the activation function is a complex exponential function. Let expression$([\mathbf{x}])$ denote the function $x \mapsto$ expression(x). Then $e^{iw \cdot [\mathbf{x}]} = \otimes_{i=1}^n e^{iw_i \cdot [\mathbf{x}]_i}$ is a product of single-particle functions, so antisymmetrizing it yields a Slater determinant

$$\mathcal{A}\left(e^{iw \cdot [\mathbf{x}]}\right) = \mathcal{A}\left(\otimes_{i=1}^n e^{iw_i \cdot [\mathbf{x}]_i}\right) = \wedge_{i=1}^n e^{iw_i \cdot [\mathbf{x}]_i}, \tag{5.1}$$

where the RHS is defined as $\det((e^{iw_i \cdot [\mathbf{x}]_j})_{ij})/\sqrt{n!}$. The overlap between two Slater determinants is the determinant of the overlap matrix [7], meaning that

$$\left\langle \mathcal{A}e^{iv \cdot [\mathbf{x}]} \,\middle|\, \mathcal{A}e^{iw \cdot [\mathbf{x}]} \right\rangle_\rho = \left\langle \wedge_i e^{iv_i \cdot [\mathbf{x}]} \,\middle|\, \wedge e^{iw_i \cdot [\mathbf{x}]} \right\rangle_\rho = \det B^{(v,w)}, \tag{5.2}$$

where $B^{(v,w)} \in \mathbb{R}^{n \times n}$ is given by $B_{ij}^{(v,w)} = \langle e^{iv_i \cdot [\mathbf{x}]} \,|\, e^{iw_j \cdot [\mathbf{x}]} \rangle_\rho$. We can evaluate this as

$$B_{ij}^{(v,w)} = \mathbb{E}_{X \sim \rho}\left[e^{-i(v_i - w_j) \cdot X}\right] = \mathfrak{F}\rho(v_i - w_j). \tag{5.3}$$

Here we have defined the un-normalized Fourier transform $\mathfrak{F}$ (also denoted by $\mathfrak{F}_d$) on $\mathbb{R}^d$ by

$$\mathfrak{F}\rho(v) = \int_{\mathbb{R}^d} e^{-iv \cdot x} d\rho(x).$$

In particular $\mathfrak{F}\rho(0) = 1$, and $\mathfrak{F}\rho = (2\pi)^{d/2}\widehat{\rho}$.

By the Fourier inversion formula (3.1) we have the identity of functions on $\mathbb{R}^{nd}$

$$\tau_w(x) = \frac{1}{\sqrt{2\pi}} \int_{|\theta| > \epsilon} \widehat{\tau}(\theta) e^{i\theta w \cdot x} d\theta + p(w \cdot x) + C_\epsilon + \mathcal{O}\left(\epsilon g(w \cdot x)\right). \tag{5.4}$$

We use the fact that low-degree polynomials vanish upon antisymmetrization:

Lemma 5.1 ([1, Lemma 7]). *If $f : \mathbb{R}^{nd} \to \mathbb{C}$ is a polynomial of degree $\deg f \leq n - 2$, then $\mathcal{A}f \equiv 0$. In particular, $\mathcal{A}\tau_w \equiv 0$ if τ is an activation function which is a polynomial of degree $\deg \tau \leq n - 2$.*

We antisymmetrize (5.4) and apply Lemma 5.1 which yields that for $n \geq \deg p + 2$,

$$\mathcal{A}\tau_w = \lim_{\epsilon \to 0} \mathcal{A}\tau^{\mathrm{HP}(\epsilon)} = \lim_{\epsilon \to 0} \frac{1}{\sqrt{2\pi}} \int_{|\theta| > \epsilon} \widehat{\tau}(\theta)\mathcal{A}\big(e^{i\theta w \cdot [\mathbf{x}]}\big)d\theta, \tag{5.5}$$

where convergence is in the $\mathcal{L}^2(\mathbb{R}^{nd}; \rho_n)$-norm.

Definition 5.1 (Overlap Kernel). *Define* $\mathrm{D}_\rho : \mathbb{R}^{nd} \times \mathbb{R}^{nd} \to \mathbb{C}$ *by* $\mathrm{D}_\rho(v, w) = \det B^{(v,w)}$, *where* $B^{(v,w)} \in \mathbb{C}^{n \times n}$ *is given by (5.3). Given a vector of weights* $w \in \mathbb{R}^{nd}$ *define* $D_\rho^{(w)} : \mathbb{R}^2 \to \mathbb{C}$ *by* $D_\rho^{(w)}(\theta, \tilde{\theta}) = \mathrm{D}_\rho(\theta w, \tilde{\theta} w)$.

Fig. 5.1 shows the overlap kernel $D_\rho^{(w)}(\theta, \tilde{\theta})$ averaged over Gaussian samples of w. Note that $D_\rho^{(w)}$ depends on the envelope function. Eq. (5.2) shows that

$$D_\rho^{(w)}(\theta, \tilde{\theta}) = \big\langle \mathcal{A}e^{i\theta w \cdot [\mathbf{x}]} \,\big|\, \mathcal{A}e^{i\tilde{\theta} w \cdot [\mathbf{x}]} \big\rangle_\rho.$$

The expansion (5.5) then yields Lemma 5.2:

Lemma 5.2. *If*

$$\iint_{\mathbb{R}^2} \big|\widehat{\tau}(\theta)\widehat{\tau}(\tilde{\theta})D_\rho^{(w)}(\theta, \tilde{\theta})\big|d\theta d\tilde{\theta} < \infty,$$

then

$$\|\mathcal{A}\tau_w\|_\rho^2 = \frac{1}{2\pi} \iint_{\mathbb{R}^2} \overline{\widehat{\tau}(\theta)}\widehat{\tau}(\tilde{\theta})D_\rho^{(w)}(\theta, \tilde{\theta})d\theta d\tilde{\theta}. \tag{5.6}$$
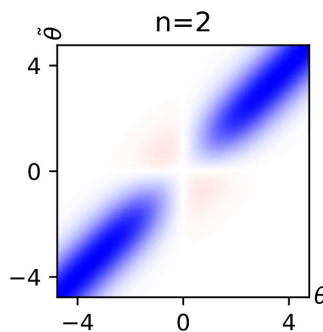

Figure 5.1: Heatmap of $\mathbb{E}_w D_\rho^{(w)}(\theta, \tilde{\theta})$, where $w \sim \mathcal{N}(0, 2/(nd)I_{nd})$ for $d = 3$ and $n = 2$. Positive values are blue and negative values red.

6 Properties of the overlap kernel

On the diagonal $v = w$ we have $0 \leq \mathrm{D}_\rho(w, w) \leq 1$. Indeed, $B_{ij}^{(w,w)}$ is the Gram matrix of vectors $\mathcal{A}e^{iw_i \cdot [\mathbf{x}]} \in \mathcal{L}^2(\mathbb{R}^d; \rho)$, so it satisfies

$$\det\big(B^{(w,w)}\big) \leq \prod_i B_{ii}^{(w,w)} = \big(\mathfrak{F}_d\rho(0)\big)^n = 1$$

by Hadamard's theorem for positive semidefinite matrices. $D_\rho(v, w)$ is a positive semidefinite kernel because it is a (infinite) Gram matrix of vectors $\mathcal{A}e^{iw\cdot[\mathbf{x}]} \in \mathcal{L}^2(\mathbb{R}^{nd}; \rho_n)$. In particular,

$$|D_\rho(v, w)|^2 \leq D_\rho(v, v) D_\rho(w, w) \leq 1$$

for all $v, w \in \mathbb{R}^{nd}$. Specializing these observations to $D_\rho^{(w)}$ yields $0 \leq D_\rho^{(w)}(\theta, \theta) \leq 1$ on the diagonal and $|D_\rho^{(w)}(\theta, \tilde{\theta})|^2 \leq 1$ everywhere.

To approximate the behavior of $D_\rho^{(w)}$ at large θ and $\tilde{\theta}$ we define a probability distribution ρ_w on $\mathbb{R}$.

Definition 6.1. *For any $w \in \mathbb{R}^{nd}$, let ρ_w on $\mathbb{R}$ be the distribution of $w \cdot X$ where $X \sim \rho_n$.*

We expect $\mathcal{A}e^{i\theta v\cdot[\mathbf{x}]}$ to be roughly orthogonal to $\mathcal{A}e^{i\tilde{\theta}v\cdot[\mathbf{x}]}$ when the frequencies θ and $\tilde{\theta}$ are sufficiently different. We therefore expect $D_\rho^{(w)}(\theta, \tilde{\theta}) = \langle \mathcal{A}e^{i\theta w\cdot[\mathbf{x}]} \mid \mathcal{A}e^{i\tilde{\theta}w\cdot[\mathbf{x}]} \rangle_\rho$ to vanish away from the diagonal $\tilde{\theta} = \theta$. Moreover, when θ is large we expect the $n!$ terms in (2.1) to be roughly orthogonal, so by Pythagoras' theorem and considering the normalization factor $1/\sqrt{n!}$ we expect

$$D_\rho^{(w)}(\theta, \theta) = \left\| \mathcal{A}e^{i\theta w\cdot[\mathbf{x}]} \right\|_\rho^2 \approx \left\| e^{i\theta w\cdot[\mathbf{x}]} \right\|_\rho^2 = 1$$

for large θ. We formalize this idea by approximating $D_\rho^{(w)}(\theta, \tilde{\theta})$ with a convolution kernel $\mathfrak{F}\rho_v(\theta - \tilde{\theta})$ when $\max\{|\theta|, |\tilde{\theta}|\} \geq T$ for sufficiently large $T > 0$. We then apply the convolution theorem to obtain

Lemma 6.1. *Let $h = \tau^{\mathrm{HP}(T)}$ be the high-passed activation function at threshold $T > 1$. Then*

$$\|\mathcal{A}h_w\|_\rho^2 = \|h_w\|_\rho^2 + \varepsilon(T) \tag{6.1}$$

with the error term

$$|\varepsilon(T)| = \mathcal{O}\left(n! \delta_w^{-2} \int_{\delta_w T}^\infty \epsilon_\rho(t) t \, dt \right), \quad \epsilon_\rho(\theta) = \sup\left\{ |\mathfrak{F}\rho(y)| : y \in \mathbb{R}^d, \|y\| \geq \theta \right\}.$$

Lemma 6.1 states that at sufficiently large threshold T the magnitude of the antisymmetrization of the high-passed activation function is approximately the magnitude before antisymmetrization.

The δ_w in the error term was defined in Definition 4.2. The fast decay of $\epsilon_\rho(\theta)$ compensates for the growth of the factor $n!$ in (6.1), making the approximation meaningful at polynomially large T.

6.1 Proof of Lemma 6.1

Lemma 6.2. *Let $\delta_w = \min_{i\neq j} \|w_i \pm w_j\|/2$ and let ρ_w be as in Definition 6.1. Then*

$$\left| D_\rho^{(w)}(\theta, \tilde{\theta}) - \mathfrak{F}\rho_w(\theta - \tilde{\theta}) \right| \leq n! \, \epsilon_\rho(\delta_w \theta_{\max}), \tag{6.2}$$

where

$$\theta_{\max} = \max\{|\theta|, |\tilde{\theta}|\}, \quad \epsilon_\rho(\theta) = \sup\left\{|\mathfrak{F}_d\rho(y)| : y \in \mathbb{R}^d, \|y\| \geq \theta\right\}.$$

$\mathfrak{F}\rho_w$ *can be expressed in terms of the nd-dimensional Fourier transform as* $\mathfrak{F}\rho_w(\theta) = \mathfrak{F}_{nd}\rho_n(\theta w).$

Proof. By definition we have

$$D_\rho^{(w)}(\theta, \tilde{\theta}) = \det\left(\mathfrak{F}_d\rho(\theta w_i - \tilde{\theta}w_j)_{ij}\right) = \sum_{\pi \in S_n} (-1)^\pi \mathfrak{F}_{nd}\rho_n(\theta w - \tilde{\theta}\pi(w)).$$

The $\pi = 1$ term equals

$$\mathfrak{F}_{nd}\rho_n((\theta - \tilde{\theta})w) = \mathfrak{F}\rho_w(\theta - \tilde{\theta}). \tag{6.3}$$

To verify the identity (6.3) and the claim at the end of the lemma, write

$$\mathfrak{F}_{nd}\rho_n(\theta w) = \mathbb{E}_{X \sim \rho_n}\left[e^{-i(\theta w)\cdot X}\right] = \mathbb{E}_{X \sim \rho_n}\left[e^{-i\theta(w\cdot X)}\right] = \mathbb{E}_{y \sim \rho_w}\left[e^{-i\theta y}\right] = \mathfrak{F}\rho_w(\theta).$$

The difference on the LHS of (6.2) is exactly

$$\varepsilon = \sum_{\pi \neq 1} (-1)^\pi \mathfrak{F}_{nd}\rho_n(\theta w - \tilde{\theta}\pi(w)).$$

Apply the triangle inequality and use the fact that $\|\epsilon_\rho\|_\infty = \|\mathfrak{F}_{nd}\rho_n\|_\infty \leq 1$ because ρ is a probability distribution to obtain

$$|\varepsilon| \leq n! \max_{\pi \neq 1} \left|\mathfrak{F}_{nd}\rho_n(\theta w - \tilde{\theta}\pi(w))\right|$$

$$\leq n! \max_{\pi \neq 1} \prod_{i=1}^n \epsilon_\rho\left(\|\theta w_i - \tilde{\theta}\pi(w)_i\|\right)$$

$$\leq n! \max_{\pi \neq 1} \min_i \epsilon_\rho\left(\|\theta w_i - \tilde{\theta}\pi(w)_i\|\right) \leq n! \, \epsilon_\rho\left(|\theta|\tilde{\delta}(w)\right), \tag{6.4}$$

where

$$\tilde{\delta}(v) = \min_{\tilde{\theta}} \min_{\pi \neq 1} \max_i \|v_i - \tilde{\theta}\pi(v)_i\|. \tag{6.5}$$

We show that $\tilde{\delta}_w \geq \delta_w$ in Lemma 6.3. The result follows from symmetry in θ and $\tilde{\theta}$. $\qquad \square$

Lemma 6.3. *Let* $\tilde{\delta}_w := \min_{\tilde{\theta} \in \mathbb{R}} \min_{\pi \neq 1} \max_i \|w_i - \tilde{\theta}\pi(w)_i\|$. *Then* $\tilde{\delta}_w \geq \delta_w$.

Proof. Let $\tilde{\theta}$ and $\sigma = \pi^{-1} \neq 1$ be arbitrary. We need to show that $\|w_i - \tilde{\theta}w_{\sigma(i)}\| \geq \delta_w$ for some i.

Since $\sigma \neq 1$ there exists i such that $\sigma(i) \neq i$, therefore $\|w_i - sw_{\sigma(i)}\| \geq 2\delta_w$ for $s = \pm 1$. If $\|w_i - \tilde{\theta}w_{\sigma(i)}\| \geq \delta_w$ then we have found i and we are done. Otherwise, let $s = \text{sign}\,\tilde{\theta}$. Then,

$$2\delta_w \leq \|w_i - sw_{\sigma(i)}\| \leq \|w_i - \tilde{\theta}w_{\sigma(i)}\| + \|\tilde{\theta}w_{\sigma(i)} - sw_{\sigma(i)}\|$$

$$= \|w_i - \tilde{\theta}w_{\sigma(i)}\| + |\tilde{\theta} - s| \cdot \|w_{\sigma(i)}\| < \delta_w + |\tilde{\theta} - s| \max_j \|w_j\|. \tag{6.6}$$

Rearranging (6.6) we yields

$$|\tilde{\theta} - s|R > \delta_w, \quad R = \max_j \|w\|_j. \tag{6.7}$$

Pick j such that $\|w_j\| = R$. If $|\tilde{\theta}| \geq 1$ then (6.7) says $(|\tilde{\theta}| - 1)R > \delta_w$. Let $i = \pi(j)$ so that $\sigma(i) = j$ and $\|w_{\sigma(i)}\| = R$. Then,

$$\|w_i - \tilde{\theta}w_{\sigma(i)}\| \geq \|\tilde{\theta}w_{\sigma(i)}\| - \|w_i\| = |\tilde{\theta}|R - \|w_i\| \geq (|\tilde{\theta}| - 1)R \geq \delta_w.$$

If instead $|\tilde{\theta}| < 1$ then (6.7) says $(1 - |\tilde{\theta}|)R > \delta_w$. We then use

$$\|w_j - \tilde{\theta}w_{\sigma(j)}\| \geq \|w_j\| - \|\tilde{\theta}w_{\sigma(j)}\| = R - \|\tilde{\theta}w_{\sigma(j)}\| \geq (1 - |\tilde{\theta}|)R \geq \delta_w.$$

The proof is complete. $\qquad\square$

Proof of Lemma 6.1. By (5.6) we have

$$\|\mathcal{A}h_w\|_\rho^2 = \frac{1}{2\pi} \iint_{|\theta|,|\tilde{\theta}| \geq T} D_\rho^{(w)}(\theta, \tilde{\theta}) \overline{\hat{\tau}(\theta)}\hat{\tau}(\tilde{\theta})d\theta d\tilde{\theta}.$$

Apply Lemma 6.2 to approximate $D_\rho^{(w)}(\theta, \tilde{\theta})$ by $\mathfrak{F}\rho_w(\theta - \tilde{\theta})$. Recall that $\hat{h} = \hat{\tau} \odot 1_{\mathbb{R}\backslash[-T,T]}$. The resulting approximation is, by the convolution theorem and Plancherel's identity,

$$\frac{1}{2\pi} \iint \mathfrak{F}\rho_w(\theta - \tilde{\theta})\overline{\hat{h}(\theta)}\hat{h}(\tilde{\theta})d\theta d\tilde{\theta}$$
$$= \frac{1}{2\pi} \int \overline{\hat{h}(\theta)}\left(\mathfrak{F}\rho_w * \hat{h}\right)(\theta)d\theta$$
$$= \int |h(y)|^2 \rho_w(y)dy = \mathbb{E}_{X \sim \rho_n}\left[|h(w \cdot X)|^2\right] = \|h_w\|_\rho^2. \tag{6.8}$$

$\hat{\tau}$ is bounded on $\mathbb{R}\backslash[-T,T]$ by assumption. The error of the approximation is then bounded by a constant times

$$\iint_{|\theta|,|\tilde{\theta}| \geq T} \left|D_\rho^{(w)}(\theta, \tilde{\theta}) - \mathfrak{F}\rho_w(\theta - \tilde{\theta})\right|d\theta d\tilde{\theta}$$
$$\leq 8 \int_T^\infty \int_T^\theta n!\epsilon_\rho(\delta_w \theta)d\tilde{\theta}d\theta \leq 8n! \int_T^\infty \epsilon_\rho(\delta_w \theta)\,\theta d\theta.$$

The error bound follows by substituting $t = \delta_w\theta$. $\qquad\square$

6.2 Upper bound for Gaussian envelopes

Lemma 6.1 explains the behavior of $D_\rho^{(w)}(\theta, \tilde{\theta})$ at large θ. We now establish that it vanishes for small θ. When the envelope is the standard Gaussian $\rho = \mathcal{N}$, the overlap kernel takes the following form:

$$D_\mathcal{N}(v, w) = e^{-\frac{\|v\|^2 + \|w\|^2}{2}} \det\left((e^{v_i \cdot w_j})_{ij}\right). \tag{6.9}$$

An upper bound on $D_\mathcal{N}$ was obtained in [1, Proposition 11]:

Proposition 6.1. *Let $v = 2\sqrt{d\|v\|_\infty\|w\|_\infty}$. Then, $\det((e^{v_i \cdot w_j})_{ij}) \leq (v/2)^{pn}$ for $v \leq 1$, where p is any integer such that*

$$\binom{p+d-1}{d} \leq \frac{n}{2} \quad and \quad p! \geq 4n^2.$$

The proof of Proposition 6.1 is recalled from [1] in Appendix B. It works by decomposing $(e^{v_i \cdot w_j})_{ij} = \sum_{k=0}^\infty Q_k$ and bounding the ranks and operator norms of the terms Q_k. For $L = \sum_{k=1}^{p-1} \text{rank } Q_k$, the L-th eigenvalue is then bounded as the tail sum $\sum_{k=p}^\infty \|Q_k\|$. Taking the product of the eigenvalues yields the bound on (6.9).

Combining Proposition 6.1 with the triangle inequality yields:

Lemma 6.4. *Let $\ell = \tau^{\text{LP}(t)}$ be the low-pass at threshold $t = (2\sqrt{d}\|w\|_\infty)^{-1}$. If w is typical then $t = \Omega(\sqrt{n/\log n})$ and $\|\mathcal{A}\ell_w\|_\mathcal{N} = \mathcal{O}(2^{-\Omega(n^{1+1/d})})$.*

7   Proof of Theorem 3.1

We specialize the quantities of Lemma 6.1 to the case of a Gaussian envelope. For the Gaussian envelope $\rho = \mathcal{N}(0, I_d)$ we have

$$\mathfrak{F}_d\rho(v) = e^{-\frac{\|v\|^2}{2}}, \quad \mathfrak{F}\rho_w(\theta) = e^{-\frac{\|w\|^2\theta^2}{2}}, \quad \epsilon_\rho(\theta) = e^{-\frac{\theta^2}{2}}.$$

The integral in the error term of Lemma 6.1 becomes

$$\int_{\delta_w T}^\infty e^{-\frac{t^2}{2}} t \, dt = e^{-\frac{(\delta_w T)^2}{2}},$$

so we get the approximation

$$\|\mathcal{A}h_w\|_\mathcal{N}^2 = \|h_w\|_\mathcal{N}^2 + \mathcal{O}\left(\delta_w^{-2} e^{-\frac{\delta_w^2 T^2}{2} + n\log n}\right). \tag{7.1}$$

We have the two following expressions for $D_\mathcal{N}^{(w)}$:

$$D_\mathcal{N}^{(w)}(\theta, \tilde{\theta}) = \det\left[\left(e^{-\frac{1}{2}\|\theta w_i - \tilde{\theta} w_j\|^2}\right)_{ij}\right] = e^{-\frac{\theta^2 + \tilde{\theta}^2}{2}\|w\|^2} \det\left[\left(e^{\theta\tilde{\theta}w_i \cdot w_j}\right)_{ij}\right]. \tag{7.2}$$

It follows from (7.2) that $D_\mathcal{N}^{(w)}(\theta, \tilde{\theta})$ can be determined from its values on the diagonal and anti-diagonal $\tilde{\theta} = \pm\theta$. More precisely we have

$$D_\mathcal{N}^{(w)}(\theta, \tilde{\theta}) = e^{-\frac{\|w\|^2}{2}(|\theta| - |\tilde{\theta}|)^2} D_\mathcal{N}^{(w)}(\theta_{\text{g.m.}}, \pm\theta_{\text{g.m.}}), \tag{7.3}$$

where the geometric mean $\theta_{\text{g.m.}} = |\theta\tilde{\theta}|^{1/2}$ and $\pm$ is the sign of $\theta\tilde{\theta}$. To show (7.3) apply the rightmost expression of (7.2) on both sides and note that the determinants are equal. In particular we have that $D_\mathcal{N}^{(w)}$ decays away from the diagonal and anti-diagonal

$$\left|D_\mathcal{N}^{(w)}(\theta, \tilde{\theta})\right| \leq e^{-\frac{\|w\|^2}{2}(|\theta| - |\tilde{\theta}|)^2}. \tag{7.4}$$

The bound (7.4) gives concentration around the diagonal everywhere and not only for large $\theta, \tilde{\theta}$ as Lemma 6.2. Integrating (7.4) yields:

Lemma 7.1. *Let $h = \tau^{\mathrm{HP}(t)}$ be the high-passed activation function at $t > 1$. Then,*

$$\|\mathcal{A}h_w\|_{\mathcal{N}}^2 \leq \frac{4}{\sqrt{2\pi}\|w\|}\mathfrak{T}_{\widehat{\tau}}(t).$$

So $\|\mathcal{A}h_w\|_{\mathcal{N}}^2 = \mathcal{O}(t^{-K})$ for typical w.

Proof. Write the LHS as a double integral over $|\theta|, |\tilde{\theta}| \geq t$ as in (5.6). By (7.4),

$$
\begin{aligned}
&\frac{1}{2\pi} \iint_{\tilde{\theta} > \theta > t} D_{\mathcal{N}}^{(w)}(\theta, \tilde{\theta}) |\widehat{\tau}(\theta)\widehat{\tau}(\tilde{\theta})| d\theta d\tilde{\theta} \\
&= \frac{1}{2\pi} \int_0^\infty e^{-\frac{\|w\|^2 s^2}{2}} \int_t^\infty |\widehat{\tau}(\theta)\widehat{\tau}(s+\theta)| d\theta ds \\
&\leq \frac{1}{2\pi} \int_0^\infty e^{-\frac{\|w\|^2 s^2}{2}} ds \int_t^\infty |\widehat{\tau}(\theta)|^2 d\theta \\
&= \frac{1}{2\sqrt{2\pi}\|w\|} \int_t^\infty |\widehat{\tau}(\theta)|^2 d\theta,
\end{aligned}
\tag{7.5}
$$

where (7.5) follows from Cauchy-Schwartz. The same bound holds for each of 8 regions in the $\theta, \tilde{\theta}$-plane. $\qquad\square$

For typical w choose threshold $t = \Omega(\sqrt{n/\log n})$ as in Lemma 6.4 and decompose $\mathcal{A}\tau_w = \mathcal{A}\ell_w + \mathcal{A}h_w$. Apply Lemmas 6.4 and 7.1 to obtain the following corollary, which implies Theorem 3.1.

Corollary 7.1. *If τ has frequency tail decay $K < \infty$ then $\|\mathcal{A}\tau_w\|_{\mathcal{N}}^2 = \tilde{O}(n^{-K/2})$ for typical w. If τ is smooth ($K = \infty$), then $\|\mathcal{A}\tau_w\|_{\mathcal{N}}^2 = n^{-\omega(1)}$ for typical w.*

8 Proof of Theorem 3.2

For the Gaussian envelope function we get a more explicit form of the error term in Lemma 6.1. To prove the lower bound we pick an appropriate T based on this expression and decompose τ into $\ell = \tau^{\mathrm{LP}(T)}$ and $h = \tau^{\mathrm{HP}(T)}$. We then expand $\|\mathcal{A}\ell + \mathcal{A}h\|_{\mathcal{N}}^2$ and use that $\|\mathcal{A}\ell\|_{\mathcal{N}}^2 \geq 0$ to get

$$\|\mathcal{A}\tau_w\|_{\mathcal{N}}^2 \geq \|\mathcal{A}h_w\|_{\mathcal{N}}^2 - 2|\langle \mathcal{A}\ell_w \,|\, \mathcal{A}h_w\rangle_{\mathcal{N}}|.$$

The result follows by bounding $|\langle \mathcal{A}\ell_w \,|\, \mathcal{A}h_w\rangle_{\mathcal{N}}|$ and applying Lemma 6.1 to lower-bound $\|\mathcal{A}h_w\|_{\mathcal{N}}$.

Lemma 8.1. *For activation functions satisfying Item 1 of Definition C.1 we have that*

$$\|\mathcal{A}\tau_w\|_{\mathcal{N}}^2 \geq \|h_w^{(T)}\|_N^2 - \mathcal{O}(e^{-n\log n}) + o(\mathfrak{T}_{\widehat{\tau}}(T)) \tag{8.1}$$

for

$$T = \frac{2}{\delta_v}\left(\sqrt{n\log n + \log\frac{1}{\delta_v}}\right)$$

and typical w.

Proof. Let $\ell^{(T)} = \tau^{\mathrm{LP}(T)}$ and $h^{(T)} = \tau^{\mathrm{HP}(T)}$. As discussed above we have the lower bound

$$\begin{aligned}
\|\mathcal{A}\tau_w\|_{\mathcal{N}}^2 &\geq \|\mathcal{A}h_w^{(T)}\|_{\mathcal{N}}^2 - 2\big|\langle \mathcal{A}\ell_w^{(T)} \mid \mathcal{A}h^{(T)}\rangle_{\mathcal{N}}\big| \\
&\geq \|h_w^{(T)}\|_{\mathcal{N}}^2 - e^{-n\log n} - 2\big|\langle \mathcal{A}\ell_w^{(T)} \mid \mathcal{A}h_w^{(T)}\rangle_{\mathcal{N}}\big|.
\end{aligned} \tag{8.2}$$

We further divide the low-frequency part into $|\theta|$ in $[0,1]$, $[1, T/2]$, and $[T/2, T]$. We write $\alpha^{(t,T)} = h^{(t)} - h^{(T)}$. Then,

$$\begin{aligned}
\big|\langle \mathcal{A}\ell_w^{(T)} \mid \mathcal{A}h_w^{(T)}\rangle_{\mathcal{N}}\big| &\leq \|\mathcal{A}\ell_w^{(1)}\|_{\mathcal{N}}\|\mathcal{A}h_w^{(T)}\|_{\mathcal{N}} \\
&\quad + \big|\langle \mathcal{A}\alpha_w^{(1,T/2)} \mid \mathcal{A}h_w^{(T)}\rangle_{\mathcal{N}} + \langle \mathcal{A}\alpha_w^{(T/2,T)} \mid \mathcal{A}h_w^{(T)}\rangle_{\mathcal{N}}\big|.
\end{aligned} \tag{8.3}$$

By applying the polarization identity to (5.6) we obtain the overlap between the antisymmetrization with different activation functions, so by the bound (7.4) on $D_{\mathcal{N}}^{(w)}(\theta, \tilde{\theta})$

$$\begin{aligned}
\big|\langle \mathcal{A}\alpha_w^{(T/2,T)} \mid \mathcal{A}h_w^{(T)}\rangle_{\mathcal{N}}\big| &\leq \frac{\big(\sup_{|\theta|\geq T/2}|\widehat{\tau}(\theta)|\big)^2}{2\pi}\int_{T/2<|\theta|<T}\int_{T<|\tilde{\theta}|}e^{-\frac{\|w\|^2}{2}(|\tilde{\theta}|-|\theta|)^2}d\tilde{\theta}d\theta \\
&\leq \frac{4\big(\sup_{|\theta|\geq T/2}|\widehat{\tau}(\theta)|\big)^2}{2\pi}\int_{T/2<\theta<T}\int_{T-\theta}^{\infty}e^{-\frac{\|w\|^2 t^2}{2}}dt d\theta \\
&= \frac{4\big(\sup_{|\theta|\geq T/2}|\widehat{\tau}(\theta)|\big)^2}{2\pi}\int_0^{\infty}\min\{t, T/2\}e^{-\frac{\|w\|^2 t^2}{2}}dt \\
&\leq \frac{4\sup_{|\theta|\geq T/2}|\widehat{\tau}(\theta)|^2}{2\pi}\int_0^{\infty}t e^{-\frac{\|w\|^2 t^2}{2}}dt \\
&= \frac{2\sup_{|\theta|\geq T/2}|\widehat{\tau}(\theta)|^2}{\pi\|w\|^2},
\end{aligned} \tag{8.4}$$

where in (8.4) the factor 4 comes from the choice of signs of $\theta, \tilde{\theta}$, and we have substituted $t = \tilde{\theta} - \theta$. We apply Item 1 of Definition C.1 to obtain that

$$\big|\langle \mathcal{A}\alpha_w^{(T/2,T)} \mid \mathcal{A}h_w^{(T)}\rangle_{\mathcal{N}}\big| = o\left(\frac{\mathfrak{T}_{\widehat{\tau}}(T)}{\|w\|^2}\right). \tag{8.5}$$

For the first term on the RHS of (8.3) we have for T sufficiently large (so that $|\widehat{\tau}(\theta)\widehat{\tau}(\tilde{\theta})| \leq 2\pi$ for $|\theta| > 1$ and $|\tilde{\theta}| > T$),

$$\big|\langle \mathcal{A}\alpha_w^{(1,T/2)} \mid \mathcal{A}h_w^{(T)}\rangle_{\mathcal{N}}\big| \leq \int_{1<|\theta|<T/2}\int_{T<|\tilde{\theta}|}e^{-\frac{\|w\|^2}{2}(|\tilde{\theta}|-|\theta|)^2}d\tilde{\theta}d\theta$$

$$\leq \frac{T}{2} \int_{T/2}^{\infty} e^{-\frac{\|w\|^2 t^2}{2}} dt \leq \frac{1}{\|w\|^2} e^{-\frac{\|w\|^2 T^2}{8}}. \tag{8.6}$$

By (8.5) and (8.6),

$$\left| \left\langle \alpha_w^{(1,T/2)} \mid h_w^{(T)} \right\rangle_{\mathcal{N}} + \left\langle \alpha_w^{(T/2,T)} \mid h_w^{(T)} \right\rangle_{\mathcal{N}} \right| = \mathcal{O}\left(e^{-\Omega(T^2)}\right) + o\left(\mathfrak{T}_{\widehat{\tau}}(T)\right)$$

for typical w. Finally, for typical w we have $t > 1$ in Lemma 6.4 and

$$\left\| h_w^{(T)} \right\|_{\mathcal{N}} = \mathcal{O}(t^{-K}) = \mathcal{O}(1)$$

by Lemma 7.1, so

$$\left\| \ell_w^{(1)} \right\|_{\mathcal{N}} \left\| h_w^{(T)} \right\|_{\mathcal{N}} = \mathcal{O}\left(\left\| \ell_w^{(t)} \right\|_{\mathcal{N}} \right) = 2^{-\Omega(n^{1+\frac{1}{d}})}.$$

The claim follows by substituting back into (8.2) and (8.3). $\qquad\qquad\square$

Proof of Theorem 3.2. By Lemma 4.2 it suffices to show the claim for $f = \tau_w$ where w is typical and has typical separation.

By the convolution theorem (6.8) and the expression $\mathfrak{F}\rho_v(\theta) = e^{-\|v\|^2\theta^2/2}$ for the Gaussian case we have

$$\left\| h_w^{(T)} \right\|_{\mathcal{N}}^2 = \frac{1}{2\pi} \iint_{|\theta|,|\tilde\theta|>T} e^{-\frac{\|w\|^2(\theta-\tilde\theta)^2}{2}} \overline{\widehat{\tau}(\theta)} \widehat{\tau}(\tilde\theta) d\theta d\tilde\theta. \tag{8.7}$$

We bound the contribution of the first quadrant to (8.7) from below. The same argument holds for the third quadrant ($\theta, \tilde\theta < -T$). Let $\sigma = 1/\|w\|$. Then

$$\iint_{\theta,\tilde\theta>T} e^{-\frac{\|w\|^2(\theta-\tilde\theta)^2}{2}} \overline{\widehat{\tau}(\theta)} \widehat{\tau}(\tilde\theta) d\theta d\tilde\theta$$

$$= 2\,\mathrm{Re} \int_T^{\infty} \overline{\widehat{\tau}(\theta)} \int_{\theta}^{\infty} e^{-\frac{\|w\|^2(\tilde\theta-\theta)^2}{2}} \widehat{\tau}(\tilde\theta) d\tilde\theta d\theta$$

$$= 2\,\mathrm{Re} \int_T^{\infty} \overline{\widehat{\tau}(\theta)} \int_{\theta}^{\infty} e^{-\frac{(\tilde\theta-\theta)^2}{2\sigma^2}} \widehat{\tau}(\tilde\theta) d\tilde\theta d\theta$$

$$= \sqrt{2\pi}\sigma \int_T^{\infty} \mathrm{Re}\left(\overline{\widehat{\tau}(\theta)} \mathbb{E}[\widehat{\tau}(\theta + |Y|)] \right) d\theta, \tag{8.8}$$

where $Y \sim \mathcal{N}(0,\sigma^2)$. To obtain a lower bound on the integrand we write

$$\overline{\widehat{\tau}(\theta)} \cdot \mathbb{E}[\widehat{\tau}(\theta + |Y|)] = \|\widehat{\tau}(\theta)\|^2 \cdot \frac{\mathbb{E}[\widehat{\tau}(\theta + |Y|)]}{\widehat{\tau}(\theta)}. \tag{8.9}$$

So from Item 2 of Definition C.1 we then have a lower bound

$$\mathrm{Re}\left(\overline{\widehat{\tau}(\theta)} \cdot \mathbb{E}[\widehat{\tau}(\theta + |Y|)] \right) = \|\widehat{\tau}(\theta)\|^2 \,\mathrm{Re}\left(\frac{\mathbb{E}[\widehat{\tau}(\theta + |Y|)]}{\widehat{\tau}(\theta)} \right) = \Omega\left(|\widehat{\tau}(\theta)|^2\right), \tag{8.10}$$

when $1/2 \le \sigma \le 2$, i.e. when $1/4 \le \|w\|^2 \le 4$. This holds because w is typical (and because $1/2 \le \tilde{c} \le 2$). For typical w (8.7) and (8.10) then show that

$$\iint_{\theta,\tilde{\theta}>T} e^{-\frac{\|w\|^2(\theta-\tilde{\theta})^2}{2}} \overline{\widehat{\tau}(\theta)}\widehat{\tau}(\tilde{\theta})d\theta d\tilde{\theta} = \Omega\left(\int_T^\infty |\widehat{\tau}(\theta)|^2 d\theta \right) = \Omega\big(\mathfrak{T}_{\widehat{\tau}}(T)\big), \tag{8.11}$$

and the same lower bound holds for the integral over $\theta,\tilde{\theta} < T$. Finally we bound the second quadrant $\theta < -T, \tilde{\theta} > t$ (the fourth quadrant $\theta > T, \tilde{\theta} < -T$ is analogous) by writing

$$\left| \iint_{\theta<-T,\tilde{\theta}>T} e^{-\frac{\|w\|^2(\theta-\tilde{\theta})^2}{2}} \overline{\widehat{\tau}(\theta)}\widehat{\tau}(\tilde{\theta})d\theta d\tilde{\theta}\right|$$

$$= o\left(\iint_{\theta<-T,\tilde{\theta}>T} e^{-\frac{\|w\|^2(\tilde{\theta}-\theta)^2}{2}} d\theta d\tilde{\theta} \right)$$

$$= o\left(\int_{2T}^\infty t e^{-\frac{\|w\|^2 t^2}{2}} dt \right) = o\frac{e^{-2\|w\|^2 T^2}}{\|w\|^2} = o\big(e^{-T^2}\big), \tag{8.12}$$

where the last expression os for typical w. The identity (8.7), the lower bound for the diagonal part (8.11), and the bound on the magnitude of the off-diagonal part (8.12) show that

$$\big\|h_w^{(T)}\big\|_{\mathcal{N}}^2 = \Omega\big(\mathfrak{T}_{\widehat{\tau}}(T)\big) = \Omega(T^{-K}).$$

For typical w with typical separation we have that $\delta_w \ge n^{-(1/2+2/d)}/\sqrt{\log n}$ and $T \sim 2n^{1+2/d}\log n$, so

$$\|h_w^{(T)}\|_{\mathcal{N}}^2 = \tilde{\Omega}\big(n^{-(1+\frac{2}{d})}\big).$$

The proof is complete. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \square$

9 Efficient algorithm

Recall (5.1) which gives the antisymmetrization with an exponential activation function as a determinant

$$\mathcal{A}\big(e^{iw\cdot[\mathbf{x}]}\big) = \frac{1}{\sqrt{n!}} \det\left(\big(e^{iw_i\cdot[\mathbf{x}]_j}\big)_{ij} \right).$$

Let $\ell^{(t)} = \tau^{\mathrm{LP}(t)}$ and $h^{(T)} = \tau^{\mathrm{HP}(T)}$ be the low-pass and high-pass of τ at thresholds $t < T$. We approximate $\mathcal{A}\tau_w$ by removing the low-passed and high-passed components: Apply (5.5) to $\tau^{\mathrm{HP}(t)} - \tau^{\mathrm{HP}(T)}$ to obtain

$$\mathcal{A}\tau_w(x) = \alpha_w + \mathcal{A}\ell_w^{(t)} + \mathcal{A}h_w^{(T)},$$

where

$$\alpha_w(x) = \frac{1}{\sqrt{2\pi n!}} \int_{[-T,T]\setminus[-t,t]} \widehat{\tau}(\theta) \det\left(\big(e^{i\theta w_i\cdot x_j}\big)_{ij} \right) d\theta. \tag{9.1}$$

The integrand can be computed at a single θ in time $\mathcal{O}(n^3)$. Apply Lemmas 6.4 and 7.1 (with a different choice of threshold for the high-pass) to bound the truncation error.

Lemma 9.1 (Truncation Error Bound). *Suppose $\widehat{\tau}$ has tail decay K. Let*

$$t = \max\left\{\frac{1}{2\sqrt{d}\|w\|_\infty}, 1\right\}$$

and, given $\epsilon > 0$, let $T = \epsilon^{-1/K}$. Then

$$\|\mathcal{A}\tau_w - \alpha_w\|_\mathcal{N}^2 = \mathcal{O}\left(2^{-\Omega(n^{1+\frac{1}{d}})} + \epsilon\right)$$

for typical w.

We approximate $\alpha_w(x)$ by a sum

$$S_w(x) = \frac{1}{\sqrt{2\pi}n!}\sum_{p=\pm 1,\ldots,\pm N} c_p \det\left(\left(e^{i\theta_p w_i \cdot x_j}\right)_{ij}\right),$$

where $\theta_1, \ldots, \theta_N$ are a discretization of $[t, T]$.

9.1 Discretization error bound

Let $t = \Omega(1)$ and let $t = t_0 < t_1 < \cdots < t_N = T$ be evenly spaced and for each $p = 1, \ldots, N$ write

$$I_{-p} = [-t_p, -t_{p-1}], \quad I_p = [t_{p-1}, t_p].$$

For $p = \pm 1, \ldots, \pm N$, let $\theta_p \in I_p$ and define

$$c_p := \int_{I_p} \widehat{\tau}(\theta) d\theta.$$

For these evenly spaced θ_p and coefficients c_p, define

$$\left\|\partial_\theta D_\mathcal{N}^{(w)}\right\|_\infty := \sup_{\theta,\tilde{\theta}}\left|\frac{\partial}{\partial\tilde{\theta}}D_\mathcal{N}^{(w)}(\theta,\tilde{\theta})\right|.$$

We then have

Lemma 9.2.

$$\|S_w - \alpha_w\|_\mathcal{N}^2 \le \frac{2}{\pi}\frac{T}{N}\left\|\partial_\theta D_\mathcal{N}^{(w)}\right\|_\infty\left(\int_{t \le |\theta| \le T}|\widehat{\tau}(\theta)|\right)^2 = \mathcal{O}\left(\frac{T}{N}\left\|\partial_\theta D_\mathcal{N}^{(w)}\right\|_\infty\right).$$

Proof. S_w is exactly the antisymmetrization $\mathcal{A}s_w$ where

$$s(t) = \frac{1}{\sqrt{2\pi}}\sum_{\pm q=1}^{N} c_q e^{i\theta_q t}.$$

Passing from the truncated activation to its discretization incurs the error

$$\|S_w - \tilde{\alpha}_w\|_{\mathcal{N}}^2 = \|\mathcal{A}(s - \tau_{\mathrm{trc}})_w\|_{\mathcal{N}}^2 = \frac{1}{2\pi}\iint_{\mathbb{R}^2} D_{\mathcal{N}}^{(w)}(\theta,\tilde{\theta})d\bar{\mu}(\theta)d\mu(\tilde{\theta}), \qquad (9.2)$$

where $\tau_{\mathrm{trc}} = \tau - \tau^{\mathrm{LP}(1)} - \tau^{\mathrm{HP}(T)}$, and μ is the complex-valued measure

$$\mu(\theta) = \sum_q c_q \delta(\theta - \theta_q) - \left(\widehat{\tau}(\theta)d\theta\right).$$

For each square $S_{pq} = I_p \times I_q$,

$$\iint_{S_{pq}} D_{\mathcal{N}}^{(w)}(\theta,\tilde{\theta})d\bar{\mu}(\theta)d\mu(\tilde{\theta})$$

$$= \overline{c_p}c_q D_{\mathcal{N}}^{(w)}(\theta_p,\theta_q) + \iint_{S_{pq}} D_{\mathcal{N}}^{(w)}(\theta,\tilde{\theta})\overline{\widehat{\tau}(\theta)}\widehat{\tau}(\tilde{\theta})d\theta d\tilde{\theta}$$

$$\quad - \overline{c_p}\int_{I_q} D_{\mathcal{N}}^{(w)}(\theta_p,\tilde{\theta})\widehat{\tau}(\tilde{\theta})d\tilde{\theta} - c_q \int_{I_p} D_{\mathcal{N}}^{(w)}(\theta,\theta_q)\overline{\widehat{\tau}(\theta)}d\theta$$

$$= \iint_{S_{pq}} \left[D_{\mathcal{N}}^{(w)}(\theta_p,\theta_q) + D_{\mathcal{N}}^{(w)}(\theta,\tilde{\theta}) - D_{\mathcal{N}}^{(w)}(\theta_p,\tilde{\theta}) - D_{\mathcal{N}}^{(w)}(\theta,\theta_q) \right] \overline{\widehat{\tau}(\theta)}\widehat{\tau}(\tilde{\theta})d\theta d\tilde{\theta}$$

$$\leq \iint_{S_{pq}} \left[2\max D_{\mathcal{N}}^{(w)}(I_p,I_q) - 2\min D_{\mathcal{N}}^{(w)}(I_p,I_q) \right] |\widehat{\tau}(\theta)\widehat{\tau}(\tilde{\theta})|d\theta d\tilde{\theta}$$

$$= 2\left(\mathrm{diam}\, D_{\mathcal{N}}^{(w)}(I_p,I_q)\right) \int_{I_p} |\widehat{\tau}(\theta)|d\theta \int_{I_q} |\widehat{\tau}(\tilde{\theta})|d\tilde{\theta}, \qquad (9.3)$$

where $\mathrm{diam}(D_{\mathcal{N}}^{(w)}(I_p,I_q))$ is the diameter of the set

$$D(I_p,I_q) = \left\{ D_{\mathcal{N}}^{(w)}(\theta,\tilde{\theta}) \mid (\theta,\tilde{\theta}) \in I_p \times I_q \right\}.$$

Let

$$\left\|\partial D_{\mathcal{N}}^{(w)}\right\|_\infty = \sup_{\theta,\tilde{\theta}} \max\left\{ \left|\frac{\partial}{\partial\theta}D_{\mathcal{N}}^{(w)}(\theta,\tilde{\theta})\right|, \left|\frac{\partial}{\partial\tilde{\theta}}D_{\mathcal{N}}^{(w)}(\theta,\tilde{\theta})\right| \right\},$$

and let $t_0 = 1,\ldots,t_N = T$ be evenly spaced so that $|I_p| \leq T/N$. Then by (9.3),

$$\iint_{S_{pq}} D_{\mathcal{N}}^{(w)}(\theta,\tilde{\theta})d\mu(\theta)d\mu(\tilde{\theta}) \leq \frac{4T}{N}\left\|\partial D_{\mathcal{N}}^{(w)}\right\|_\infty \int_{I_p} |\widehat{\tau}(\theta)|d\theta \int_{I_q} |\widehat{\tau}(\tilde{\theta})|d\tilde{\theta}. \qquad (9.4)$$

Summing over p,q and applying (9.2) we get

$$\|S_w - \tilde{\alpha}_w\|_{\mathcal{N}}^2 \leq \frac{2}{\pi}\frac{T}{N}\left\|\partial D_{\mathcal{N}}^{(w)}\right\|_\infty \left(\int_{t\leq|\theta|\leq T} |\widehat{\tau}(\theta)|\right)^2.$$

The proof is complete. $\qquad\square$

Lemma 9.3. *Let $B_{ij}^{(v,w)} = \mathfrak{F}\rho(v_i - w_j)$ as in Definition 5.1. Then, $|(\partial/\partial B_{ij})\det B| \leq 1$ at any $B = B^{(v,w)}$.*

Proof.

$$\frac{\partial}{\partial B_{ij}} \det B = (-1)^{i+j} m_{i,j}^B,$$

where $m_{i,j}^B$ is the i, j-th minor of B. But $m_{ij}^B = \det B^{(\tilde{v}, \tilde{w})}$ where $\tilde{v} = (v_{i'})_{i' \neq i}$ and $\tilde{w} = (w_{j'})_{j' \neq j}$, and

$$|\det B^{(\tilde{v}, \tilde{w})}| = |D_{\mathcal{N}}(\tilde{v}, \tilde{w})| \leq 1$$

by the properties mentioned in Section 6.1. $\qquad\square$

Corollary 9.1.

$$\left| \frac{\partial}{\partial \theta} D_{\mathcal{N}}^{(w)}(\theta, \tilde{\theta}) \right| \leq \frac{n^{\frac{3}{2}} \|w\|}{\sqrt{e}}.$$

Proof. By the chain rule,

$$\frac{\partial}{\partial \theta} D_{\mathcal{N}}^{(w)}(\theta, \tilde{\theta}) = \sum_{ij} \frac{\partial B_{ij}(\theta w, \tilde{\theta} w)}{\partial \theta} \frac{\partial \det B}{\partial B_{ij}},$$

where

$$\left| \frac{\partial B_{ij}(\theta w, \tilde{\theta} w)}{\partial \theta} \right| = \left| w_i \cdot (\nabla \mathfrak{F}_d \rho)(\theta w_i - \tilde{\theta} w_j) \right| \leq \|w_i\| \cdot \|\nabla \mathfrak{F}_d \rho\|_\infty.$$

Lemma 9.3 and the triangle inequality then imply that

$$\left| \frac{\partial}{\partial \theta} D_{\mathcal{N}}^{(w)}(\theta, \tilde{\theta}) \right| \leq n \|\nabla \mathfrak{F}_d \rho\|_\infty \sum_i \|w_i\| \leq n^{\frac{3}{2}} \|\nabla \mathfrak{F}_d \rho\|_\infty \|w\|,$$

where the last inequality is by Cauchy-Schwarz. For $\rho = \mathcal{N}(0, I_d)$ we have

$$\mathfrak{F}_d \mathcal{N}(w) = e^{-\frac{\|w\|^2}{2}}, \quad \|\nabla \mathfrak{F}_d \mathcal{N}\|_\infty = \frac{1}{\sqrt{e}}.$$

The proof is complete. $\qquad\square$

In the presence of bias terms b_k in (2.3) we can efficiently compute an approximation $S_{w,b}$ to the function $\mathcal{A}\tau(w^{(k)} \cdot [\mathbf{x}] + b_k)$ with the same error bound. Indeed, the bias term results in a shift of the activation function which corresponds to multiplying the Fourier transform by an oscillating phase. Since the upper bounds do not depend on the phase of the Fourier transform, the same truncation error bound applies. Theorem 3.1 holds for arbitrary bias terms for the same reason.

Proof of Theorem 3.3. Because τ is rough we have

$$\mathbb{E}[\|\mathcal{A}f\|_{\mathcal{N}} | W] = n^{-\mathcal{O}(1)}$$

with probability $1 - o(1)$. Given target relative error $\epsilon = n^{-\mathcal{O}(1)}$ it suffices to achieve absolute error

$$\epsilon' = \epsilon \cdot \mathbb{E}[\|\mathcal{A}f\|_{\mathcal{N}}|W] = n^{-\mathcal{O}(1)}.$$

By Lemmas 9.1 and 9.2

$$\|S_w - \mathcal{A}\tau_w\|_{\mathcal{N}}^2 = \mathcal{O}\left(\epsilon' + n^{\mathcal{O}(1)}\frac{\|\partial_\theta D_{\mathcal{N}}^{(w)}\|_\infty}{N}\right).$$

Corollary 9.1 shows that $\|\partial_\theta D_{\mathcal{N}}^{(w)}\|_\infty = n^{\mathcal{O}(1)}$, so

$$\|S_w - \mathcal{A}\tau_w\|_{\mathcal{N}}^2 = \mathcal{O}\left(\epsilon' + \frac{n^{\mathcal{O}(1)}}{N}\right).$$

It then suffices to pick $N = n^{\mathcal{O}(1)}$. Let

$$S_{W,a,b} = \sum_{k=1}^m a_k S_{w^{(k)},b_k}$$

and apply Lemma 4.1 to the difference $f_{W,a,b} - S_{W,a,b}$ to extend the termwise error bound to the sum (2.3). The computational cost of evaluating $S_w(x)$ is $\mathcal{O}(n^3 N)$. $\square$

9.2 Numerical demonstration of Theorem 3.3

We numerically demonstrate Theorem 3.3 by approximating the anti-symmetrization of a single neuron with the ReLU activation function (Figs. 9.1 and 9.2). We compare the
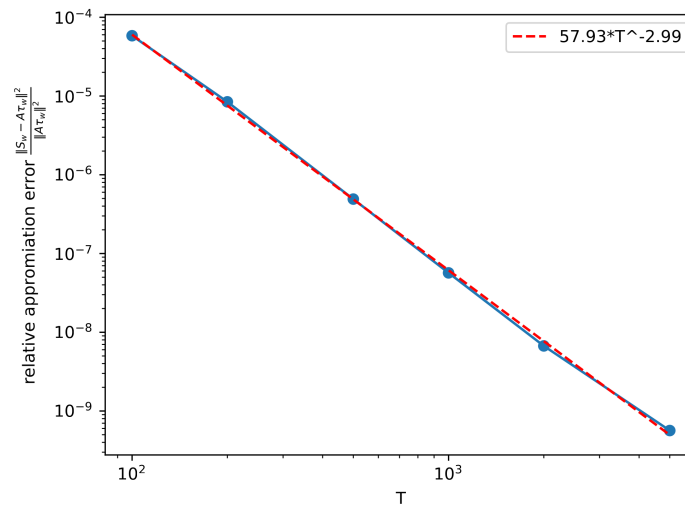

Figure 9.1: Approximation error in Theorem 3.3 for a single ReLU neuron as a function of the ultraviolet truncation T. Values plotted are $T = 100, 200, 500, 1000, 2000, 5000$. Here, $n = 8$ and $d = 3$. Here, the number of quadrature points $N = 10^4$ and the infra-red cutoff $t = 0.1$ is kept constant.
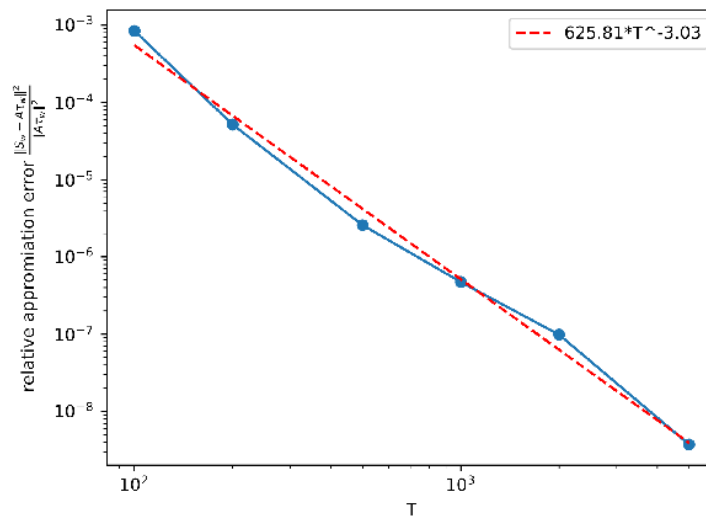
Figure 9.2: The approximation error of Theorem 3.3 as in Fig. 9.1, except the number of quadrature points N is chosen to grow with the ultraviolet cutoff T (specifically, we take $N = T$).

approximation S_w given by Theorem 3.3 with the explicit anti-symmetrization $\mathcal{A}\tau_w$. We use 100 sample points $x \sim \mathcal{N}(0, I)$ to estimate the norm of the anti-symmetrized function and the distance between the true anti-symmetrization and its efficient approximation. In our implementation we used Gauss-Legendre quadrature to estimate the integral in (9.1).

The cubic convergence with T observed in Figs. 9.1 and 9.2 is in accordance with our theory, because the anti-symmetrization operator $\mathcal{A}$ is approximately an isometry for highly oscillating functions, i.e. in the within the ultraviolet part. By Plancherel's equality we can approximate the squared error introduced by the tail truncation as the squared L^2 norm of the truncated tail, which is of order $\int_{|\theta|>T} |\hat{\tau}(\theta)|^2 d\theta \propto \int_{|\theta|>T} |\theta|^{-4} \propto T^{-3}$ when τ is the ReLU activation.

10   Empirical generalization to multi-layer networks

It is natural to ask whether the advantage of rough activation functions against cancellations remains as the depth of the neural network grows. We consider networks of depth $L = 3, 4, 5$ and compare $\|\mathcal{A}f\|_{\mathcal{N}}^2$ between two choices of activation functions: The smooth tanh and the rough normalized double ReLU (DReLU) $\tau_\kappa(y) = \kappa \max\{-1, \min\{1, y\}\}$ where $\kappa \approx 0.875$ is chosen such that $\mathbb{E}[|\tau_\kappa(Z)|^2] = \mathbb{E}[|\tanh(Z)|^2]$ for standard-Gaussian $Z \sim \mathcal{N}(0, 1)$. Fig. 10.1 shows a comparison between the DReLU and tanh activation functions. We take $d = 3$, let all layers have width $m = 3n$, and instantiate NNs from the Xavier initialization (independent Gaussian weights with variance $1/m$ where m is the number of neurons in the preceding layer). Fig. 10.2 shows that the rough activation function maintains its advantage for networks with more layers.
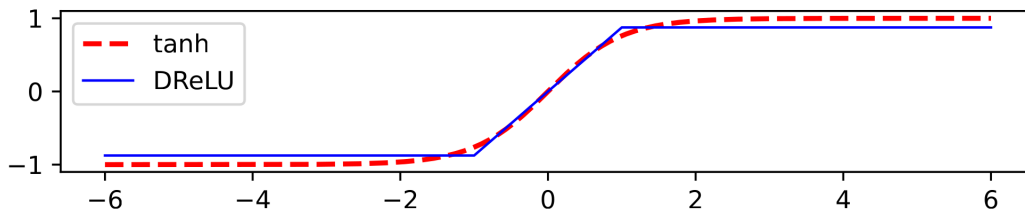
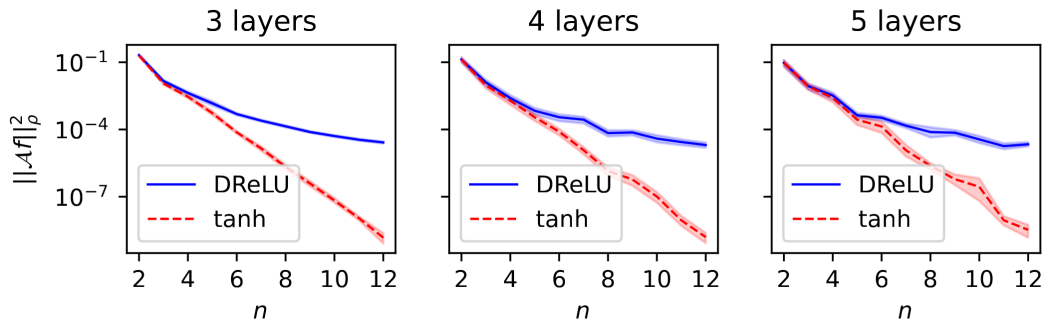Figure 10.1: Tanh and the normalized DReLU.

Figure 10.2: Log-scale comparison between $\mathbb{E}[\|\mathcal{A}f\|_{\mathcal{N}}^2]$ for smooth and rough activation functions (tanh vs normalized DReLU) for antisymmetrized neural networks of different depths. Shaded areas show 90% confidence regions.

11 Conclusion

Using the Fourier representation of the activation function, we observe that a rough activation function is necessary to tame the near-exact cancellations when antisymmetrizing two-layer NNs initialized with the standard initializations. Equivalently, an architecture based on a smooth activation function would require an initialization of the weights in the first layer distinct from the standard Xavier/He initializations to avoid the sign problem in the antisymmetric setting. The Fourier perspective also provides a polynomial-time algorithm for approximately evaluating explicitly antisymmetrized two-layer NNs. It may be possible that explicitly antisymmetrized two-layer NNs provides a path towards universal approximation of a class of antisymmetric functions without suffering from curse of dimensionality. Our work also raises intriguing open questions about how the cancellations and efficient algorithms generalize to antisymmetrized multi-layer NNs as well as to the training regime.

Appendix A. Lower-bounding the typical separation

Lemma A.1. *For $w \sim \mathcal{N}(0, (2/nd)I_{nd})$ sampled from the He initialization,*

$$P(\delta_w < \delta) \le 2 \binom{n}{2} \left(\frac{2nd\delta^2}{\tilde{c}\pi} \right)^{\frac{d}{2}} |B_d|, \tag{A.1}$$

where $|B_d|$ is the volume of the unit ball in $\mathbb{R}^d$. For constant d this is $\mathcal{O}(n^{2+d/2}\delta^d)$.

Proof. If w_i are sampled independently from a distribution $\mathcal{W}$ on $\mathbb{R}^d$, then

$$
\begin{aligned}
P(\delta_w < \delta) &\leq \sum_{i<j} P\left(\frac{1}{2}\|w_i - w_j\| < \delta\right) + P\left(\frac{1}{2}\|w_i + w_j\| < \delta\right) \\
&\leq \sum_{i<j} \max_{w' \in \mathbb{R}^d} P(\|w_i - w_j\| < 2\delta \,|\, w_j = w') \\
&\quad + \max_{w' \in \mathbb{R}^d} P(\|w_i + w_j\| < 2\delta \,|\, w_j = w') \\
&\leq 2\binom{n}{2}(2\delta)^d |B_d| \, \|\mathcal{W}\|_\infty,
\end{aligned}
\tag{A.2}
$$

where $\|\mathcal{W}\|_\infty$ is the supremum of the density. For the He initialization $w_i \sim \mathcal{N}(0, 2I_d/(nd))$ we have

$$
\|\mathcal{W}\|_\infty = \left(2\pi \cdot \frac{\tilde{c}}{nd}\right)^{-\frac{d}{2}} = \left(\frac{2\tilde{c}\pi}{nd}\right)^{-\frac{d}{2}}.
$$

The proof is complete. $\qquad\square$

For fixed d, $v = w^{(k)}$ satisfies that $P(\delta_v < \delta) = \mathcal{O}(n^{2+d/2}\delta^d)$. Then with probability $1 - o(1)$, v has typical separation as defined in Definition 4.3.

Proof of Lemma 4.2. Recall that by definition, W is typical if each $w^{(k)}$ is typical and at least half the $w^{(k)}$ have typical separation. The distribution of $(nd/\tilde{c})\|w^{(k)}\|^2 \sim \chi^2(nd)$ implies that for each $k = 1, \ldots, m$,

$$
P\left(\|w^{(k)}\|^2 < \frac{\tilde{c}}{2}\right) \leq \left(\frac{\sqrt{e}}{2}\right)^{\frac{nd}{2}}, \quad P(\|w^{(k)}\|^2 > 2\tilde{c}) \leq \left(\frac{2}{e}\right)^{\frac{nd}{2}}
$$

since $\mathbb{E}[\|w^{(k)}\|^2] = \tilde{c}$. By a union bound we have $\tilde{c}/2 \leq \|w^{(k)}\|^2 \leq 2\tilde{c}$ for all $k = 1, \ldots, m$ with probability $1 - 2m2^{-\Omega(nd)}$. Furthermore, $w_{ij}^{(k)} \sim \mathcal{N}(0, \tilde{c}/(nd))$ and a union bound imply that

$$
P\left(|w_{ij}^{(k)}| \geq t \text{ for some } i, j, k\right) \leq 2mnde^{-\frac{ndt^2}{2\tilde{c}}} = \mathcal{O}\left(e^{-\frac{ndt^2}{2\tilde{c}} + (C'+1)\log n + \log d}\right).
$$

Given any $C'' > 0$ we may let

$$
t = 2\sqrt{C' + C'' + 1}\sqrt{\frac{\log(nd)}{nd}}
$$

and obtain that $\|w^{(k)}\|_\infty \leq t$ for all k with probability at least $1 - n^{-C''}$. This shows that each w is typical with probability $1 - 1/n$ for appropriate C in Definition 4.1.

If $P(\delta_{w^{(k)}} \geq \delta) \to 1$ for each fixed k then the probability that $\delta(w^{(k)} \geq \delta)$ for at least half the $k = 1, \ldots, m$ also converges to 1. Eq. (4.3) follows from Lemma 4.1. $\qquad\square$

Appendix B. Bound on complex-exponential overlap kernel for the Gaussian envelope

In this section, we recall the proof of the bound on a determinant of exponentials given in [1].

Lemma B.1. *Let $v = (v_1, \ldots, v_n)^T \in \mathbb{R}^{n \times d}$ and $w = (w_1, \ldots, w_n)^T \in \mathbb{R}^{n \times d}$. Then*

$$(e^{v_i \cdot w_j})_{ij} = \sum_{k=0}^{\infty} Q_k,$$

where

$$\operatorname{rank} Q_k \leq \binom{k+d-1}{d-1}, \quad \|Q_k\| \leq \frac{n(\|v\|_\infty \|w\|_\infty d)^k}{k!}. \tag{B.1}$$

Proof. Let $(c_1, \ldots, c_d)$ and $(\tilde{c}_1, \ldots, \tilde{c}_d)$ be the columns of v and w and let $\odot$ denote elementwise operations

$$(e^{v_i \cdot w_j})_{ij} = e^{\odot \sum_{i=1}^d c_i \tilde{c}_i^T} = \odot_{i=1}^d e^{\odot c_i \tilde{c}_i^T}. \tag{B.2}$$

We first consider each factor $e^{\odot c_i \tilde{c}_i^T}$ separately. Elementwise multiplication of rank-one matrices given as outer products corresponds to elementwise multiplication of the vectors, $ab^T \odot \tilde{a}\tilde{b}^T = (a \odot \tilde{a})(b \odot \tilde{b})^T$. Therefore, applying the Taylor expansion entrywise,

$$e^{\odot c \tilde{c}^T} = \sum_{k=0}^{\infty} \frac{(c\tilde{c}^T)^{\odot k}}{k!} = \sum_{k=0}^{\infty} \frac{(c^{\odot k})(\tilde{c}^{\odot k})^T}{k!}, \tag{B.3}$$

where $c = c_i$, $\tilde{c} = \tilde{c}_i$ are column vectors. Apply (B.3) to each factor of (B.2),

$$\odot_{i=1}^d e^{\odot c_i \tilde{c}_i^T} = \sum_{k_1, \ldots, k_d = 0}^{\infty} \frac{\left(\odot_{i=1}^d c_i^{\odot k_i}\right)\left(\odot_{i=1}^d \tilde{c}_i^{\odot k_i}\right)^T}{\prod_{i=1}^d k_i!}$$

$$= \sum_{k=0}^{\infty} \sum_{k_1 + \cdots + k_d = k}^{\infty} \frac{1}{k!} \binom{k}{k_1, \ldots, k_d} \left(\odot_{i=1}^d c_i^{\odot k_i}\right)\left(\odot_{i=1}^d \tilde{c}_i^{\odot k_i}\right)^T. \tag{B.4}$$

Let Q_k be the innermost sum of (B.4). We estimate the maximum over the entries

$$\|Q_k\|_{\max} \leq \frac{\|v\|_\infty^k \|w\|_\infty^k}{k!} \sum_{k_1 + \cdots + k_d = k}^{\infty} \binom{k}{k_1, \ldots, k_d} = \frac{\|v\|_\infty^k \|w\|_\infty^k d^k}{k!}$$

and apply the inequality $\|Q_k\| \leq n\|Q_k\|_{\max}$. $\qquad\square$

Lemma B.2. *Let $\lambda_0 \geq \lambda_1 \geq \cdots$ be the absolute values of the eigenvalues of $(e^{v_i \cdot w_j})_{ij}$ and let $\mu = \|v\|_\infty \|w\|_\infty d$. Then $\lambda_0 \leq ne^\mu$, and for $\mu \leq 1/2$,*

$$\lambda_L \leq \frac{2n}{p!} \mu^p, \quad L = \binom{p+d-1}{d}, \tag{B.5}$$

where the case $p = 0$ of (B.5) *holds with the interpretation*

$$L = \binom{d-1}{d} = 0, \quad \lambda_0 \leq ne^{\frac{1}{2}} \leq 2n.$$

Proof. From the identity

$$\binom{p+d-1}{d} = 1 + d + \binom{d+1}{d-1} + \cdots + \binom{p+d-2}{d-1},$$

there are

$$1 + d + \binom{d+1}{d-1} + \cdots + \binom{p+d-2}{d-1} \geq \operatorname{rank} Q_0 + \cdots + \operatorname{rank} Q_{p-1}$$

eigenvalues in front of λ_L where $L = \binom{p+d-1}{d}$, and we have used Lemma B.1. By the min-max principle,

$$\lambda_L \leq \left\| \sum_{k=p}^{\infty} Q_k \right\| \leq n \sum_{k=p}^{\infty} \frac{\mu^k}{k!} = \frac{n}{p!} \sum_{k=p}^{\infty} \mu^k = \frac{n}{p!} \frac{\mu^p}{1-\mu} \leq \frac{2n}{p!} \mu^p.$$

The proof is complete. □

Proof of Proposition 6.1. By Lemma B.2 and the assumptions on p we have

$$\lambda_{\lfloor \frac{n}{2} \rfloor} \leq \frac{2n}{p!} \mu^p \leq \frac{\mu^p}{2n},$$

and $\lambda_0 \leq 2n$ where $\mu = d\|v\|_\infty \|w\|_\infty$, so it follows that

$$\left| \det \left(\left(e^{v_i \cdot w_j}\right)_{ij} \right) \right| \leq \lambda_0^{\frac{n}{2}} \lambda_{\frac{n}{2}}^{\frac{n}{2}} \leq (\mu^p)^{\frac{n}{2}} = \left(\frac{v}{2}\right)^{pn}.$$

The proof is complete. □

Apply the bound to $D_\mathcal{N}(\theta w, \theta w)$ with $p = \Theta(n^{1/d})$ to get that for $\theta \leq t := (2\sqrt{d}\|w\|_\infty)^{-1}$,

$$D_\mathcal{N}^{(w)}(\theta, \theta) = \left(\frac{\theta}{2t}\right)^{\Omega(n^{1+\frac{1}{d}})}. \tag{B.6}$$

Eq. (5.5) implies the triangle inequality

$$\|\mathcal{A}\tau_w\|_\rho \leq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |\widehat{\tau}(\theta)| \sqrt{D_\mathcal{N}^{(w)}(\theta, \theta)} d\theta,$$

because $\sqrt{D_\rho^{(w)}(\theta, \theta)} = \|\mathcal{A}e^{i\theta w \cdot [\mathbf{x}]}\|_\rho$ by definition. Apply this triangle inequality to the low-pass part and bound the integrand using (B.6) to cancel the pole of $|\widehat{\tau}|$ at 0. We then obtain Lemma 6.4.

Lemma 6.4. *Let $\ell = \tau^{\mathrm{LP}(t)}$ be the low-pass at threshold $t = (2\sqrt{d}\|w\|_\infty)^{-1}$. If w is typical then $t = \Omega(\sqrt{n/\log n})$ and $\|\mathcal{A}\ell_w\|_\mathcal{N} = \mathcal{O}(2^{-\Omega(n^{1+1/d})})$.*

Proof. The lower bound on t follows directly from its definition and the definition of w being typical.

Bound $D_\mathcal{N}^{(w)}$ as in (B.6) and write $|\widehat{\tau}(\theta)| = \mathcal{O}(|\theta|^{-r} + 1)$. The triangle inequality yields

$$\left\|\tau^{\mathrm{LP}(t)}\right\|_\mathcal{N} = \mathcal{O}\left(2^{-\frac{pn}{2}} \int_{-t}^{t} \left(\frac{|\theta|}{t}\right)^{\frac{pn}{2}} (|\theta|^{-r} + 1)d\theta\right) = \mathcal{O}\left(2^{-\frac{pn}{2}} t(t^{-r} + 1)\right),$$

where $p = \Omega(n^{1/d})$. Here we have cancelled the pole $|\theta|^{-r}$ by writing

$$\left(\frac{|\theta|}{t}\right)^{\frac{pn}{2}} |\theta|^{-r} \leq \left(\frac{|\theta|}{t}\right)^{r} |\theta|^{-r} = t^{-r},$$

so that the integrand is bounded by $t^{-r} + 1$.   □

Appendix C. Generalized definition of roughness

Definition C.1 (Generalized Rough Activation Functions). *$\tau : \mathbb{R} \to \mathbb{C}$ is generalized rough if its Fourier transform $\widehat{\tau}$ has tail decay $K < \infty$, i.e. if $\mathfrak{T}_{\widehat{\tau}}(t) = t^{-\mathcal{O}(1)}$, and if additionally,*

1. *$|\widehat{\tau}(\theta)|^2/\mathfrak{T}_{\widehat{\tau}}(\theta) \to 0$ as $\theta \to \infty$.*

2. *There exists $\gamma > 0$ such that for $1/2 \leq \sigma \leq 2$ and all $\theta \in \mathbb{R}\backslash[-1, 1]$ where $\widehat{\tau}(\theta) \neq 0$,*

$$\mathrm{Re}\left(\frac{\mathbb{E}[\widehat{\tau}(\theta \pm |Z|)]}{\widehat{\tau}(\theta)}\right) \geq \gamma, \tag{C.1}$$

   where $Z \sim \mathcal{N}(0, \sigma^2)$. Here, "$\pm$" is taken the sign of θ, and Re is the real part of a complex number.

The assumptions of Items 1 and 2 prevent activation functions with excessive negative correlations between nearby frequencies. The constraint on the standard deviation $1/2 \leq \sigma \leq 2$ is related to the Xavier/He initialization in Definition 3.4.

Appendix D. The Fourier inversion formula holds for ReLU and tanh

We include the derivation (see [1]) to verify Eq. (3.1) which asserts that $\tau^{\mathrm{LP}(\epsilon)} = p + C_\epsilon + \mathcal{O}(\epsilon g)$ where p is a low-degree polynomial and g is bounded by a polynomial. Here we have defined $\tau^{\mathrm{LP}(\epsilon)} = \tau - \tau^{\mathrm{HP}(\epsilon)}$.

1. $\tau = $ ReLU: We first evaluate the high-pass part

$$\tau^{\text{HP}(\epsilon)}(y) = \frac{|y|}{2} - \frac{\cos(\epsilon y)}{\pi \epsilon} - \frac{y \, \text{Si}(\epsilon y)}{\pi},$$

where

$$\text{Si}(y) = \int_0^y \frac{\sin s}{s} ds.$$

Since $\text{ReLU}(y) = |y|/2 + y/2$,

$$\tau^{\text{LP}(\epsilon)}(y) = \frac{y}{2} + \frac{\cos(\epsilon y)}{\pi \epsilon} + \frac{y \, \text{Si}(\epsilon y)}{\pi}. \tag{D.1}$$

Write $\tau^{\text{LP}(\epsilon)} = p + C_\epsilon + \varepsilon$, where

$$p(y) = \frac{y}{2}, \quad C_\epsilon = \frac{1}{\pi \epsilon}.$$

Then the remainder satisfies

$$|\varepsilon| \leq \left| \frac{\cos(\epsilon y) - 1}{\pi \epsilon} \right| + \left| \frac{y \, \text{Si}(\epsilon y)}{\pi} \right|$$

$$\leq \frac{(\epsilon y)^2}{2\pi \epsilon} + \frac{y \cdot (\epsilon y)}{\pi} = \epsilon g(y), \quad g(y) := \frac{3}{2\pi} y^2. \tag{D.2}$$

2. $\tau = $ tanh: We can write the low-pass part as an absolutely convergent integral as

$$\tau^{\text{LP}(\epsilon)}(y) = \frac{1}{\sqrt{2\pi}} \int_{-\epsilon}^{\epsilon} \frac{-i\sqrt{\pi/2}}{\sinh(\pi\theta/2)} (e^{i\theta y} - 1) d\theta.$$

Let $p, C_\epsilon \equiv 0$ and bound

$$\left| \tau^{\text{LP}(\epsilon)}(y) \right| \leq \frac{1}{\sqrt{2\pi}} \int_{-\epsilon}^{\epsilon} \frac{\sqrt{\pi/2}}{|\pi\theta/2|} |\theta y| d\theta = \epsilon g(y), \quad g(y) := \frac{2}{\pi} |y|. \tag{D.3}$$

Acknowledgments

References

[1] N. Abrahamsen and L. Lin, Anti-symmetric Barron functions and their approximation with sums of determinants, *arXiv:2303.12856*, 2023.

[2] D. M. Ceperley, Fermion nodes, *J. Stat. Phys.*, **63**(5-6):1237–1267, 1991.

[3] J. Han, L. Zhang, and W. E, Solving many-electron Schrödinger equation using deep neural networks, *J. Comput. Phys.*, **399**:108929, 2019.

[4] J. Hermann, Z. Schatzle, and F. Noe, Deep neural network solution of the electronic Schrödinger equation, *Nature Chem.*, **12**:891–897, 2020.

[5] J. Lin, G. Goldshlager, and L. Lin, Explicitly antisymmetrized neural network layers for variational Monte Carlo simulation, *J. Comput. Phys.*, **474**:111765, 2023.

[6] E. Loh, J. Gubernatis, R. Scalettar, S. White, D. Scalapino, and R. Sugar, Sign problem in the numerical simulation of many-electron systems, *Phys. Rev. B*, **41**(13):9301, 1990.

[7] P.-O. Löwdin, Quantum theory of many-particle systems. I. Physical interpretations by means of density matrices, natural spin-orbitals, and convergence problems in the method of configurational interaction, *Phys. Rev.*, **97**:1474–1489, 1955.

[8] D. Luo and B. K. Clark, Backflow transformations via neural networks for quantum many-body wave functions, *Phys. Rev. Lett.*, **122**:226401, 2019.

[9] D. Pfau, J. S. Spencer, A. G. D. G. Matthews, and W. M. C. Foulkes, Ab initio solution of the many-electron Schrödinger equation with deep neural networks, *Phys. Rev. Research*, **2**:033429, 2020.

[10] M. Reed and B. Simon, *I: Functional Analysis. Methods of Modern Mathematical Physics*, Elsevier Science, 1981.

[11] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, A simple neural network module for relational reasoning, *Adv. Neural Inf. Process. Syst.*, **30**, 2017.

[12] S. Sorella, S. Baroni, R. Car, and M. Parrinello, A novel technique for the simulation of interacting fermion systems, *EPL*, **8**(7):663–668, 1989.

[13] J. Stokes, J. R. Moreno, E. A. Pnevmatikakis, and G. Carleo, Phases of two-dimensional spinless lattice fermions with first-quantized deep neural-network quantum states, *Phys. Rev. B*, **102**:205122, 2020.

[14] D. Yarotsky, Universal approximations of invariant maps by neural networks, *Constr. Approx.*, **55**(1):407–474, 2022.

[15] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, Deep sets, *Adv. Neural Inf. Process. Syst*, **30**, 2017.

[16] A. Zweig and J. Bruna, A Functional perspective on learning symmetric functions with neural networks. In: *Proceedings of the 38th International Conference on Machine Learning*, PMLR, **139**:13023–13032, 2021.

[17] A. Zweig and J. Bruna, Towards antisymmetric neural Ansatz separation, *arXiv:2208.03264*, 2022.